5-3-2013

Bayes multiple decision functions

Wensong Wu
Department of Mathematics and Statistics, Florida International University, wenswu@fiu.edu

Edsel A. Pena
University of South Carolina

Electronic Journal of Statistics
Vol. 7 (2013) 1272–1300
ISSN: 1935-7524
DOI: 10.1214/13-EJS813

Bayes multiple decision functions

Wensong Wu

Division of Statistics, Department of Mathematics and Statistics
Florida International University, Miami, Florida 33199
e-mail: wenswu@fiu.edu

and

Edsel A. Peña

Department of Statistics
University of South Carolina, Columbia, SC 29208
e-mail: pena@stat.sc.edu

Abstract: This paper deals with the problem of simultaneously making many (M) binary decisions based on one realization of a random data matrix $\mathbf{X}$. M is typically large and $\mathbf{X}$ will usually have M rows associated with each of the M decisions to make, but for each row the data may be low dimensional. Such problems arise in many practical areas such as the biological and medical sciences, where the available dataset is from microarrays or other high-throughput technology and with the goal being to decide which among of many genes are relevant with respect to some phenotype of interest; in the engineering and reliability sciences; in astronomy; in education; and in business. A Bayesian decision-theoretic approach to this problem is implemented with the overall loss function being a cost-weighted linear combination of Type I and Type II loss functions. The class of loss functions considered allows for use of the false discovery rate (FDR), false nondiscovery rate (FNR), and missed discovery rate (MDR) in assessing the quality of decision. Through this Bayesian paradigm, the Bayes multiple decision function (BMDF) is derived and an efficient algorithm to obtain the optimal Bayes action is described. In contrast to many works in the literature where the rows of the matrix $\mathbf{X}$ are assumed to be stochastically independent, we allow a dependent data structure with the associations obtained through a class of frailty-induced Archimedean copulas. In particular, non-Gaussian dependent data structure, which is typical with failure-time data, can be entertained. The numerical implementation of the determination of the Bayes optimal action is facilitated through sequential Monte Carlo techniques. The theory developed could also be extended to the problem of multiple hypotheses testing, multiple classification and prediction, and high-dimensional variable selection. The proposed procedure is illustrated for the simple versus simple hypotheses setting and for the composite hypotheses setting through simulation studies. The procedure is also applied to a subset of a microarray data set from a colon cancer study.

Keywords and phrases: Archimedean copula, Bayesian framework, decision theoretic framework, false discovery rate, frailty, multiple testing, sequential Monte Carlo.

Received December 2012.

1. Introduction

The advent of computer-automated high-throughput data-gathering technology, epitomized by the microarray, has led to the generation of so-called "large M, small n" data sets, which are those characterized by a large number, M, of variables (hereon called *genes* for historical reasons), which are observed or measured on a relatively small number, n, of subjects or units. Examples of such data sets in different scientific fields could, for instance, be found in [10] and [6].

For such data sets a typical goal is to choose an action, a_m, associated with gene m, from a set of possible actions, $\mathcal{A}_m$, for each of the M genes. For example, in a two-group microarray data set, one may want to decide, for each gene, whether it is differentially expressed between the two groups (action is $a = 1$), or whether it is not differentially expressed between the two groups (action is $a = 0$). This situation corresponds to the problem of simultaneously testing multiple pairs of null and alternative hypotheses.

In this paper we shall focus on these two-point action spaces for each of the genes, that is, those with $\mathcal{A}_m = \{0, 1\}$. Of interest therefore is to choose a vector of actions

$$\mathbf{a} = (a_1, a_2, \ldots, a_M)^{\mathsf{T}} \in \mathcal{A} = \{0, 1\}^M$$

based on the observed "large M, small n" data set. For the mth gene there will be associated a $\theta_m \in \{0, 1\}$, which is unknown, representing the *correct* action to take. Thus, for the M genes there will be an unknown vector

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_M)^{\mathsf{T}} \in \Theta = \{0, 1\}^M$$

representing the vector of actions that *ought* to be taken. This $\boldsymbol{\theta}$ will be referred to as the *state of reality*. In light of this state of reality vector $\boldsymbol{\theta}$, a chosen action vector $\mathbf{a}$ will have consequences quantified through a loss. That is, there will be a mapping

$$(\mathbf{a}, \boldsymbol{\theta}) \mapsto L(\mathbf{a}, \boldsymbol{\theta})$$

where $L(\mathbf{a}, \theta)$ is the loss that is incurred with the action $\mathbf{a}$ when reality is $\boldsymbol{\theta}$. Such a loss must take into account the loss incurred when the action is $a = 1$ when reality is $\theta = 0$, called a Type I error, as well as the loss incurred when the action is $a = 0$ when reality is $\theta = 1$, called a Type II error. There could be a variety of ways of measuring the overall Type I and Type II errors in such multiple decision problems, which will be formally described in Section 2.

These multiple decision problems appertaining to such "large M, small n" data sets lend naturally to a decision-theoretic framework discussed in more detail in Section 2. In addition to this decision-theoretic framework, we implement a Bayesian approach to decision-making by putting a prior probability distribution on the unknown state of reality $\boldsymbol{\theta}$. Coupled with the appropriate loss function, we obtain the Bayes multiple decision action. To achieve this, we obtain the mathematical form of the Bayes multiple decision function, abbreviated BMDF, and describe an efficient computational implementation of this BMDF under varied combinations of loss functions and data structures.

A decision-theoretic and a Bayesian approach to these multiple decision problems with high-dimensional data is certainly not new as can be seen in [15, 21, 20, 4, 3, 17]. Other approaches are in [22, 23] and [18]. See also the monograph [7]. An innovative major contribution of this paper is the use of a general class of loss functions that encompasses many of the loss functions that have been used in earlier works, thereby leading to a unified treatment of the multiple decision problem. For instance, the general class of loss functions introduced in Section 2 includes as special cases those that involve false positives and false negatives as well as the commonly-used false discovery rates and false nondiscovery rates. Another major contribution is an efficient algorithm for computationally finding the Bayes multiple decision action, an algorithm that has computational order of at most $O(M^2 \log M)$. Many papers have dealt with the situation where the observables from each of the genes are stochastically independent. We go beyond this usual assumption by incorporating dependencies among these observables, with the dependence structure induced by frailty-type models, which also takes the form of Archimedean copulas. This dependent modeling approach utilizes ideas from survival analysis where frailty and copula models have been used to model associations (see, for instance, [11]).

The statistical models governing these multiple decision problems in fact possess more complications than the simplistic description above. This is so since, even though the parameter of main interest is the state of reality vector $\boldsymbol{\theta}$, there will be unknown model parameters that are present which are nuisance, and we need to deal with them in constructing Bayes multiple decision functions and specifying prior probabilities. In our development of the BMDF we therefore first consider the situation of a simple null hypothesis versus a simple alternative hypothesis setting, wherein the distributional model for the random observable for the mth gene is completely known under either $\theta_m = 0$ or $\theta_m = 1$. We then utilize the results for this setting to solve the problem for a composite null hypothesis versus a composite alternative hypothesis setting, which are settings with nuisance parameters. An interesting development is the use of Sequential Monte Carlo (SMC) techniques to numerically approximate the Bayes multiple decision action especially in the presence of associations among the observables and in the prior probability specification.

We outline the contents of this paper. In Section 2 we will introduce the mathematical setting and elements of the multiple decision problem, including the general class of loss functions. Section 3 will demonstrate the general form of the BMDF along with a computationally efficient algorithm of finding this BMDF in both simple and composite hypotheses testing settings. Section 4 will give the expressions of the BMDF under three concrete loss functions. We will introduce the frailty-based dependent data models in Section 5, and in Section 6 we will discuss computational aspects of the posterior calculations under the dependent models, and give algorithms using Sequential Importance Sampling (SIS). In Section 7 we will illustrate the BMDF in some concrete multiple decision problems, and compare the performance with currently used procedures via simulation studies. We will also apply the BMDF to a subset of a microarray data set. We will conclude the paper in Section 8 with some remarks.

2. Elements of decision theory

2.1. Multiple decision problem

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a basic statistical model with $\mathcal{P}$ being a collection of probability measures on $(\Omega, \mathcal{F})$. For $m = 1, 2, \ldots, M$, where $M \geq 1$ is a known integer, let $X_m : (\Omega, \mathcal{F}) \to (\mathcal{X}_m, \mathcal{B}_m)$, where $\mathcal{X}_m$ is some space and $\mathcal{B}_m$ is an associated σ-field of subsets of $\mathcal{X}_m$. In applications, X_m represents the vector of observables for the mth gene. Let $\mathbf{X} = (X_1, X_2, \ldots, X_M) : (\Omega, \mathcal{F}) \to (\mathcal{X}, \mathcal{B})$, where $\mathcal{X} = \bigotimes_{m=1}^{M} \mathcal{X}_m$ is the product sample space and $\mathcal{B}$ is the associated product σ-field. A realization $\mathbf{X} = \mathbf{x}$ will be called a (sample) data. For any $P \in \mathcal{P}$, the induced joint probability measure of $\mathbf{X}$ is $Q = P\mathbf{X}^{-1}$, whereas the marginal probability measure of X_m is $Q_m = PX_m^{-1}$. Let $\mathcal{Q} = \{P\mathbf{X}^{-1} : P \in \mathcal{P}\}$ denote the collection of all probability measures of $\mathbf{X}$ induced by $\mathcal{P}$. Consider a mapping $\boldsymbol{\vartheta} = (\vartheta_1, \vartheta_2, \ldots, \vartheta_M)^{\mathrm{T}} : \mathcal{Q} \to \{0,1\}^M$, where $\vartheta_m : \mathcal{Q} \to \{0,1\}$ only depends on the mth marginal probability measure. In essence, the parameter of main interest is $\boldsymbol{\vartheta}(Q) = \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_M)^{\mathrm{T}}$, which takes values in the parameter space $\Theta = \{0,1\}^M$. Observe that $\mathcal{Q}$ can be decomposed via $\mathcal{Q} = \biguplus_{\boldsymbol{\theta} \in \Theta} \mathcal{Q}_{\boldsymbol{\theta}}$, where $\mathcal{Q}_{\boldsymbol{\theta}} = \{Q \in \mathcal{Q} : \boldsymbol{\vartheta}(Q) = \boldsymbol{\theta}\}, \forall \boldsymbol{\theta} \in \Theta$. Let $\mathbf{a} = (a_1, a_2, \ldots, a_M)^{\mathrm{T}}$ be an action in the action space $\mathcal{A} = \{0,1\}^M$. Let $L^\circ : \mathcal{A} \times \mathcal{Q} \to \mathbb{R}$ be a loss function such that $L^\circ(\mathbf{a}, Q) = L(\mathbf{a}, \boldsymbol{\vartheta}(Q)), \forall \mathbf{a} \in \mathcal{A}, \forall Q \in \mathcal{Q}$, where $L : \mathcal{A} \times \Theta \to \mathbb{R}$ is a loss function belonging to a class $\mathcal{L}$ discussed in Section 2.2.

Since we are implementing a Bayesian framework we may restrict the space of decision functions, $\mathcal{D}$, to be nonrandomized [2]. Such decision functions are measurable maps δ from $\mathcal{X}$ to $\mathcal{A}$ with $\delta(\mathbf{x}) = (\delta_1(\mathbf{x}), \delta_2(\mathbf{x}), \ldots, \delta_M(\mathbf{x}))$ being the action taken when observed data is $\mathbf{x}$. For a $\delta \in \mathcal{D}$, its risk function (with respect to Q) is

$$R^\circ(\delta, Q) = \sum_{\boldsymbol{\theta} \in \Theta} \left[\int_{\mathcal{X}} L(\delta(\mathbf{x}), \boldsymbol{\theta}) Q(\mathrm{d}\mathbf{x}) \right] I(\boldsymbol{\vartheta}(Q) = \boldsymbol{\theta}).$$

We shall assume that for any $\boldsymbol{\theta} \in \Theta$, $\mathcal{Q}_{\boldsymbol{\theta}}$ is an identifiable parametric class given by $\mathcal{Q}_{\boldsymbol{\theta}} = \{Q_{\boldsymbol{\theta}}(\cdot\,; \gamma_{\boldsymbol{\theta}}) : \gamma_{\boldsymbol{\theta}} \in \Gamma_{\boldsymbol{\theta}}\}$, where $\gamma_{\boldsymbol{\theta}}$ is a nuisance parameter. This implies that $\mathcal{Q}$ is an identifiable model with respect to the parameter $(\boldsymbol{\theta}, \gamma_{\boldsymbol{\theta}})$ which belongs to the enlarged parameter space $\Theta^\circ = \cup_{\boldsymbol{\theta} \in \Theta} [\{\boldsymbol{\theta}\} \times \Gamma_{\boldsymbol{\theta}}]$. Then, for any $\delta \in \mathcal{D}$ and $Q \in \mathcal{Q}$, the risk function is given by

$$R^\circ(\delta, Q) = \sum_{\boldsymbol{\theta} \in \Theta} R(\delta, (\boldsymbol{\theta}, \gamma_{\boldsymbol{\theta}})) I(\boldsymbol{\vartheta}(Q) = \boldsymbol{\theta}, Q = Q_{\boldsymbol{\theta}}(\cdot\,; \gamma_{\boldsymbol{\theta}})),$$

where $R : \mathcal{D} \times \Theta^\circ \to \mathbb{R}$ is given by

$$R(\delta, (\boldsymbol{\theta}, \gamma_{\boldsymbol{\theta}})) = E_{\mathbf{X} \sim Q_{\boldsymbol{\theta}}(\cdot; \gamma_{\boldsymbol{\theta}})} L(\delta(\mathbf{X}), \boldsymbol{\theta}) = \int_{\mathcal{X}} L(\delta(\mathbf{x}), \boldsymbol{\theta}) Q_{\boldsymbol{\theta}}(\mathrm{d}\mathbf{x}; \gamma_{\boldsymbol{\theta}}).$$

