Bayesian screening for feature selection

A. Lawrence Gould, Richard Baumgartner, and Amanda Zhao

Biostatistics and Research Decision Sciences Merck & Co Inc Kenilworth, New Jersey, USA

ABSTRACT

Biomedical applications such as genome-wide association studies screen large databases with high-dimensional features to identify rare, weakly expressed, and important continuous-valued features for subsequent detailed analysis. We describe an exact, rapid Bayesian screening approach with attractive diagnostic properties using a Gaussian random mixture model focusing on the missed discovery rate (the probability of failing to identify potentially informative features) rather than the false discovery rate ordinarily used with multiple hypothesis testing. The method provides the likelihood that a feature merits further investigation, as well as distributions of the effect magnitudes and the proportion of features with the same expected responses under alternative conditions. Important features include the dependence of the critical values on clinical and regulatory priorities and direct assessment of the diagnostic properties.

1. Introduction

Many investigations, especially those using modern data collection, generate large numbers of continuous-valued data measurements (“features”) per experimental unit. Most of these features are unlikely to be relevant. Obtaining large numbers of good observational units is often difficult. Classical methods such as discriminant analysis are unlikely to be useful for identifying relevant features when there may be thousands of measurements and few observational units (Donoho and Jin 2009b). Typical investigations include biomedical applications such as genomics ensuing from gene microarrays (Ahdesmäki and Strimmer 2010) and genome-wide association studies (GWAS) (Barnett et al. 2017), metabolomics (Wehrens and Franceschi 2012a, 2012b), drug safety (Gould 2008, 2013, 2018), industrial applications (Li et al. 2019), and, recently, economic applications (Cai and Sun 2017a). The effects of rare and weakly expressed important features may be obscured by the noise due to features that are unrelated to the attributes being studied. Such investigations exemplify large-scale inference (LSI), which addresses two basic issues: determining which of a large collection of potential explanatory variables or regressors should be retained in a model, and identifying the elements of very large observation vectors based on relatively few independent observations to consider further (Donoho and Jin 2015), which this paper addresses.

Screening an extensive series of measurements to identify relatively rare features that express potentially interesting differences (“signals”) entails two basic aims: detection to determine if an ensemble of features distinguishes between the attributes being studied, and identification to identify which features might be informative. Depending on the number of features and the sparsity and weakness of the signals, it may not be possible to detect, let alone identify, potential signals (Donoho and Jin 2004, 2008, 2009b, 2015). Subsequent analyses become more efficient and informative when the data dimensionality is reduced by discarding non-informative features. There is a substantial literature
on the properties of methods for detecting signals from large databases when the signals are rare and weak. The missed discovery rate (MDR) becomes important in this setting because of the difficulty of identifying true “signals” using conventional tools such as FDR (False Discovery Rate) that may be more effective in the rare signal setting when the signals are not weak (Donoho and Jin 2015; Klaus and Strimmer 2013; Li et al. 2019). The gene expression data set described by Golub et al. (1999) provides a suitable illustrative example. The goal is to find genes of acute lymphoblastic leukemia (ALL) that are differentially expressed from genes of acute myeloid leukemia (AML). Each feature represents a gene as measured by the gene microarray technology. A successful screening procedure reduces a large collection of genes to a manageable number of potentially “interesting” genes while retaining an acceptably small probability of missing the “interesting” genes. This well-known data set has been used in various publications, and we use it to show the utility of our method in the Examples section.

1.1. Conventional screening methods

Frequentist-based screening methods have been studied extensively, including variations on the positive False Discovery Rate (pFDR) (Storey 2002, 2003, 2007; Storey et al. 2019, 2004; Storey and Tibshirani 2003), an empirical Bayes method ashr (Stephens 2017; Stephens et al. 2019) that enables estimation of the FDR as well as the corresponding effect sizes, and methods tailored to the use of penalty functions to address issues arising in the evaluation of high-dimensional data. (Fan et al. 2014; Fan and Lv 2010; Fan and Song 2010) Decision-theoretic approaches have also been described, differing largely in how the loss functions are defined (Longford 2013, 2014; Wu and Pena 2013).

The “Higher Criticism” strategy suggested originally by John Tukey for determining if an aggregate collection of hypothesis test results provides sufficient evidence to warrant concluding that at least some of the tests may reflect real differences has been addressed extensively in the literature. (Donoho and Jin 2004, 2006, 2008, 2009b, 2015; Ingster 1999; Jin and Ke 2016; Li and Siegmund 2015) Different statistics can be used to implement the strategy, including a “Higher Criticism” (HC) statistic (Donoho and Jin 2015) and its variations, the Berk-Jones (BJ) statistic (Berk and Jones 1979; Wellner and Koltchinskii 2003; Zhang and Wu 2018), and the Average Likelihood (AL) statistic (Walther 2013). HC statistics can also be used to identify features for further evaluation. Whether individual features can be identified or whether one can even decide if any feature differences can be detected depends on the sparsity of individual signals and on their strength, which can be clarified using a phase diagram (Cai et al. 2007; Donoho and Jin 2004; Ingster 1999; Ji and Jin 2012; Jin and Ke 2016; Klaus and Strimmer 2013; Xie et al. 2011). Detectability is not an issue for the examples considered here.

The HC and BJ statistics are examples of a general class of statistics called φ-divergence statistics. (Jager and Wellner 2007) Let X₁, . . . , Xₙ denote the elements of an observation vector that are independently distributed on the real line according to F(x). The objective is to test H₀: F = F₀ against H₁: F ≠ F₀ where F₀ denotes the cdf of a uniform (0,1) distribution. The empirical cdf of any observation vector is

\[ F_N(x) = \frac{\# \ X_i \leq x}{N}. \]

Let \( p_i, i = 1, \ldots, N \) denote 1-sided or 2-sided p-values \([1 - F_0(x_i)] \text{ or } 2(1 - F_0(|x_i|)) \) obtained by testing H₀ against H₁ for each of the Xᵢ. Following Zhang et al. (2017), let \( p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(N)} \) denote the order statistics, define a supremum domain \( R \) by

\[ R = \{i : k_0 \leq i \leq k_1\} \cap \{p_{(i)} : \alpha_0 \leq i \leq \alpha_1\} \]

where \( k_0, k_1, \alpha_0, \) and \( \alpha_1 \) are specified constants, and consider a general family of statistics for testing H₀ \( p_i \sim U(0,1), \)

\[ S_{n,R} = \sup_{R} f\left(1/N, p