Furthermore, the decomposition of $\mathcal{Q}$ becomes

$$\mathcal{Q} = \biguplus_{\boldsymbol{\theta} \in \Theta} \left\{ Q_{\boldsymbol{\theta}}(\cdot; \gamma_{\boldsymbol{\theta}}) : \boldsymbol{\vartheta}(Q_{\boldsymbol{\theta}}(\cdot; \gamma_{\boldsymbol{\theta}})) = \boldsymbol{\theta}, \gamma_{\boldsymbol{\theta}} \in \Gamma_{\boldsymbol{\theta}} \right\}.$$

A prior probability measure on $\mathcal{Q}$ can be constructed by specifying a prior probability measure on $(\Theta^\circ, \sigma(\Theta^\circ))$, where $\sigma(\Theta^\circ)$ is the σ-field generated by a semi-ring $\mathcal{C} = \{\{\boldsymbol{\theta}\} \times C_{\boldsymbol{\theta}} : C_{\boldsymbol{\theta}} \in \sigma(\Gamma_{\boldsymbol{\theta}}), \boldsymbol{\theta} \in \Theta\}$ with $\sigma(\Gamma_{\boldsymbol{\theta}})$ a σ-field on $\Gamma_{\boldsymbol{\theta}}$. Define a probability measure Π^* on $\mathcal{C}$ such that for any $\boldsymbol{\theta}_0 \in \Theta$ and any $C_{\boldsymbol{\theta}_0} \in \sigma(\Gamma_{\boldsymbol{\theta}_0})$,

$$\Pi^*(\boldsymbol{\theta} = \boldsymbol{\theta}_0, \gamma_{\boldsymbol{\theta}} \in C_{\boldsymbol{\theta}_0}) = \Pi(\boldsymbol{\theta}_0) \int_{C_{\boldsymbol{\theta}_0}} P_{\boldsymbol{\theta}_0}(\mathrm{d}\gamma_{\boldsymbol{\theta}_0}) = \Pi(\boldsymbol{\theta}_0) P_{\boldsymbol{\theta}_0}(C_{\boldsymbol{\theta}_0}),$$

where $\Pi(\cdot)$ is a prior probability measure on Θ and $P_{\boldsymbol{\theta}}(\cdot)$ is a prior probability measure on $\Gamma_{\boldsymbol{\theta}}$. This induces, by Caratheodory's extension theorem, a prior probability measure Π° on $(\Theta^\circ, \sigma(\mathcal{C}) = \sigma(\Theta^\circ))$. Since the elements of $\mathcal{Q}$ are identified by $(\boldsymbol{\theta}, \gamma_{\boldsymbol{\theta}})$, there is a one-to-one and onto mapping $h : \Theta^\circ \to \mathcal{Q}$ with $(\boldsymbol{\theta}, \gamma_{\boldsymbol{\theta}}) \overset{h}{\leftrightarrow} Q_{\boldsymbol{\theta}}(\cdot; \gamma_{\boldsymbol{\theta}})$. Therefore, Π° determines a prior probability measure on $(\mathcal{Q}, \sigma(\mathcal{Q}))$, where $\sigma(\mathcal{Q}) = h\,\sigma(\Theta^\circ)$. The Bayes risk function of a decision function $\delta \in \mathcal{D}$ for the prior Π° is defined via

$$r_{\Pi^\circ}(\delta) = E_{Q \sim \Pi^\circ} R^\circ(\delta, Q) = \int_\Theta \int_{\Gamma_{\boldsymbol{\theta}}} R(\delta, (\boldsymbol{\theta}, \gamma_{\boldsymbol{\theta}})) P_{\boldsymbol{\theta}}(\mathrm{d}\gamma_{\boldsymbol{\theta}}) \Pi(\mathrm{d}\boldsymbol{\theta}).$$

A decision function δ^* is called a Bayes multiple decision function (BMDF) if

$$\delta^* = \arg\min_{\delta \in \mathcal{D}} r_{\Pi^\circ}(\delta).$$

The multiple decision problem is to find the BMDF, which is the Bayes optimal procedure for choosing the M-dimensional action vector. More practically, there is the issue of finding the Bayes optimal action, which is the realization of δ^*, in a computationally efficient manner.

2.2. Class of loss functions

The value $L(\mathbf{a}, \boldsymbol{\theta})$ of a loss function $L : \mathcal{A} \times \Theta \to \mathbb{R}$ quantifies the error that is committed when action $\mathbf{a}$ is chosen and $\boldsymbol{\theta}$ is the state of reality. We shall consider a class of cost-weighted loss functions whose members are of form

$$L(\mathbf{a}, \boldsymbol{\theta}) = C_0 L_0(\mathbf{a}, \boldsymbol{\theta}) + C_1 L_1(\mathbf{a}, \boldsymbol{\theta}), \tag{1}$$

where $C_0 \geq 0$ and $C_1 \geq 0$ are pre-determined costs for loss functions L_0 and L_1, respectively. L_0 will quantify the loss from the Type I errors, whereas L_1 will quantify the loss from the Type II errors. The general forms of L_0 and L_1 are, respectively,

$$L_0(\mathbf{a}, \boldsymbol{\theta}) = [\alpha_0(\mathbf{a}^\mathsf{T}\mathbf{1})\mathbf{g}_0(\mathbf{a})]^\mathsf{T} [\beta_0(\boldsymbol{\theta}^\mathsf{T}\mathbf{1})\mathbf{h}_0(\boldsymbol{\theta})];$$

$$L_1(\mathbf{a}, \boldsymbol{\theta}) = [\alpha_1(\mathbf{a}^\mathsf{T}\mathbf{1})\mathbf{g}_1(\mathbf{a})]^\mathsf{T} [\beta_1(\boldsymbol{\theta}^\mathsf{T}\mathbf{1})\mathbf{h}_1(\boldsymbol{\theta})],$$

where $\alpha_j : \mathbb{R} \to \mathbb{R}$, $\beta_j : \mathbb{R} \to \mathbb{R}$, $\mathbf{g}_j : \mathcal{A} \to \mathcal{A}$, and $\mathbf{h}_j : \Theta \to \Theta$ for $j = 0, 1$. We assume further that, for $j = 0, 1$, $\mathbf{g}_j$ is τ_j-invariant with respect to the

TABLE 1
Some examples of loss functions used in this paper. FP and FDP are Type I loss functions, while FN, MDP, FNP and AMDP are Type II loss functions

Descriptive Name	Abbreviation	$L(\mathbf{a}, \boldsymbol{\theta})$
False Positive Proportion	FP	$\frac{\mathbf{a}^{\mathrm{T}}(\mathbf{1}-\boldsymbol{\theta})}{M}$
False Negative Proportion	FN	$\frac{(\mathbf{1}-\mathbf{a})^{\mathrm{T}}\boldsymbol{\theta}}{M}$
False Discovery Proportion	FDP	$\frac{\mathbf{a}^{\mathrm{T}}(\mathbf{1}-\boldsymbol{\theta})}{(\mathbf{a}^{\mathrm{T}}\mathbf{1})\vee 1}$
Missed Discovery Proportion	MDP	$\frac{(\mathbf{1}-\mathbf{a})^{\mathrm{T}}\boldsymbol{\theta}}{(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{1})\vee 1}$
False Nondiscovery Proportion	FNP	$\frac{(\mathbf{1}-\mathbf{a})^{\mathrm{T}}\boldsymbol{\theta}}{((\mathbf{1}-\mathbf{a})^{\mathrm{T}}\mathbf{1})\vee 1}$
Adjusted Missed Discovery Proportion	AMDP	$\frac{(\mathbf{1}-\mathbf{a})^{\mathrm{T}}\boldsymbol{\theta}}{(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{1})+1}$

sub-action space $\mathcal{A}_k \equiv \{\mathbf{a} \in \mathcal{A} : \mathbf{a}^{\mathrm{T}}\mathbf{1} = k\}$ for $k \in \mathcal{M} \equiv \{0, 1, \ldots, M\}$, in the sense that there exists a mapping $\tau_j : \mathcal{M} \to \mathcal{M}$ associated with $\mathbf{g}_j$ such that $\mathbf{a} \in \mathcal{A}_k$ implies $\mathbf{g}_j(\mathbf{a}) \in \mathcal{A}_{\tau_j(k)}$. Examples of τ include the identity mapping with $\tau_0(k) = k$ and also $\tau_1(k) = M - k$. Then $\mathbf{g}_0$ with $\mathbf{g}_0(\mathbf{a}) = \mathbf{a}$ is τ_0-invariant, while $\mathbf{g}_1$ with $\mathbf{g}_1(\mathbf{a}) = \mathbf{1} - \mathbf{a}$ is τ_1-invariant. With $a \vee b = \max(a, b)$, some examples of loss functions L_0 and L_1 on $\mathcal{A} \times \Theta$ are given in Table 1.

Besides the aforementioned properties of L_0 and L_1, we also assume that L_0 and L_1 possess a complementarity property given by

$$\mathbf{g}_1(\mathbf{a}) = \mathbf{a}_0 - A_1\mathbf{g}_0(\mathbf{a})$$

for some $\mathbf{a}_0 \in \mathcal{A}$ and with $A_1 > 0$. In the multiple hypotheses testing settings considered in this paper, we will have $A_1 = 1$, $\mathbf{a}_0 = \mathbf{1}$, and $\mathbf{g}_0(\mathbf{a}) = \mathbf{a}$. This property describes a relation between L_0 and L_1 which indicates that they are loss functions having complementary behaviors. For example, FP and FDP are proportions of false discoveries, where a discovery at the mth coordinate is having $a_m = 1$, whereas FN, MDP, and FNP are proportions of false nondiscoveries. In the sequel, we will consider the pairs (FP, FN), (FDP, FNP), and (FDP, MDP) for (L_0, L_1) in the multiple decision problem. The cost constants C_0 and C_1 will generally be determined by the decision maker or subject matter specialist, and they reflect the consequences of false discoveries and false nondiscoveries.

2.3. Multiple testing problems

2.3.1. Simple hypotheses setting

The simple-versus-simple multiple hypotheses testing problem is a particular case of this multiple decision problem. Suppose that the marginal probability distribution of X_m satisfies $Q_m \in \{Q_{m0}, Q_{m1}\}$ with $Q_{m0} \neq Q_{m1}$, and the parameter vector is $\boldsymbol{\theta} = (I(Q_m = Q_{m1}), m = 1, 2, \ldots, M)$. We may consider simultaneously the M pairs of simple versus simple hypotheses $H_{m0} : Q_m = Q_{m0}$ versus $H_{m1} : Q_m = Q_{m1}$ for $m = 1, 2, \ldots, M$. In this case, for $m = 1, 2, \ldots, M$, $\theta_m = 1(0)$ indicates whether H_{m1} is (not) true, and the action $a_m = 1(0)$ means rejecting (not rejecting) H_{m0}.

Usually, independent Bernoulli priors are assigned to $\boldsymbol{\theta}$. Let $\pi_{m0}, \pi_{m1} \in (0, 1)$ be such that $\pi_{m0} + \pi_{m1} = 1$ for $m = 1, 2, \ldots, M$. The prior probability mass function π on $\boldsymbol{\theta}$ is specified by

$$\pi(\boldsymbol{\theta}) = \prod_{m=1}^{M} \pi_{m0}^{1-\theta_m} \pi_{m1}^{\theta_m} I(\theta_m \in \{0, 1\}). \tag{2}$$

In this situation, the Bayes risk function of a decision function δ for a prior mass function π of $\boldsymbol{\theta}$ is

$$r_\pi(\delta) = E_{\boldsymbol{\theta}\sim\pi}\, E_{X\sim Q_{\boldsymbol{\theta}}}\, L(\delta(\mathbf{X}), \boldsymbol{\theta}),$$

associated with a loss function $L \in \mathcal{L}$, where $Q_{\boldsymbol{\theta}}$ is the joint probability function of $\mathbf{X}$, given $\boldsymbol{\theta}$, and whose mth marginal distribution function is $Q_m = Q_{m\theta_m}$.

2.3.2. Composite hypotheses setting

Suppose that Q_m, the marginal distribution of X_m, is in a class of distributions $\mathcal{Q}_m$ given by $\mathcal{Q}_m = \{Q_m(\cdot\,; \gamma_m, \xi_m) : \gamma_m \in \Gamma_m, \xi_m \in \Xi_m\}$; Assume, for $m = 1, 2, \ldots, M$, $\Gamma_m = \Gamma_{m0} \cup \Gamma_{m1}$ and $\Gamma_{m0} \cap \Gamma_{m1} = \varnothing$. Then $\mathcal{Q}_m$ has two subclasses denoted by

$$\mathcal{Q}_{m0} = \{Q_m(\cdot\,; \gamma_m, \xi_m) : \gamma_m \in \Gamma_{m0}, \xi_m \in \Xi_m\}$$
$$\mathcal{Q}_{m1} = \{Q_m(\cdot\,; \gamma_m, \xi_m) : \gamma_m \in \Gamma_{m1}, \xi_m \in \Xi_m\}.$$

Consider the M pairs of composite hypotheses $H_{m0} : Q_m \in \mathcal{Q}_{m0}$ versus $H_{m1} : Q_m \in \mathcal{Q}_{m1}$, for $m = 1, 2, \ldots, M$. Note that Γ_{m0}, Γ_{m1}, or Ξ_m could be the same for all m, though in general they may be different. Let $\boldsymbol{\theta} = (I(Q_m \in \mathcal{Q}_{m1}), m = 1, 2, \ldots, M) = (I(\gamma_m \in \Gamma_{m1}), m = 1, 2, \ldots, M) \in \Theta = \{0, 1\}^M$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_M) \in \Gamma \equiv \bigotimes_{m=1}^{M} \Gamma_m$, and $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_M) \in \Xi \equiv \bigotimes_{m=1}^{M} \Xi_m$. Also, for any $\boldsymbol{\theta} \in \Theta$, let $\Gamma_{\boldsymbol{\theta}} \equiv \bigotimes_{m=1}^{M} \Gamma_{m\theta_m}$. Then the extended parameter vector is

$$(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) \in \Theta^\circ \equiv \biguplus_{\boldsymbol{\theta}\in\Theta} \{\{\boldsymbol{\theta}\} \times \Gamma_{\boldsymbol{\theta}} \times \Xi\}, \tag{3}$$

where $\boldsymbol{\theta} \in \Theta$ is the parameter of main interest in the multiple decision problem, while $(\boldsymbol{\gamma}, \boldsymbol{\xi}) \in \Gamma \times \Xi$ are nuisance parameters. Note that the value of $\boldsymbol{\gamma}$ determines

the value of $\boldsymbol{\theta}$, but we only want to determine whether $\gamma_m \in \Gamma_{m0}$ or $\gamma_m \in \Gamma_{m1}$ rather than estimating the exact values of γ_ms. The parameter $\boldsymbol{\theta} \in \Theta$ and the action $\mathbf{a} \in \mathcal{A}$ have the same interpretation in this composite hypotheses testing problem as in the simple-vs-simple hypotheses setting.

Assume that the prior distribution on the enlarged parameter space Θ° is

$$\pi(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) = \prod_{m=1}^{M} (\pi_{m0} p_{m0}(\gamma_m, \xi_m))^{1-\theta_m} (\pi_{m1} p_{m1}(\gamma_m, \xi_m))^{\theta_m}, \tag{4}$$

where p_{m0} and p_{m1} are prior densities on $\Gamma_{m0} \times \Xi_m$ and $\Gamma_{m1} \times \Xi_m$, respectively, and with $\pi_{m0}, \pi_{m1} \in (0,1)$ and $\pi_{m0} + \pi_{m1} = 1$. The Bayes risk function of a decision function δ for a prior density π of $(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\xi})$ associated with a loss function $L \in \mathcal{L}$ is

$$r_\pi(\delta) = E_{(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) \sim \pi} \, E_{\mathbf{X} \sim Q_{\boldsymbol{\theta}}(\cdot; \gamma, \boldsymbol{\xi})} \, L(\delta(\mathbf{X}), \boldsymbol{\theta}),$$

where $Q_{\boldsymbol{\theta}}(\cdot; \boldsymbol{\gamma}, \boldsymbol{\xi})$ is the joint probability function of $\mathbf{X}$, given $(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\xi})$, and the marginal probability measures are $Q_m = Q_m(\cdot; \gamma_m, \xi_m)$, $m = 1, 2, \ldots, M$. To indicate that the marginal $Q_m \in \mathcal{Q}_{m\theta_m}$, we shall denote it by $Q_{m\theta_m}(\cdot; \gamma_m, \xi_m)$.

3. Bayes multiple decision functions

3.1. BMDF in simple hypotheses

Let $\pi(\cdot)$ be a prior probability mass function of $\boldsymbol{\theta} \in \Theta$. Then the Bayes risk function of δ for the prior π is given by

$$r_\pi(\delta) = E_{\boldsymbol{\theta} \sim \pi} \, R(\delta, \boldsymbol{\theta}) = E_{\boldsymbol{\theta}} \, E_{\mathbf{X}|\boldsymbol{\theta}} \, L(\delta(\mathbf{X}), \boldsymbol{\theta}) = E_{\mathbf{X}} \, E_{\boldsymbol{\theta}|\mathbf{X}} \, L(\delta(\mathbf{X}), \boldsymbol{\theta}),$$

where $E_{\boldsymbol{\theta}|\mathbf{X}}$ is the expectation with respect to the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{X}$. For $\mathbf{a} \in \mathcal{A}$ and $\mathbf{X} = \mathbf{x} \in \mathcal{X}$, define the posterior expected loss by

$$\tilde{L}(\mathbf{a}, \mathbf{x}) = E_{\boldsymbol{\theta}|\mathbf{X}=\mathbf{x}} \, L(\mathbf{a}, \boldsymbol{\theta}), \tag{5}$$

and denote the optimal action when $\mathbf{X} = \mathbf{x}$ by

$$\mathbf{a}^*(\mathbf{x}) = \arg\min_{\mathbf{a} \in \mathcal{A}} \tilde{L}(\mathbf{a}, \mathbf{x}). \tag{6}$$

Then the BMDF is

$$\delta^*(\mathbf{X}) = \mathbf{a}^*(\mathbf{X}). \tag{7}$$

Notice that finding the optimal action a^*, and thus the BMDF δ^*, via equation (6) involves searching for the minimizer of the function $\tilde{L}$ among all 2^M elements of $\mathcal{A}$. When M is relatively large, the number of operations required to find the optimal action (hereon called the searching order) is of order $O(2^M)$, which would be practically infeasible to implement. Furthermore, note that this computational problem does not yet include the problem of computing the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{X} = \mathbf{x}$.

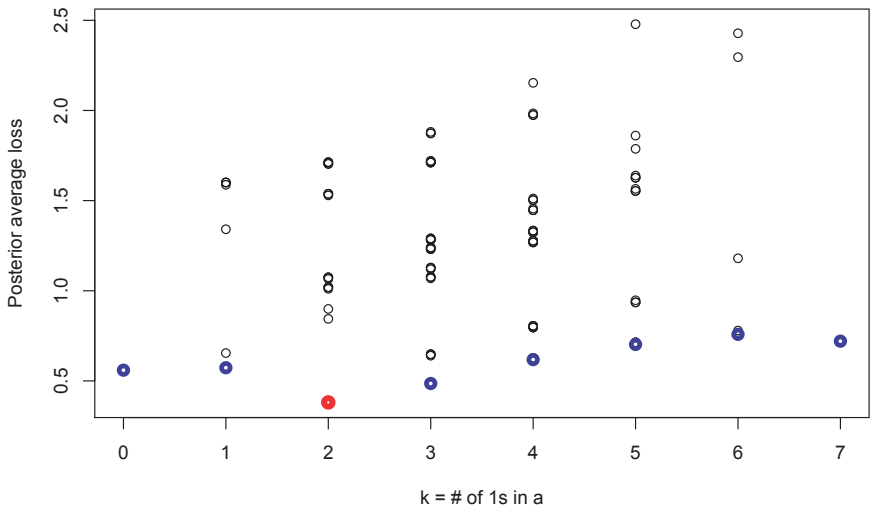

FIG 1. *Graph of $k = \mathbf{a}^{\mathsf{T}}\mathbf{1}$ versus the posterior average loss functions when $M = 7$ for a simulated data. Blue circles are restricted optimal actions within sub-action spaces; Red circle is the optimal action. Black circles represent all other actions.*

The idea for obtaining a computationally efficient algorithm to find the optimal action is to first find the restricted optimal action over the sub-action space $\mathcal{A}_k = \{\mathbf{a} \in \mathcal{A} : \mathbf{a}^{\mathsf{T}}\mathbf{1} = k\}$ for each $k \in \mathcal{M}$, and then to find the optimal action among these restricted optimal actions. Figure 1 shows an example for a simulated data when $M = 7$. The $2^M = 128$ actions in the action spaces are stratified into eight sub-action spaces according to $k = \mathbf{a}^{\mathsf{T}}\mathbf{1} = 0, 1, \ldots, 7$. The search of each restricted optimal action with the least $\tilde{L}$ within each sub-action space is computationally easier because of the form of the loss function, so that a direct search of the optimal action with the least $\tilde{L}$ among all actions is not needed.

Before presenting the results, we first define some relevant quantities that will be used. With the notation of the loss $L \in \mathcal{L}$ described in Section 2.2, for $k \in \mathcal{M}$, let

$$\mathbf{d}_0(k, \mathbf{x}) = C_0 \alpha_0(k) E_{\boldsymbol{\theta}|\mathbf{X}=\mathbf{x}} [\beta_0(\boldsymbol{\theta}^{\mathsf{T}}\mathbf{1}) \mathbf{h}_0(\boldsymbol{\theta})]; \tag{8}$$

$$\mathbf{d}_1(k, \mathbf{x}) = C_1 \alpha_1(k) E_{\boldsymbol{\theta}|\mathbf{X}=\mathbf{x}} [\beta_1(\boldsymbol{\theta}^{\mathsf{T}}\mathbf{1}) \mathbf{h}_1(\boldsymbol{\theta})]; \tag{9}$$

$$\mathbf{e}(k, \mathbf{x}) = \mathbf{d}_0(k, \mathbf{x}) - A_1 \mathbf{d}_1(k, \mathbf{x}); \tag{10}$$

and let $\mathbf{r}(k, \mathbf{x}) = (r_m(k, \mathbf{x}),\ m = 1, 2, \ldots, M)$ be the rank vector of $\mathbf{e}(k, \mathbf{x})$. Also, define

$$H(k, \mathbf{x}) = \mathbf{a}_0^{\mathsf{T}} \mathbf{d}_1(k, \mathbf{x}) + \sum_{m=1}^{M} I(r_m(k, \mathbf{x}) \le \tau_0(k)) e_m(k, \mathbf{x}), \tag{11}$$

where we recall that $\tau_0 : \mathcal{M} \to \mathcal{M}$ is a mapping such that $\mathbf{a} \in \mathcal{A}_k$ implies $\mathbf{g}_0(\mathbf{a}) \in \mathcal{A}_{\tau_0(k)}$.

Theorem 1. *For a multiple decision problem with loss function $L \in \mathcal{L}$ and prior probability mass function π on $\boldsymbol{\theta}$, let $k^* : \mathcal{X} \to \mathcal{M}$ be defined via*

$$k^*(\mathbf{x}) = \arg\min_{k \in \mathcal{M}} H(k, \mathbf{x}), \quad \mathbf{x} \in \mathcal{X},$$

where $H : \mathcal{M} \times \mathcal{X} \to \mathbb{R}$ is defined in (11). Then the BMDF is of the form (7) with $a^*(\mathbf{x})$ satisfying

$$\mathbf{g}_0(\mathbf{a}^*(\mathbf{x})) = \Big(I\{r_m(k^*(\mathbf{x}), \mathbf{x}) \le \tau_0(k^*(\mathbf{x}))\}, \quad m = 1, 2, \ldots, M \Big). \tag{12}$$

The searching order for obtaining the Bayes optimal action associated with the BMDF is no more than $O(M^2 \log M)$.

Proof. Associated with the loss function L, $\tilde{L}$ defined in equation (5) has a specific form given by

$$\tilde{L}(\mathbf{a}, \mathbf{x}) = \sum_{j=0}^{1} C_j [\alpha_j(\mathbf{a}^{\mathsf{T}}\mathbf{1}) \mathbf{g}_j(\mathbf{a})]^{\mathsf{T}} E_{\boldsymbol{\theta}|\mathbf{X}=\mathbf{x}} [\beta_j(\boldsymbol{\theta}^{\mathsf{T}}\mathbf{1}) \mathbf{h}_j(\boldsymbol{\theta})].$$

Restricting $\mathbf{a}$ on $\mathcal{A}_k$,

$$\begin{aligned} \tilde{L}(\mathbf{a}, \mathbf{x}) &= \mathbf{g}_0(\mathbf{a})^{\mathsf{T}} \mathbf{d}_0(k, \mathbf{x}) + \mathbf{g}_1(\mathbf{a})^{\mathsf{T}} \mathbf{d}_1(k, \mathbf{x}) \\ &= \mathbf{a}_0^{\mathsf{T}} \mathbf{d}_1(k, \mathbf{x}) + \mathbf{g}_0(\mathbf{a})^{\mathsf{T}} [\mathbf{d}_0(k, \mathbf{x}) - A_1 \mathbf{d}_1(k, \mathbf{x})] \\ &= \mathbf{a}_0^{\mathsf{T}} \mathbf{d}_1(k, \mathbf{x}) + \mathbf{g}_0(\mathbf{a})^{\mathsf{T}} \mathbf{e}(k, \mathbf{x}), \end{aligned}$$

where $d_0(k, \mathbf{x})$, $d_1(k, \mathbf{x})$, and $\mathbf{e}(k, \mathbf{x})$ are as defined in (8)-(10).

Since for $\mathbf{a} \in \mathcal{A}_k$, $\mathbf{g}_0(\mathbf{a})^{\mathsf{T}}\mathbf{1} = \tau_0(k)$, the optimal action on $\mathcal{A}_k$, denoted by $\mathbf{a}_k^*(\mathbf{x})$, which minimizes $\tilde{L}(\mathbf{a}, \mathbf{x})$ for $\mathbf{a} \in \mathcal{A}_k$, therefore satisfies

$$\begin{aligned} &\mathbf{g}_0(\mathbf{a}_k^*(\mathbf{x})) \\ &\quad = \Big(I\{r_1(k, \mathbf{x}) \le \tau_0(k)\}, I\{r_2(k, \mathbf{x}) \le \tau_0(k)\}, \ldots, I\{r_M(k, \mathbf{x}) \le \tau_0(k)\} \Big)^{\mathsf{T}}, \end{aligned}$$

where we recall that, for $m = 1, 2, \ldots, M$, $r_m(k, \mathbf{x})$ is the rank of $e_m(k, \mathbf{x})$ among the elements of $\mathbf{e}(k, \mathbf{x})$, $m = 1, 2, \ldots, M$. Thus,

$$\tilde{L}(\mathbf{a}_k^*(\mathbf{x}), \mathbf{x}) = \mathbf{a}_0^{\mathsf{T}} \mathbf{d}_1(k, \mathbf{x}) + \sum_{m=1}^{M} I(r_m(k, \mathbf{x}) \le \tau_0(k)) e_m(k, \mathbf{x}),$$

which equals the function $H(k, \mathbf{x})$. Therefore, for the $k^*(\mathbf{x})$ in the statement of Theorem 1, $\mathbf{a}_{k^*(\mathbf{x})}^*(\mathbf{x})$ minimizes $\tilde{L}(\mathbf{a}, \mathbf{x})$ over all actions $\mathbf{a} \in \mathcal{A}$. The optimal action, given $\mathbf{X} = \mathbf{x}$, is therefore $\mathbf{a}^*(\mathbf{x}) = \mathbf{a}_{k^*(\mathbf{x})}^*(\mathbf{x})$, which satisfies (12).

For the computational order of the algorithm, observe that for $k \in \mathcal{M}$, in order to find $\mathbf{a}_k^*(\mathbf{x})$, it is only necessary to know which are the $\tau_0(k)$ smallest among all the elements of $\mathbf{e}(k, \mathbf{x})$, and the order is bounded by $O(M + \tau_0(k) \log M)$ [12].

Upon obtaining $\mathbf{a}_k^*(\mathbf{x})$, one only needs to search the minimum of $H(k, \mathbf{x})$ for $k \in \mathcal{M}$. Therefore, the searching order is bounded by $\sum_{k=1}^{M} O(M+\tau_0(k)\log M)$. The worst-case scenario is when $\tau_0(k) \equiv M$, $k = 0, 1, \ldots, M$, which leads to an upper bound of $O(M^2 \log M)$. □

Observe that the searching order of $O(M^2 \log M)$ is a considerable improvement over $O(2^M)$. This is due to the special form of the loss function and the nature of the parameter space, action space, and multiple decision function space. In a lot of cases, including the specific pairs of loss functions discussed in Section 4, the searching order may still be lower than $O(M^2 \log M)$.

Observe also that in the BMDF described in Theorem 1, we need to obtain the posterior expectation of form $E_{\boldsymbol{\theta}|\mathbf{X}=\mathbf{x}}[\beta(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{1})\mathbf{h}(\boldsymbol{\theta})]$. Recall that X_m, given $\boldsymbol{\theta}$, has the marginal distribution $Q_{m\theta_m}$. We assume for now that the X_ms, given $\boldsymbol{\theta}$, are independent for $m = 1, 2, \ldots, M$. In general, we may also model the X_ms to be dependent as will be discussed in Section 5, in which case the computation of the posterior expectation will be discussed in Section 6. Denote the density of $Q_{m\theta_m}$ by $q_{m\theta_m}$. Suppose an independent prior distribution of the form described in (2) is used. Then, the posterior distribution of $\boldsymbol{\theta}$ has the θ_ms also independent, with

$$\pi_m(\theta_m|\mathbf{x}) = \frac{\pi_{m\theta_m} q_{m\theta_m}(x_m)}{\pi_{m0} q_{m0}(x_m) + \pi_{m1} q_{m1}(x_m)}, m = 1, 2, \ldots, M.$$

Therefore,

$$E_{\boldsymbol{\theta}|\mathbf{x}}[\beta(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{1})\mathbf{h}(\boldsymbol{\theta})] = \sum_{\boldsymbol{\theta}\in\Theta} \beta(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{1})\mathbf{h}(\boldsymbol{\theta}) \prod_{m=1}^{M} \frac{\pi_{m\theta_m} q_{m\theta_m}(x_m)}{\pi_{m0} q_{m0}(x_m) + \pi_{m1} q_{m1}(x_m)}.$$

In general, this is a sum of $|\Theta| = 2^M$ terms. A particular case is $E(\boldsymbol{\theta}|\mathbf{x}) = (E(\theta_1|\mathbf{x}), E(\theta_2|\mathbf{x}), \ldots, E(\theta_M|\mathbf{x}))$ where, for $m = 1, 2, \ldots, M$,

$$E(\theta_m|\mathbf{x}) = P(\theta_m = 1|\mathbf{x}) = \frac{\pi_{m1} q_{m1}(x_m)}{\pi_{m0} q_{m0}(x_m) + \pi_{m1} q_{m1}(x_m)}.$$

An important aspect to note is that in general, each component of the posterior expectation $E_{\boldsymbol{\theta}|\mathbf{x}}[\beta(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{1})\mathbf{h}(\boldsymbol{\theta})]$ is needed to obtain the BMDF, so that each component of δ^* may depend on *all* components of $\mathbf{X}$. This makes the BMDF a *compound* decision function [23]. In essence the decision for the mth component borrows information from the other components, or as mentioned in [8], the decision makes use of direct evidence from the mth component of the data as well as indirect evidence from the other components.

3.2. BMDF in composite hypotheses

Let π be a prior density function on the enlarged parameter space Θ° (see (3) on page 1278). Then the Bayes risk of a decision function $\delta \in \mathcal{D}$ is given by

$$r_\pi(\delta) = E_{(\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\xi})} E_{\mathbf{X}|(\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\xi})} L(\delta(\mathbf{X}), \boldsymbol{\theta}) = E_{\mathbf{X}} E_{\boldsymbol{\theta}|\mathbf{X}} L(\delta(\mathbf{X}), \boldsymbol{\theta}).$$

Observe that the final form of the Bayes risk is exactly the same as in the simple-vs-simple hypotheses setting. This implies that the results in Theorem 1 apply directly. However, since the parameter space where the prior distribution is defined is enlarged, the posterior expectation $E_{\boldsymbol{\theta}|\mathbf{x}}[L(\delta(\mathbf{x}),\boldsymbol{\theta})]$ in the Bayes risk, or $E_{\boldsymbol{\theta}|\mathbf{x}}[\beta(\boldsymbol{\theta}^{\mathsf{T}}\mathbf{1})\mathbf{h}(\boldsymbol{\theta})]$ in Theorem 1 is now taken with respect to the *marginal* posterior distribution of $\boldsymbol{\theta}$, given $\mathbf{X}=\mathbf{x}$.

Assume the X_ms, given θ, are independent, for $m=1,2,\ldots,M$. Denote the density of $Q_m(\cdot\,;\gamma_m,\xi_m)$, under $H_{m\theta_m}$, by $q_{m\theta_m}(\cdot\,;\gamma_m,\xi_m)$. If an independent prior of the form in (4) is assigned, then the marginal posterior distribution of $\boldsymbol{\theta}$ also makes θ_ms independent, with

$$\pi_m(\theta_m|\mathbf{x})=\frac{\pi_{m\theta_m}\tilde{q}_{m\theta_m}(x_m)}{\pi_{m0}\tilde{q}_{m0}(x_m)+\pi_{m1}\tilde{q}_{m1}(x_m)},$$

where

$$\tilde{q}_{m\theta_m}(x_m)=\int_{\Gamma_{m\theta_m}}\int_{\Xi_m}p_{m\theta_m}(\gamma_m,\xi_m)q_{m\theta_m}(x_m;\gamma_m,\xi_m)\mathrm{d}\xi_m\mathrm{d}\gamma_m. \tag{13}$$

Therefore,

$$E_{\boldsymbol{\theta}|\mathbf{x}}[\beta(\boldsymbol{\theta}^{\mathsf{T}}\mathbf{1})\mathbf{h}(\boldsymbol{\theta})]=\sum_{\boldsymbol{\theta}\in\Theta}\beta(\boldsymbol{\theta}^{\mathsf{T}}\mathbf{1})\mathbf{h}(\boldsymbol{\theta})\prod_{m=1}^{M}\frac{\pi_{m\theta_m}\tilde{q}_{m\theta_m}(x_m)}{\pi_{m0}\tilde{q}_{m0}(x_m)+\pi_{m1}\tilde{q}_{m1}(x_m)}.$$

In particular, $E(\boldsymbol{\theta}|\mathbf{x})=(E(\theta_1|\mathbf{x}),E(\theta_2|\mathbf{x}),\ldots,E(\theta_M|\mathbf{x}))$ where, for $m=1,2,\ldots,M$,

$$E(\theta_m|\mathbf{x})=\frac{\pi_{m1}\tilde{q}_{m1}(x_m)}{\pi_{m0}\tilde{q}_{m0}(x_m)+\pi_{m1}\tilde{q}_{m1}(x_m)}.$$

Notice that the integral in $\tilde{q}_{m\theta_m}(x_m)$ may not be in closed form. Thus Monte Carlo techniques may be needed even in this independent setting to approximately compute $\tilde{q}_{m\theta_m}(x_m)$ and hence $E(\theta_m|\mathbf{x})$. Similarly to the simple-vs-simple hypotheses setting, the BMDF in this composite hypotheses setting is compound.

4. Loss functions: Special cases

4.1. Combination of FP and FN loss functions

Consider the loss function

$$L_{(FP,FN)}(\mathbf{a},\boldsymbol{\theta})=C_0L_{FP}(\mathbf{a},\boldsymbol{\theta})+C_1L_{FN}(\mathbf{a},\boldsymbol{\theta})=C_0\frac{\mathbf{a}^{\mathsf{T}}(\mathbf{1}-\boldsymbol{\theta})}{M}+C_1\frac{(\mathbf{1}-\mathbf{a})^{\mathsf{T}}\boldsymbol{\theta}}{M},$$

where L_{FP} and L_{FN} are the false positive proportion and false negative proportion.

This is the loss function that has been studied extensively in previous papers. It is clear that the optimal action minimizing $\tilde{L}_{(FP,FN)}$ is $\mathbf{a}^*_{(FP,FN)}(\mathbf{x}) = ((\mathbf{a}^*_{(FP,FN)}(\mathbf{x}))_m, m = 1, 2, \ldots, M)$, where, for $m = 1, 2, \ldots, M$,

$$(a^*_{(FP,FN)}(\mathbf{x}))_m = I\left(\frac{E(\theta_m|\mathbf{x})}{1 - E(\theta_m|\mathbf{x})} > \frac{C_0}{C_1}\right) = I\left(E(\theta_m|\mathbf{x}) > \frac{C_0}{C_0 + C_1}\right).$$

The corresponding BMDF δ^* is such that $\delta^*(\mathbf{X}) = a^*(\mathbf{X})$. This BMDF is of intuitive form in that the decision on each component is based only on $E(\theta_m|\mathbf{x}) = P(\theta_m = 1|\mathbf{x})$ and the threshold is just $C_0/(C_0 + C_1)$, though it should be pointed out that $E(\theta_m|\mathbf{x})$ may depend on all the data. Note that, when $M = 1$, this is just the Bayes test corresponding to a C_0/C_1-loss function [5].

4.2. Combination of FDP and FNP loss function

Consider the loss function

$$\begin{aligned} L_{(FDP,FNP)}(\mathbf{a}, \boldsymbol{\theta}) &= C_0 L_{FDP}(\mathbf{a}, \boldsymbol{\theta}) + C_1 L_{FNP}(\mathbf{a}, \boldsymbol{\theta}) \\ &= C_0 \frac{\mathbf{a}^{\mathsf{T}}(\mathbf{1} - \boldsymbol{\theta})}{(\mathbf{a}^{\mathsf{T}}\mathbf{1}) \vee 1} + C_1 \frac{(\mathbf{1} - \mathbf{a})^{\mathsf{T}}\boldsymbol{\theta}}{((\mathbf{1} - \mathbf{a})^{\mathsf{T}}\mathbf{1}) \vee 1}, \end{aligned}$$

where L_{FDP} and L_{FNP} are the false discovery proportion and the false non-discovery proportion. Note that for this $L \in \mathcal{L}$, $h_0(\boldsymbol{\theta}) = \mathbf{1} - \boldsymbol{\theta}$, $h_1(\boldsymbol{\theta}) = \boldsymbol{\theta}$, $\mathbf{g}_0(\mathbf{a}) = \mathbf{a}$, $\mathbf{a}_0 = \mathbf{1}$, $A_1 = 1$, and $\tau_0(k) = k$, $\alpha_0(k) = 1/(k \vee 1)$, and $\alpha_1(k) = 1/[(M - k) \vee 1]$. This loss function has also been studied previously.

Let $(\phi_{(1)}(\mathbf{x}), \phi_{(2)}(\mathbf{x}), \ldots, \phi_{(M)}(\mathbf{x}))$ denote the ordered vector associated with $E(\boldsymbol{\theta}|\mathbf{x})$. Then, in Theorem 1, we have

$$\begin{aligned} H(k, \mathbf{x}) &= C_1 \frac{\sum_{m=1}^{M} \phi_m(\mathbf{x})}{(M - k) \vee 1} \\ &\quad + \left[C_0 \frac{\sum_{i=1}^{k} (1 - \phi_{(M-i+1)}(\mathbf{x}))}{k \vee 1} - C_1 \frac{\sum_{i=1}^{k} \phi_{(M-i+1)}(\mathbf{x})}{(M - k) \vee 1}\right] \\ &= C_0 \frac{\sum_{i=1}^{k} (1 - \phi_{(M-i+1)}(\mathbf{x}))}{k \vee 1} + C_1 \frac{\sum_{i=k+1}^{M} \phi_{(M-i+1)}(\mathbf{x})}{(M - k) \vee 1}. \end{aligned}$$

Letting $k^*(\mathbf{x}) = \arg\min_{k \in \mathcal{M}} H(k, \mathbf{x})$, by Theorem 1 the optimal action is

$$\begin{aligned} \mathbf{a}^*_{(FDP,FNP)}(\mathbf{x}) &= (I(r_m(k^*(\mathbf{x}), \mathbf{x}) \le k^*(\mathbf{x})),\ m = 1, 2, \ldots, M) \\ &= (I(\operatorname{rank}(E(\theta_m|\mathbf{x})) \ge M - k^*(\mathbf{x}) + 1), m = 1, 2, \ldots, M) \\ &= (I(M + 1 - \operatorname{rank}(P(\theta_m = 1|\mathbf{x})) \le k^*(\mathbf{x})), m = 1, 2, \ldots, M). \end{aligned}$$

Notice that for all $k = 1, 2, \ldots, M - 1$, $H(k, \mathbf{x})$ depends only on the ordered vector of $E(\boldsymbol{\theta}|\mathbf{x})$, which means that in order to select $k^*(\mathbf{x})$ we only need to sort $E(\boldsymbol{\theta}|\mathbf{x})$ once. Also, the optimal action only requires the rank vector of $E(\boldsymbol{\theta}|\mathbf{x})$ after $k^*(\mathbf{x})$ has been obtained. Therefore, the searching order is reduced to $O(M \log M)$.

For this case, the posterior means of the θ_m's are still the main basis of choosing the optimal actions, but in contrast to the previous case, the decision at any particular component depends on *all* posterior means. One may initially conclude that the Bayes multiple decision function does not depend on the magnitudes of the $E(\theta_m|\mathbf{x})$, but rather only on their relative ranks. However, this is *not* the case since their magnitudes are actually needed to determine $k^*(\mathbf{x})$.

Observe also that this Bayes optimal action has similarities to the Benjamini-Hochberg (BH) procedure [1], since it can be considered as a step-up procedure which compares ordered posterior probabilities that $\theta_m = 1$ to a data-dependent critical cutoff value, whereas the BH procedure also compares ordered p-values to a data-dependent critical cutoff value. Analogous observations of the similarities of Bayes procedures to the BH procedure have also been pointed out in [14] and [4].

4.3. Combination of FDP and MDP loss functions

Consider the loss function

$$\begin{aligned} L_{(FDP,MDP)}(\mathbf{a},\boldsymbol{\theta}) &= C_0 L_{FDP}(\mathbf{a},\boldsymbol{\theta}) + C_1 L_{MDP}(\mathbf{a},\boldsymbol{\theta}) \\ &= C_0 \frac{\mathbf{a}^{\mathrm{T}}(\mathbf{1}-\boldsymbol{\theta})}{(\mathbf{a}^{\mathrm{T}}\mathbf{1}) \vee 1} + C_1 \frac{(\mathbf{1}-\mathbf{a})^{\mathrm{T}}\boldsymbol{\theta}}{(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{1}) \vee 1}, \end{aligned}$$

where L_{FDP} and L_{MDP} are the false discovery proportion and the missed discovery proportion. Analogously for the pair of (FDP, FNP), $h_0(\boldsymbol{\theta}) = \mathbf{1} - \boldsymbol{\theta}$, $h_1(\boldsymbol{\theta}) = \boldsymbol{\theta}$, $\mathbf{g}_0(\mathbf{a}) = \mathbf{a}$, $\mathbf{a}_0 = \mathbf{1}$, $A_1 = 1$, and $\tau_0(k) = k$, $\alpha_0(k) = 1/(k \vee 1)$, and $\alpha_1(k) = 1$. Let

$$\mathbf{e}(k,\mathbf{x}) = \frac{C_0}{k} E(\mathbf{1}-\boldsymbol{\theta}|\mathbf{x}) - C_1 E\left(\frac{\boldsymbol{\theta}}{(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{1}) \vee \mathbf{1}} \middle| \mathbf{x} \right),$$

and denote by $(e_{(1)}(k,\mathbf{x}), e_{(2)}(k,\mathbf{x}), \ldots, e_{(M)}(k,\mathbf{x}))$ the ordered vector of $\mathbf{e}(k,\mathbf{x})$. Then

$$H(k,\mathbf{x}) = C_1 \mathbf{1}^{\mathrm{T}} E\left(\frac{\boldsymbol{\theta}}{(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{1}) \vee \mathbf{1}} | \mathbf{x} \right) + \sum_{i=1}^{k} e_{(i)}(k,\mathbf{x})$$

and

$$k^*(\mathbf{x}) = \begin{cases} 0 & \text{if } \min\limits_{k \in \mathcal{M}\setminus\{0\}} \sum\limits_{i=1}^{k} e_{(i)}(k,\mathbf{x}) > 0 \\ \arg\min\limits_{k \in \mathcal{M}\setminus\{0\}} \sum\limits_{i=1}^{k} e_{(i)}(k,\mathbf{x}) & \text{otherwise} \end{cases}.$$

By Theorem 1, the optimal Bayes action is

$$\mathbf{a}^*_{(FDP,MDP)}(\mathbf{x}) = (I(\mathrm{rank}(e_m(k^*(\mathbf{x}),\mathbf{x})) \le k^*(\mathbf{x}),\ m = 1,2,\ldots,M).$$

Observe that $\mathbf{a}^*(\mathbf{x})$ and $k^*(\mathbf{x})$ depend on the values and ranks of $\tilde{e}_m(k,\mathbf{x})$ for $k \in \mathcal{M}$. The searching order in this case is $O(M^2 \log M)$. It can be shown that when the posterior probability distribution of $\boldsymbol{\theta}$ specifies independent components, the searching order is reduced to $O(M \log M)$.

4.4. Discussion of choice of loss functions

As shown in this section, different loss functions lead to different forms of the BMDF, that is, given the same data $\mathbf{x}$, it is possible that different Bayes optimal actions will arise by varying the loss functions. The loss function pair of (FP, FN) is the most straight-forward as it is equivalent to counting the number of errors. Some theoretical studies, such as [3], have shown that the BH false discovery rate controlling procedure possesses an asymptotic Bayes optimality under this loss function. However, if the goal is to minimize the Bayes risk using a loss function induced by the FDP, then the loss function pair of (FDP, FNP) or (FDP, MDP) will be more appropriate since they are both weighted sums of FDP and a specific type II error loss function. In addition, as shown in this section, the BMDF associated with these two loss functions have a flavor similar to the BH procedure. Computationally, these two have the same searching order once the necessary posterior expectations have been calculated. However, the BMDF associated with (FDP, MDP) requires the posterior expectation of $\boldsymbol{\theta}/[(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{1}) \vee \mathbf{1}]$, which involves the entire vector $\boldsymbol{\theta}$. As we will discuss in the sections dealing with sequential Monte Carlo procedures, the computation of this posterior expectation is relatively more extensive and less stable. Thus, in practice, an adjusted version of MDP, the $AMDP$ indicated in Table 1, is recommended. Even with the use of the $AMDP$, the computational cost is still higher compared to the case where the (FDP, FNP) pair is used.

5. Dependent data structure

In Section 3, formulas of the posterior expectations are given under the assumption that the X_m's, $m = 1, 2, \ldots, M$, are independent. However, in various situations this assumption may not be realistic. Recall in section 2.1 that $Q_{\boldsymbol{\theta}}(\cdot; \gamma_{\boldsymbol{\theta}})$ is the probability measure of $\mathbf{X}$, given $(\boldsymbol{\theta}, \gamma_{\boldsymbol{\theta}})$, with $\boldsymbol{\vartheta}(Q_{\boldsymbol{\theta}}(\cdot; \gamma_{\boldsymbol{\theta}})) = \boldsymbol{\theta}$. In this section, we describe $Q_{\boldsymbol{\theta}}(\cdot; \gamma_{\boldsymbol{\theta}})$ that allows for dependencies among the X_ms in $\mathbf{X}$.

In the simple hypotheses setting, the goal is to specify $Q_{\boldsymbol{\theta}}$, the joint probability distribution of $\mathbf{X}$ given $\boldsymbol{\theta}$, such that the marginal probability distribution of X_m is $Q_m = Q_{m\theta_m}$. Let $\mathcal{M}_0(\boldsymbol{\theta}) = \{m \in \{1, 2, \ldots, M\} \colon \theta_m = 0\}$ and $\mathcal{M}_1(\boldsymbol{\theta}) = \{m \in \{1, 2, \ldots, M\} \colon \theta_m = 1\}$. Assume that $X_{\mathcal{M}_0(\boldsymbol{\theta})} \equiv \{X_m \colon m \in \mathcal{M}_0(\boldsymbol{\theta})\}$ is a collection of independent random vectors, and $X_{\mathcal{M}_1(\boldsymbol{\theta})} \equiv \{X_m \colon m \in \mathcal{M}_1(\boldsymbol{\theta})\}$ is a collection of *possibly* dependent vectors, and the collections $X_{\mathcal{M}_0(\boldsymbol{\theta})}$ and $X_{\mathcal{M}_1(\boldsymbol{\theta})}$ are independent of each other. Borrowing from survival analysis ideas, we assume that the dependence structure of the collection $X_{\mathcal{M}_1(\boldsymbol{\theta})}$ is induced by an unobserved frailty variable Z. We assume that, conditionally on $Z = z$, $X_{\mathcal{M}_1(\boldsymbol{\theta})}$ are independent. Specifically, assume that

$$Q_{\boldsymbol{\theta}}\left(\bigcap_{m \in \mathcal{M}_1(\boldsymbol{\theta})} [X_m \in B_m] \,\middle|\, Z = z\right) = \prod_{m \in \mathcal{M}_1(\boldsymbol{\theta})} [\breve{Q}_m(B_m)]^z,$$

for all $B_m \in \mathcal{B}_m$, $m \in \mathcal{M}_1(\boldsymbol{\theta})$, where $\breve{Q}_m$'s are some distributions on $(\mathcal{X}_m, \mathcal{B}_m)$'s, and the frailty $Z \in \mathcal{Z}$ is assumed to have a distribution G. Such frailty models

have been widely used in survival analysis, where the frailty variable is used to account for hidden heterogeneity. We use this idea in the multiple decision setting, where the frailty variable is used to model the possible common feature for genes in the true alternative collection, for example, those genes that are differentially expressed across treatment groups.

The joint distribution of $X_{\mathcal{M}_1(\boldsymbol{\theta})}$ is

$$Q_{\boldsymbol{\theta}}\left(\bigcap_{m\in\mathcal{M}_1(\boldsymbol{\theta})}[X_m\in B_m]\right)=\int_{\mathcal{Z}}\prod_{m\in\mathcal{M}_1(\boldsymbol{\theta})}[\breve{Q}_m(B_m)]^z G(\mathrm{d}z), \tag{14}$$

for all $B_m\in\mathcal{B}_m$, $m\in\mathcal{M}_1(\boldsymbol{\theta})$. Recall that in the simple-vs-simple multiple hypotheses testing setting, under H_{m1}, $X_m\sim Q_{m1}$ marginally. So the distributions $\breve{Q}_m$, $m\in\mathcal{M}_1(\boldsymbol{\theta})$, should be such that these conditions are satisfied. Let $\mathcal{L}_G$ be the Laplace transform of the distribution function G, that is,

$$\mathcal{L}_G(u)=\int_{\mathcal{Z}}e^{-uz}G(\mathrm{d}z),\ \forall u\in\mathbb{R}.$$

Let $M_1\equiv|\mathcal{M}_1(\boldsymbol{\theta})|$. The following result gives the joint distribution of the dependent collection $X_{\mathcal{M}_1(\boldsymbol{\theta})}$ in terms of the collection of the marginal distributions $\{Q_{m1}:m\in\mathcal{M}_1(\boldsymbol{\theta})\}$ under this frailty-based model.

Proposition 1. *The frailty-based model described in (14) is an M_1-dimensional Archimedean copula C_G such that*

$$Q_{\boldsymbol{\theta}}\left(\bigcap_{m\in\mathcal{M}_1(\boldsymbol{\theta})}[X_m\in B_m]\right)=C_G(Q_{m1}(B_m),m\in\mathcal{M}_1(\boldsymbol{\theta}))$$

for all $B_m\in\mathcal{B}_m$, where $C_G\colon[0,1]^{M_1}\to[0,1]$ is defined via

$$C_G(u_1,u_2,\ldots,u_{M_1})=\mathcal{L}_G\left(\sum_{m=1}^{M_1}\mathcal{L}_G^{-1}(u_m)\right).$$

Proof. According to the model, marginally for $m\in\mathcal{M}_1(\boldsymbol{\theta})$ and all $B_m\in\mathcal{B}_m$,

$$\begin{aligned}Q_{m1}(B_m)&=\int_{\mathcal{Z}}\breve{Q}_m(B_m)^z G(\mathrm{d}z)\\&=\int_{\mathcal{Z}}\exp\big\{-z\big\{-\log\breve{Q}_m(B_m)\big\}\big\}G(\mathrm{d}z)\\&=\mathcal{L}_G(-\log\breve{Q}_m(B_m)).\end{aligned}$$

Thus $\breve{Q}_m(B_m)=\exp(-\mathcal{L}_G^{-1}(Q_{m1}(B_m))$. So,

$$Q_{\boldsymbol{\theta}}\left(\bigcap_{m\in\mathcal{M}_1(\boldsymbol{\theta})}[X_m\in B_m]\right)=\int_{\mathcal{Z}}\prod_{m\in\mathcal{M}_1(\boldsymbol{\theta})}(\exp(-\mathcal{L}_G^{-1}(Q_{m1}(B_m))))^z G(\mathrm{d}z)$$

$$= \int_{\mathcal{Z}} \exp\left(-\left[\sum_{m\in\mathcal{M}_1(\boldsymbol{\theta})} \mathcal{L}_G^{-1}(Q_{m1}(B_m))\right]\cdot z\right) G(\mathrm{d}z)$$
$$= \mathcal{L}_G\left(\sum_{m\in\mathcal{M}_1(\boldsymbol{\theta})} \mathcal{L}_G^{-1}(Q_{m1}(B_m))\right).$$

To show C_G is an Archimedean copula, it is sufficient to show that the function $\mathcal{L}_G^{-1}$ is a *strict generator* of a copula, which is that it is a continuous strictly decreasing convex function from $[0,1]$ to $[0,\infty]$ with $\mathcal{L}_G^{-1}(1)=0$ and $\mathcal{L}_G^{-1}(0)=\infty$ [16]. But these are straight-forward to verify using properties of the Laplace transform. □

Thus, a frailty-induced dependent full data model is given by

$$Q_{\boldsymbol{\theta}}\left(\bigcap_{m=1}^{M}[X_m\in B_m]\right) = \left(\prod_{m\in\mathcal{M}_0(\boldsymbol{\theta})} Q_{m0}(B_m)\right) C_G\Big[Q_{m1}(B_m), m\in\mathcal{M}_1(\boldsymbol{\theta})\Big], \tag{15}$$

for all $B_m\in\mathcal{B}_m$, $m=1,2,\ldots,M$. Notice that the distribution function G may have nuisance parameters, say, $G(\cdot)=G(\cdot,\boldsymbol{\upsilon})$, where $\boldsymbol{\upsilon}\in\Upsilon$. In this case, to calculate the posterior expectations, a prior on Υ will also be needed.

In the composite hypotheses testing setting, we are to specify $Q_{\boldsymbol{\theta}}(\cdot;\boldsymbol{\gamma},\boldsymbol{\xi})$, the joint probability distribution of $\mathbf{X}$ given $(\boldsymbol{\theta},\boldsymbol{\gamma},\boldsymbol{\xi})$, such that the marginal probability distribution of X_m is $Q_m = Q_{m\theta_m}(\cdot;\gamma_m,\xi_m)$. The result in Proposition 1 is easily extended to get

$$Q_{\boldsymbol{\theta}}\left(\bigcap_{m\in\mathcal{M}_1(\boldsymbol{\theta})}[X_m\in B_m];\boldsymbol{\gamma},\boldsymbol{\xi}\right) = C_G\Big[Q_{m1}(B_m;\gamma_m,\xi_m), m\in\mathcal{M}_1(\boldsymbol{\theta})\Big],$$

and the full data model is given by

$$Q_{\boldsymbol{\theta}}\left(\bigcap_{m=1}^{M}[X_m\in B_m];\boldsymbol{\gamma},\boldsymbol{\xi}\right)$$
$$= \left(\prod_{m\in\mathcal{M}_0(\boldsymbol{\theta})} Q_{m0}(B_m;\gamma_m,\xi_m)\right) C_G\Big[Q_{m1}(B_m;\gamma_m,\xi_m), m\in\mathcal{M}_1(\boldsymbol{\theta})\Big].$$

6. Sequential Monte Carlo

The applicability of the algorithm for finding the Bayes optimal action in Theorem 1 is contingent on an efficient way of calculating the posterior expectations

$$E(H(\boldsymbol{\theta})|\mathbf{X}=\mathbf{x}) = \sum_{\boldsymbol{\theta}\in\Theta} H(\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{x}),$$

where $H(\boldsymbol{\theta})$ takes the form $\beta(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{1})h(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta}|\mathbf{x})$ is the posterior probability mass function of $\boldsymbol{\theta}$, given data $\mathbf{x}$. For example, the specific forms of the H function desired in case of FDP and MDP loss functions are $H(\boldsymbol{\theta}) = \boldsymbol{\theta}$ and $H(\boldsymbol{\theta}) = \boldsymbol{\theta} \big/ \left[(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{1}) \vee 1\right]$. As pointed out in Sections 3 and 5, Monte Carlo integration is needed for approximating the posterior expectations. However, in the regular Importance Sampling (IS) algorithm [19], as M, the dimension of $\boldsymbol{\theta}$, increases, the computational complexity of calculating the weights also increases. So it is important to consider a sequential application of the importance sampling methods [9]. Notice that the index m, which takes values $1, 2, \ldots, M$, does not necessarily represent time or positions in an ordered sequence, and that the proposed dependent data structure is not necessarily a state space model as what is usually the case in sequential importance sampling applications. However, the sequential technique provides a visual solution through $m = 1, 2, \ldots, M$ which deals with the dimensionality and monitors the efficiency of the sampling procedure.

6.1. Simple hypotheses

Let π be a prior probability mass function of $\boldsymbol{\theta}$. Under the dependent data model described in (15), the desired posterior expectation is given by

$$I(\mathbf{x}) = E(H(\boldsymbol{\theta})|\mathbf{x}) = \frac{\sum_{\boldsymbol{\theta}\in\Theta} H(\boldsymbol{\theta})\pi(\boldsymbol{\theta})Q_{\boldsymbol{\theta}}(\mathrm{d}\mathbf{x})}{\sum_{\boldsymbol{\theta}\in\Theta} \pi(\boldsymbol{\theta})Q_{\boldsymbol{\theta}}(\mathrm{d}\mathbf{x})},$$

where

$$\begin{aligned} Q_{\boldsymbol{\theta}}(\mathrm{d}\mathbf{x}) \quad = \quad & \left(\prod_{m\in\mathcal{M}_0(\boldsymbol{\theta})} q_{m0}(x_m)\right) \\ & \times \left(\prod_{m\in\mathcal{M}_1(\boldsymbol{\theta})} q_{m1}(x_m)\right) c_G(Q_{m1}(x_m), m \in \mathcal{M}_1(\boldsymbol{\theta}))\mathrm{d}\mathbf{x}, \end{aligned}$$

with q_{m0} and q_{m1} being the density functions of Q_{m0} and Q_{m1}, respectively, and

$$c_G(u_1, u_2, \ldots, u_{M_1}) = \frac{\partial^{M_1}}{\partial u_1 \partial u_2 \ldots \partial u_{M_1}} C_G(u_1, u_2, \ldots, u_{M_1})$$

is the copula density of C_G. Consider an independent data-adaptive trial probability mass function g of $\boldsymbol{\theta}$ given by $g(\boldsymbol{\theta}) = g(\boldsymbol{\theta}|\mathbf{x}) = \prod_{m=1}^{M} g_m(\theta_m|x_m)$, where, for $m = 1, 2, \ldots, M$, $g_m(\theta_m|x_m)$ is the marginal posterior of θ_m, given $X_m = x_m$. This is a Bernoulli distribution of form

$$g_m(\theta_m|x_m) \propto \pi_m(\theta_m) q_{m0}(x_m)^{1-\theta_m} q_{m1}(x_m)^{\theta_m}, \tag{16}$$

where $\pi_m(\theta_m)$ is the marginal prior probability of θ_m. Denote by $\mathbf{x}_{1:m} = (x_1, x_2, \ldots, x_m)$ and $\boldsymbol{\theta}_{1:m} = (\theta_1, \theta_2, \ldots, \theta_m)$ for $m = 1, 2, \ldots, M$. The joint

prior distribution of $\boldsymbol{\theta}$ can be written as $\pi(\boldsymbol{\theta}) = \prod_{m=1}^{M} \pi_m(\theta_m|\boldsymbol{\theta}_{1:m-1})$, where $\pi_1(\theta_1|\boldsymbol{\theta}_{1:0})$ is the marginal distribution of θ_1, and for $m = 2, 3, \ldots, M$, $\pi_m(\theta_m|\boldsymbol{\theta}_{1:m-1})$ is the probability of θ_m, given $\boldsymbol{\theta}_{1:m-1}$, under the joint prior π. Notice that if π is independent, then $\pi_m(\theta_m|\boldsymbol{\theta}_{1:m-1}) = \pi_m(\theta_m)$. However, dependent prior structures can also be constructed using frailty-induced models similarly to the dependent data structure, in which case $\pi_m(\theta_m|\boldsymbol{\theta}_{1:m-1})$ does not necessarily reduce to $\pi_m(\theta_m)$.

Let $q_{\boldsymbol{\theta}_{1:m}}(\mathbf{x}_{1:m})$ be the marginal density function of $\mathbf{X}_{1:m}$, given $\theta_{1:m}$, under the dependent data structure proposed in (15). Since the dependent structure is induced by a frailty model, we have, for $m = 2, 3, \ldots, M$, $q_{\boldsymbol{\theta}_{1:m}}(\mathbf{x}_{1:m}) = q_{\boldsymbol{\theta}_{1:m}}(x_m|\mathbf{x}_{1:m-1}) \cdot q_{\boldsymbol{\theta}_{1:m-1}}(\mathbf{x}_{1:m-1})$, where

$$\begin{aligned} q_{\boldsymbol{\theta}_{1:m}}(x_m|\mathbf{x}_{1:m-1}) &= \frac{q_{\boldsymbol{\theta}_{1:m}}(\mathbf{x}_{1:m})}{q_{\boldsymbol{\theta}_{1:m-1}}(\mathbf{x}_{1:m-1})} \\ &= \begin{cases} q_{m0}(x_m) & \text{if } \theta_m = 0 \\ q_{m1}(x_m)\dfrac{c_G(Q_{t1}(x_t), t \in \mathcal{M}_1(\boldsymbol{\theta}_{1:m}))}{c_G(Q_{t1}(x_t), t \in \mathcal{M}_1(\boldsymbol{\theta}_{1:m-1}))} & \text{if } \theta_m = 1 \end{cases}. \end{aligned}$$

Thus, the full density of the data has the sequential form

$$q_{\boldsymbol{\theta}}(\mathbf{x}) = \prod_{m=1}^{M} q_{\boldsymbol{\theta}_{1:m}}(x_m|\mathbf{x}_{1:m-1}) \text{ with } q_{\theta_1}(x_1|\mathbf{x}_{1:0}) = 1.$$

So there is a recursive formula for calculating the importance weight function, given by

$$w_m(\boldsymbol{\theta}_{1:m}|\mathbf{x}_{1:m}) = w_{m-1}(\boldsymbol{\theta}_{1:m-1}|\mathbf{x}_{1:m-1})u_m(\boldsymbol{\theta}_{1:m}|\mathbf{x}_{1:m}),$$

where the increment

$$u_m(\boldsymbol{\theta}_{1:m}|\mathbf{x}_{1:m}) = \frac{q_{\boldsymbol{\theta}_{1:m}}(x_m|\mathbf{x}_{1:m-1})\pi(\theta_m|\boldsymbol{\theta}_{1:m-1})}{g_m(\theta_m|x_m)}$$

satisfies

$$\begin{aligned} &u_m(\boldsymbol{\theta}_{1:m}|\mathbf{x}_{1:m}) \\ &\quad \propto \begin{cases} \dfrac{\pi(\theta_m = 0|\boldsymbol{\theta}_{1:m-1})}{\pi_m(\theta_m = 0)} & \text{if } \theta_m = 0 \\ \dfrac{c_G(Q_{t1}(x_t), t \in \mathcal{M}_1(\boldsymbol{\theta}_{1:m}))}{c_G(Q_{t1}(x_t), t \in \mathcal{M}_1(\boldsymbol{\theta}_{1:m-1}))} \cdot \dfrac{\pi(\theta_m = 1|\boldsymbol{\theta}_{1:m-1})}{\pi_m(\theta_m = 1)} & \text{if } \theta_m = 1 \end{cases}. \end{aligned} \tag{17}$$

Note that $w_M(\boldsymbol{\theta}_{1:M}|\mathbf{x}_{1:M}) = w(\theta|\mathbf{x})$, the importance weight function. These results will now enable us to sample particles and calculate the importance weights in a sequential manner.

The fundamental difficulty of SIS is the degeneracy of the weights. For large values of M, the weights $w^{(r)}(\boldsymbol{\theta}|\mathbf{x})$, $r = 1, 2, \ldots, R$, where R is the number of particles, are all close to 0 except for one of them that will be close to 1, which will

eventually result in a poor estimate. A solution to this difficulty is through resampling, or using the so-called bootstrap filter [13], in which, after resampling, all the importance weights are set to $1/R$ so that all particles make important contributions to the MC estimate. But resampling at each $m = 1, 2, \ldots, M$ may be computationally expensive, so we would only resample whenever the empirical effective sample size is too low. Through this procedure, we can make sure that the weights do not diverge.

Algorithm 1. (SMC in Simple Hypotheses)

1. *Fix a large integer R and a threshold $\rho \in (0, 1]$.*
2. *Iterate for $m = 1, 2, \ldots, M$.*
   (a) *For all $r = 1, 2, \ldots, R$, generate $\theta_m^{(r)}$ independently from $g_m(\cdot|x_m)$ in (16), and set $\boldsymbol{\theta}_{1:m}^{(r)} = (\boldsymbol{\theta}_{1:m-1}^{(r)}, \theta_m^{(r)})$.*
   (b) *Compute the increments $u_m(\boldsymbol{\theta}_{1:m}^{(r)}|\mathbf{x}_{1:m})$ given in (17), and the importance weights $w_m^{(r)} \equiv w_m(\boldsymbol{\theta}_{1:m}^{(r)}|\mathbf{x}_{1:m}) = w_{m-1}(\boldsymbol{\theta}_{1:m-1}^{(r)}|\mathbf{x}_{1:m-1})\, u_m(\boldsymbol{\theta}_{1:m}^{(r)}|\mathbf{x}_{1:m})$.*
   (c) *Compute $ESS = R/\sum_{r=1}^{R}(w_m^{(r)})^2$.*
   (d) *If $ESS < \rho R$, normalize the weights, and resample, with replacement, R particles from $\{\boldsymbol{\theta}_{1:m}^{(r)} : r = 1, 2, \ldots, R\}$ according to the normalized weights, and set all the weights to $1/R$.*
3. *$I(\mathbf{x}) = E(H(\boldsymbol{\theta})|\mathbf{x})$ is approximated by $\hat{I}(\mathbf{x}) = \dfrac{\sum_{r=1}^{R} H(\boldsymbol{\theta}_{1:M}^{(r)}) w_M^{(r)}}{\sum_{r=1}^{R} w_M^{(r)}}$.*

6.2. Composite hypotheses

Let π be the prior probability function on the enlarged parameter space Θ° described in (4), and denote the independent Bernoulli marginal prior probability on Θ by π_0. Then, under the dependent data model described in (15), the desired posterior expectation is given by

$$I(\mathbf{x}) = E(H(\boldsymbol{\theta})|\mathbf{x}) = \frac{\sum_{\boldsymbol{\theta}\in\Theta} H(\boldsymbol{\theta})\pi_0(\boldsymbol{\theta})\tilde{Q}_{\boldsymbol{\theta}}(\mathrm{d}\mathbf{x})}{\sum_{\boldsymbol{\theta}\in\Theta} \pi_0(\boldsymbol{\theta})\tilde{Q}_{\boldsymbol{\theta}}(\mathrm{d}\mathbf{x})},$$

where

$$\begin{aligned}\tilde{Q}_{\boldsymbol{\theta}}(\mathrm{d}\mathbf{x}) \quad = \quad & \left(\prod_{m\in\mathcal{M}_0(\boldsymbol{\theta})} \tilde{q}_{m0}(x_m)\right) \\ & \times \left(\prod_{m\in\mathcal{M}_1(\boldsymbol{\theta})} \tilde{q}_{m1}(x_m)\right) c_G(\tilde{Q}_{m1}(x_m), m \in \mathcal{M}_1(\boldsymbol{\theta}))\mathrm{d}\mathbf{x},\end{aligned}$$

with $\tilde{q}_{m0}$ and $\tilde{q}_{m1}$ defined in (13), $\tilde{Q}_{m1}$ is the distribution function of $\tilde{q}_{m1}$, and $c_G(u_1, u_2, \ldots, u_{M_1})$ is the copula density of C_G. Consider a data-adaptive trial density $\tilde{g}$ on the enlarged parameter space, given by

$$\tilde{g}(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) = \tilde{g}(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\xi}|\mathbf{x}) = \prod_{m=1}^{M} \tilde{g}_m(\theta_m, \gamma_m|x_m),$$

where, for $m = 1, 2, \ldots, M$,

$$\begin{aligned}\tilde{g}_m(\theta_m, \gamma_m, \xi_m|x_m) \propto (\pi_{m0}\breve{q}_{m0}(x_m)p_{m0}(\gamma_m, \xi_m))^{1-\theta_m} \\ \times (\pi_{m1}\breve{q}_{m1}(x_m)p_{m1}(\gamma_m, \xi_m))^{\theta_m},\end{aligned}$$

and for $j = 0, 1$, $\breve{q}_{mj}(x_m) = q_{mj}(x_m; \hat{\gamma}_{mj}(x_m), \hat{\xi}_{mj}(x_m))$, where $\hat{\gamma}_{mj}(x_m)$ and $\hat{\xi}_{mj}(x_m)$ are some convenient estimates, for example, maximum likelihood or method-of-moments estimates, of γ_m and ξ_m, under the marginal models $X_m \sim Q_{mj}$. Notice that if $\breve{q}$ is replaced by $\tilde{q}$, $\tilde{g}_m$ would be equal to the marginal posterior of $(\theta_m, \gamma_m, \xi_m)$ given x_m. Since $\breve{q}_{mj}(x_m)$ is an approximation of $\tilde{q}_{mj}(x_m)$, $\tilde{g}_m$ also provides some guidance to the posterior distribution by making use of the data. Write the prior in (4) as $\pi(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\xi}) = \prod_{m=1}^{M} \pi_m(\theta_m, \gamma_m, \xi_m)$, where for $m = 1, 2, \ldots, M$, $\pi_m(\theta_m, \gamma_m, \xi_m) = [\pi_{m0}p_{m0}(\gamma_m, \xi_m)]^{1-\theta_m}[\pi_{m1}p_{m1}(\gamma_m, \xi_m)]^{\theta_m}$. Similarly to the simple-vs-simple hypotheses case, the full density of data has a sequential form given by $q_{\boldsymbol{\theta}}(\mathbf{x}; \boldsymbol{\gamma}, \boldsymbol{\xi}) = \prod_{m=1}^{M} q_{\boldsymbol{\theta}_{1:m}}(x_m|\mathbf{x}_{1:m-1}; \boldsymbol{\gamma}_{1:m}, \boldsymbol{\xi}_{1:m})$, where

$$\begin{aligned}&q_{\boldsymbol{\theta}_{1:m}}(x_m|\mathbf{x}_{1:m-1}; \boldsymbol{\gamma}_{1:m}, \boldsymbol{\xi}_{1:m}) \\ &\quad = \begin{cases} q_{m0}(x_m; \gamma_m, \xi_m) & \text{if } \theta_m = 0 \\ q_{m1}(x_m; \gamma_m, \xi_m)\dfrac{c_G(Q_{t1}(x_t; \gamma_t, \xi_t), t \in \mathcal{M}_1(\boldsymbol{\theta}_{1:m}))}{c_G(Q_{t1}(x_t; \gamma_t, \xi_t), t \in \mathcal{M}_1(\boldsymbol{\theta}_{1:m-1}))} & \text{if } \theta_m = 1 \end{cases}.\end{aligned}$$

So the recursive formula for calculating the importance weight function is

$$\begin{aligned}&w_m(\boldsymbol{\theta}_{1:m}|\mathbf{x}_{1:m}; \boldsymbol{\gamma}_{1:m}, \boldsymbol{\xi}_{1:m}) \\ &\quad = w_{m-1}(\boldsymbol{\theta}_{1:m-1}|\mathbf{x}_{1:m-1}; \boldsymbol{\gamma}_{1:m-1}, \boldsymbol{\xi}_{1:m-1})u_m(\boldsymbol{\theta}_{1:m}|\mathbf{x}_{1:m}; \boldsymbol{\gamma}_{1:m}, \boldsymbol{\xi}_{1:m}),\end{aligned}$$

where the increment satisfies

$$\begin{aligned}&u_m(\boldsymbol{\theta}_{1:m}|\mathbf{x}_{1:m}; \boldsymbol{\gamma}_{1:m-1}, \boldsymbol{\xi}_{1:m-1}) \\ &\quad \propto \begin{cases} \dfrac{q_{m0}(x_m; \gamma_m, \xi_m)}{\breve{q}_{m0}(x_m)} & \text{if } \theta_m = 0 \\ \dfrac{q_{m1}(x_m; \gamma_m, \xi_m)}{\breve{q}_{m1}(x_m)} \cdot \dfrac{C_G(q_{t1}(x_t; \gamma_t, \xi_t), t \in \mathcal{M}_1(\boldsymbol{\theta}_{1:m}))}{C_G(q_{t1}(x_t; \gamma_t, \xi_t), t \in \mathcal{M}_1(\boldsymbol{\theta}_{1:m-1}))} & \text{if } \theta_m = 1 \end{cases}. \quad (18)\end{aligned}$$

Therefore, the SIS algorithm is very similar to that for the simple-vs-simple hypotheses case with the increment replaced by the expression in (18).

TABLE 2
Average of empirical FDP, FNP and MDP in 1000 simulations in a composite multiple testing problem with independent Gaussian observations. True prior parameters are $\pi = 0.5$ *and* $\sigma = 4$

π^*	σ^*	λ	Loss Functions	(FP, FN)	(FDP, FNP)	(FDP,MDP) Exact	(FDP,MDP) Approx.	BH
0.5	4	1	$F\hat{D}P$	0.115	0.052	0.184	0.112	0.025
			$F\hat{N}P$	0.260	0.278	0.211	0.253	0.309
			$M\hat{D}P$	0.323	0.422	0.230	0.311	0.450
		2	$F\hat{D}P$	0.051	0.010	0.035	0.026	0.025
			$F\hat{N}P$	0.290	0.324	0.297	0.303	0.309
			$M\hat{D}P$	0.388	0.484	0.412	0.428	0.450
0.7	10	1	$F\hat{D}P$	0.098	0.049	0.165	0.100	0.027
			$F\hat{N}P$	0.270	0.287	0.225	0.263	0.304
			$M\hat{D}P$	0.332	0.424	0.248	0.324	0.448
		2	$F\hat{D}P$	0.063	0.017	0.042	0.035	0.027
			$F\hat{N}P$	0.285	0.317	0.290	0.297	0.304
			$M\hat{D}P$	0.390	0.483	0.410	0.428	0.448

7. Simulations and a data example

7.1. Composite alternatives with independent Gaussian observations

Assume X_m are independent $N(\mu_m, 1)$, $m = 1, 2, \ldots, M$. Consider testing H_{m0} : $\mu_m = 0$ versus $H_{m1} : \mu_m \neq 0$, for $m = 1, 2, \ldots, M$. Let $\theta_m = I(\mu_m \neq 0)$. Consider the independent prior $\mu_m \sim (1-\pi)I(\mu_m = 0) + \pi\phi(\mu_m; 0, \sigma^2)$, where π is the fixed prior probability of the alternative hypotheses, $\sigma > 0$ is fixed, and $\phi(\cdot; \mu, \sigma^2)$ denotes the density function of a normal distribution with mean μ and variance σ^2. Then the posterior means are

$$E(\theta_m|X_m) = \frac{\pi\phi(X_m; 0, 1+\sigma^2)}{(1-\pi)\phi(X_m; 0, 1) + \pi\phi(X_m; 0, 1+\sigma^2)}.$$

We performed 1000 simulations with $M = 12$, true $\sigma = 4$, and correct proportion of alternatives equal to $\pi = 0.5$ for all three procedures. For both correct prior parameters with $\pi^* = 0.5$, $\sigma^* = 4$ and misspecified prior parameters with $\pi^* = 0.7$, $\sigma^* = 10$, the empirical FDP, FNP, and MDP were calculated for $C_0/C_1 \in \{1, 2\}$. The results in Table 2 compare the BMDF associated with the three pairs of loss functions to the Benjamini and Hochberg (BH) procedure [1] with a false discovery rate (FDR) threshold of 0.05. To find the BMDF associated with the (FDP, MDP) pair of loss functions, the posterior expectation $E(\boldsymbol{\theta}/(\boldsymbol{\theta}^{\mathrm{T}}\mathbf{1} \vee 1)|\mathbf{x})$ was calculated *exactly* by using a recursive formula and *approximately* by a Monte Carlo approximation. Note that the use of BH procedure in the simulation study is to enable comparison with the most commonly-used, albeit frequentist, multiple testing procedure. However, it is worth mentioning that it is not totally fair to compare the BMDF to the BH procedure because they are designed under different criteria. In Table 2, the empirical risks of the three BMDFs are comparable to those of the BH procedure when the cost ratio C_0/C_1 is relatively

TABLE 3
Results in one simulation replicate in a multiple testing problem with exponential lifetimes

Loss Functions Pair	(FP, FN)		(FDP,FNP)		(FDP,MDP)	
Actions	Nulls	Alts	Nulls	Alts	Nulls	Alts
0: Accepts	334	4	331	4	338	6
1: Rejects	6	156	9	156	2	154

large (= 2). With a smaller cost ratio (= 1), the empirical FDP becomes larger and the empirical FNP and MDP become smaller since, with a smaller cost, the BMDF sacrifices larger Type I error probabilities to achieve an optimal combined risks. With misspecified prior parameters, the combined empirical risks stay almost the same as those with the correctly specified prior parameters, but the Type I error FDP becomes smaller and the Type II errors FNP and MDP become larger. This is because we misspecified a higher prior probability for the alternatives than the true prior probability, which results in more discoveries. Finally, the exact and approximately-calculated posterior expectations result in similar BMDFs associated with the (FDP, MDP) pair of loss functions in terms of empirical risks. Since the exact calculation is computationally expensive for large M, in the following illustration we utilized the Monte Carlo approximated posterior expectations.

7.2. Simple-vs-simple with dependent exponential observations

Consider a situation where, for all m, $X_{m1}, X_{m2}, \ldots, X_{mn}$ are IID with $\mathcal{Q}_{m0} = \{EXP(\lambda_m) : \lambda_m = \lambda_0\}$, and $\mathcal{Q}_{m1} = \{EXP(\lambda_m) : \lambda_m = \lambda_1\}$. A Gamma$(\kappa, \kappa)$ frailty induces dependency among $\{\mathbf{X}_m : m \in \mathcal{M}_1\}$. The independent Bernoulli prior with $P(\theta_m = 1) = \pi$ is used. Data was generated under the true model parameters: $M = 500, n = 30, \pi = .30, \lambda_0 = 1, \lambda_1 = .5, \kappa = 2$. To find the BMDF, a frailty-model with exponential marginals and a Gamma(κ, κ) frailty is used. In the Sequential Monte Carlo, $R = 1000$ particles are used with $\pi^* = .20, \kappa^* = 3$. Results for one replicate using different pairs of loss functions are shown in Table 3. The cost ratio used is $C_0/C_1 = 1$. Notice that the performance of all three BMDF are satisfactory even under misspecifications of the prior probabilities for the alternative hypotheses and the hyperparameter of the frailty distribution.

7.3. Two group composite hypothesis with independent Gaussian observations

Assume that $\mathbf{X}_m = (X_{m1}, X_{m2}, \ldots, X_{mn_1}, Y_{m1}, Y_{m2}, \ldots, Y_{mn_2})$ are independent with $X_{mi} \overset{iid}{\sim} N(\mu_{m1}, \sigma_m^2)$ and $Y_{mi} \overset{iid}{\sim} N(\mu_{m2}, \sigma_m^2)$. Consider testing $H_{0m} : \mu_{m1} = \mu_{m2}$ versus $H_{1m} : \mu_{m1} \neq \mu_{m2}$ with independent Bernoulli prior on $\boldsymbol{\theta}$ with probability π for the alternatives, and conjugate prior for nuisance parameters given by, for $m \in \mathcal{M}_1$, $\mu_{m1}, \mu_{m2} \overset{iid}{\sim} N(\nu_m, k_0\sigma_m^2)$, $\sigma_m^{-2} \sim$ Gamma(α, β), and

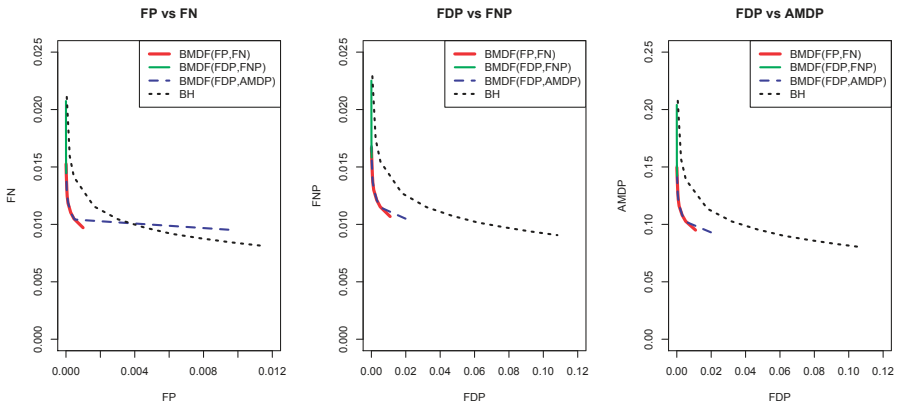

FIG 2. *Graphs of* $(\hat{FP}, \hat{FN})$, $(\hat{FDP}, \hat{FNP})$, *and* $(\hat{FDP}, \hat{AMDP})$ *for three BMDFs with different loss functions when the cost ratio varies and the BH procedure when the FDR threshold varies in a two-group composite hypothesis problem with independent Gaussian observations. Correct prior parameters are specified.*

for $m \in \mathcal{M}_0$, $\mu_{m1} = \mu_{m2} \overset{iid}{\sim} N(\nu_m, k_0\sigma_m^2)$, $\sigma_m^{-2} \sim \text{Gamma}(\alpha, \beta)$. True parameters used to generate the data are $M = 500, \pi = 0.1, k_0 = 200, \alpha = 4, \beta = 4, \nu_m = 20$. We performed 1000 simulations each with correct prior parameters and misspecified prior parameters: $\pi^* = 0.05, k_0^* = 100, \alpha^* = 20$ and with other parameters *empirically* estimated via $\nu_m = \overline{\mathbf{x}}_m, \beta = \overline{S_p^2(\mathbf{x}_m)}(\alpha - 1)/(k_0 + 1))$, where

$$\overline{S_p^2(\mathbf{x}_m)} = \frac{1}{M}\sum_{m=1}^{M} \frac{(n_1 - 1)S^2(x_{m1}, \ldots, x_{mn_1}) + (n_2 - 1)S^2(y_{m1}, \ldots, y_{mn_2})}{(n_1 + n_2)}$$

is the average of pooled variances. In order to stabilize the results when π is small, we used the adjusted version of MDP, the AMDP given by $L(\mathbf{a}, \boldsymbol{\theta}) = \frac{(\mathbf{1}-\mathbf{a})^\intercal \boldsymbol{\theta}}{(\boldsymbol{\theta}^\intercal \mathbf{1})+1}$. We implemented the three BMDFs associated with different loss functions with 10 different cost ratios C_0/C_1 and the BH procedure with 10 different FDR thresholds. Figure 2 and Figure 3 show graphs of three empirical risk pairs for all four procedures with correctly specified prior parameters and partially misspecified and partially empirically-estimated prior parameters, respectively. In Figure 2 with the correct prior specification, the empirical risk curves for all three BMDFs are well below that for the BH procedure indicating better performance, except for the BMDF associated with the (FDP, AMDP) pair of loss functions where the average FP loss is surprisingly high. This is because with small prior probability π for the alternatives, the AMDP could be large for some simulated data, and when the cost ratio C_0/C_1 is very small, the BMDF would rather sacrifice a large number of false positives to achieve the optimal combined risk of FDP and AMDP. In Figure 3, when the prior parameters are partially misspecified and partially empirically estimated, the results of all four

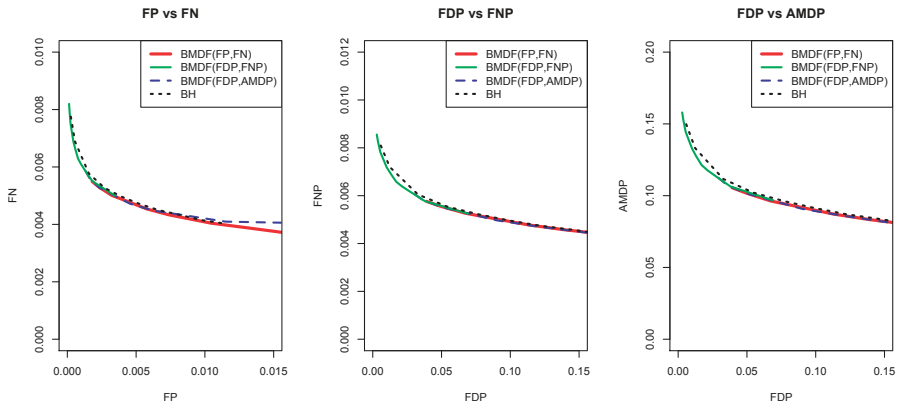

FIG 3. *Graphs of* $(\hat{FP}, \hat{FN})$, $(\hat{FDP}, \hat{FNP})$, *and* $(\hat{FDP}, \hat{AMDP})$ *for three BMDFs with different loss functions when the cost ratio varies and the BH procedure when the FDR threshold varies in a two-group composite hypothesis problem with independent Gaussian observations. The prior parameters are partially misspecified and partially empirically estimated.*

procedures almost coincide. The simulation results lend empirical support to the theoretical result in [3] that the BH procedure has the asymptotic optimality under the loss function pair (FP, FN) as a function of the cost ratio. However, more studies regarding more general loss functions and the empirical Bayes ideas are needed to obtain more reliable conclusions.

7.4. *A microarray data analysis*

In a colon cancer tumor metastasis study conducted in the laboratory of Dr. Marge Peña at the University of South Carolina, expression levels for 41268 genes from mice tissues were obtained through an Agilent Technology microarray. For each gene, five replicates were obtained for a control group and five replicates for a metastatic group. Computationally, the BMDF associated with (FP, FN) and (FDP, FNP) loss functions have relatively low cost, but the one associated with the (FDP, AMDP) loss function needs a much longer time, because it requires the computation of a posterior expectation involving all the genes. Therefore for illustration purposes, we simply randomly selected 500 genes out of the 41268 on which to apply our BMDF. We assumed the independent two-group Gaussian model, and used the partially empirically-estimated prior parameters described in Section 7.3. Cost ratios used for (FP, FN), (FDP, FNP) and (FDP, AMDP) loss function pairs are, respectively, $C_0/C_1 = 3, 0.2, 2$, and for BH procedure the FDR threshold is 0.05. These cost ratios and thresholds were chosen according to the simulation results in Section 7.3 in order for the four procedures to have similar empirical FDPs. Out of the 500 genes, the BMDF associated with the loss function pairs (FP, FN), (FDP, FNP), and (FDP, AMDP) found 15, 14, and 14 genes differentially expressed across groups, re-

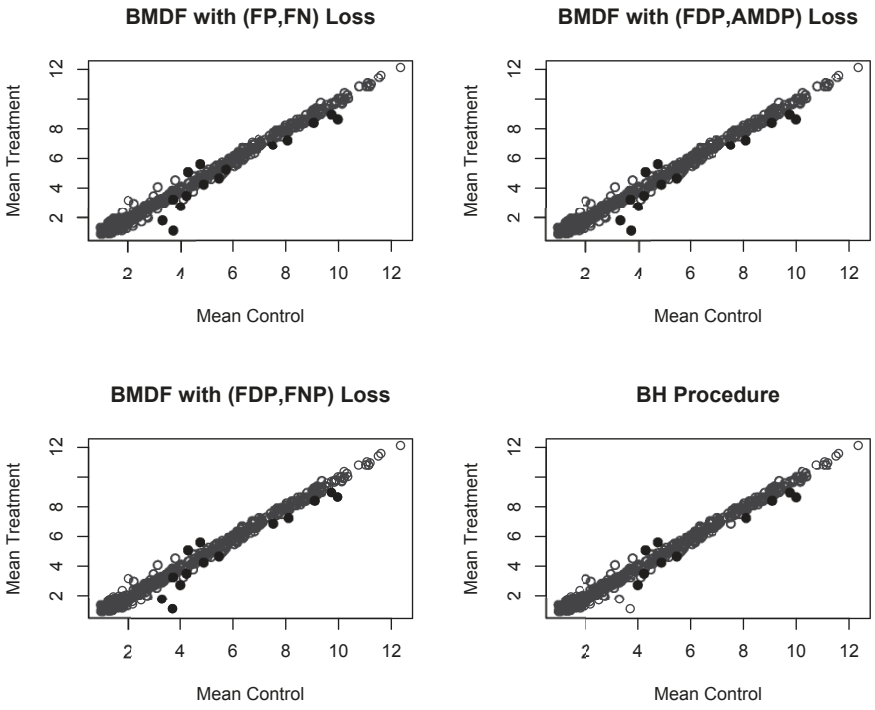

FIG 4. *Mean expression level of the control group versus the treatment group for 500 genes in a microarray data set. Gray circles are non discoveries and black circles are discoveries. Top left: BMDF with (FP, FN) loss function pair; top right: BMDF with (FDP, AMDP) loss function pair; bottom left: BMDF with (FDP,FNP) loss function pair; bottom right: BH procedure.*

spectively, and the BH procedure found 10. Figure 4 shows the mean expression level of the control group versus the treatment group of the discovered genes. Notice that all 10 genes discovered by the BH procedure were also discovered by the three BMDFs.

8. Concluding remarks

BMDF developed in this paper generates a class of multiple decision procedures which are optimal in the Bayesian framework for a general class of loss functions. The results in Theorem 1 describe the form of the BMDF and provide an efficient algorithm of finding the associated decisions in multiple testing settings. Notice that the pairs of loss functions are not limited to those described in Section 4. For example, the adjusted MDP given in Table 1 may help stabilize the computation of the the Bayes optimal actions when the prior probabilities of the alternative hypotheses, the $\pi_m = Pr(\theta_m = 1), \pi = 1, 2, \ldots, M$, are small. Also notice that

the choice of the loss function pairs and the cost ratio should be pre-determined in consultation with specialists of the scientific discipline relevant to the specific application.

The frailty-based model is a class of flexible models for dependent data structure, where the distribution of the frailty can be specified in a hierarchical manner with hyperparameters. Similarly, the prior distribution could also be dependent with frailty-based structures. Furthermore, the SMC could be easily implemented in the computations of the posterior expectations. Note, however, that not all dependent structures are frailty-based. Therefore, in real data analysis, model validation is needed to see the validity of the imposed dependent structures.

One possible extension of this research is in two-class prediction problems where the form of the BMDF could be extended with the loss functions replaced by prediction errors. Besides the usual prediction loss function pair of (FP, FN) that has been studied extensively, such as in [3], we can also consider a similar class of loss functions in multiple prediction problems, where the class memberships of many new items are to be predicted simultaneously. Future studies may also include extensions to model selection, and the empirical Bayes approach to determining prior hyperparameter values. In particular, of interest is to study whether the empirical Bayes procedures are equivalent to the non-Bayesian BH multiple decision function and the procedure in [18].

Acknowledgments

The authors thank Dr. Marge Peña for providing the microarray data set and for discussions regarding microarrays. They also thank the referees, associate editor, and editor for their constructive comments and criticisms which led to considerable improvements in the paper.

The authors acknowledge supports from National Science Foundation (NSF) Grants DMS 0805809 and DMS 1106435, National Institutes of Health (NIH) Grants RR17698 and R01CA154731, and Environmental Protection Agency (EPA) Grant RD-83241902-0 to the University of Arizona with subaward number Y481344 to the University of South Carolina.

This paper is based on a portion of the first author's PhD dissertation at the Department of Statistics, University of South Carolina, Columbia.

References

[1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 1, 289–300. MR1325392 (96d:62143)

[2] Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*, Second ed. Springer Series in Statistics. Springer-Verlag, New York. MR804611 (87i:62003)

[3] Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *Ann. Statist.* **39**, 3, 1551–1579. http://dx.doi.org/10.1214/10-AOS869. MR2850212 (2012j:62019)

[4] Bogdan, M., Ghosh, J. K., and Tokdar, S. T. (2008). A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing. In *Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Pranab K. Sen*. Inst. Math. Stat. Collect., Vol. **1**. Inst. Math. Statist., Beachwood, OH, 211–230. http://dx.doi.org/10.1214/193940307000000158. MR2462208 (2009m:62018)

[5] Casella, G. and Berger, R. L. (2001). *Statistical Inference, 2nd ed.* Duxbury Press.

[6] Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* **23**, 1, 1–22. MR2431866

[7] Efron, B. (2010a). *Large-scale inference*. Institute of Mathematical Statistics Monographs, Vol. **1**. Cambridge University Press, Cambridge. MR2724758

[8] Efron, B. (2010b). The Future of Indirect Evidence. *Statistical Science* **25**, 2, 145–157.

[9] Gordon, N. J. and Smith, A. F. M. (1993). Approximate non-Gaussian Bayesian estimation and modal consistency. *J. Roy. Statist. Soc. Ser. B* **55**, 4, 913–918. MR1229888 (94b:62077)

[10] Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*, Second ed. Springer Series in Statistics. Springer, New York. MR2722294

[11] Hougaard, P. (2000). *Analysis of multivariate survival data*. Statistics for Biology and Health. Springer-Verlag, New York. http://dx.doi.org/10.1007/978-1-4612-1304-8. MR1777022 (2001h:62003)

[12] Knuth, D. E. (1973). *The art of computer programming. Volume 3*. Addison-Wesley Publishing Co., Reading, Mass.-London-Don Mills, Ont. MR0445948 (56 #4281)

[13] Liu, J. S. (2001). *Monte Carlo strategies in scientific computing*. Springer Series in Statistics. Springer-Verlag, New York. MR1842342 (2002i:65006)

[14] Müller, P., Parmigiani, G., and Rice, K. (2007). FDR and Bayesian multiple comparisons rules. In *Bayesian statistics 8*. Oxford Sci. Publ. Oxford Univ. Press, Oxford, 349–370. MR2433200

[15] Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *J. Amer. Statist. Assoc.* **99**, 468, 990–1001. MR2109489

[16] Nelsen, R. B. (1999). *An introduction to copulas*. Lecture Notes in Statistics, Vol. **139**. Springer-Verlag, New York. MR1653203 (99i:60028)

[17] Neutial, P. and Roquain, E. On false discovery rate thresholding for classification under sparsity. *To appear in Ann. Statist.*.

[18] Peña, E. A., Habiger, J., and Wu, W. (2011). Power-enhanced multiple decision functions controlling family-wise error and false discovery rates. *Annals of Statistics* **39**, 1, 556–583.

[19] Ripley, B. D. (1987). *Stochastic simulation*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York. MR875224 (88b:68181)

[20] Sarkar, S. K., Zhou, T., and Ghosh, D. (2008). A general decision theoretic formulation of procedures controlling FDR and FNR from a Bayesian perspective. *Statist. Sinica* **18**, 3, 925–945. MR2440399

[21] Scott, J. G. and Berger, J. O. (2006). An exploration of aspects of Bayesian multiple testing. *J. Statist. Plann. Inference* **136**, 7, 2144–2162. http://dx.doi.org/10.1016/j.jspi.2005.08.031. MR2235051

[22] Storey, J. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics 31*, 2012 – 2035.

[23] Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. *J. Amer. Statist. Assoc.* **102**, 479, 901–912. http://dx.doi.org/10.1198/016214507000000545. MR2411657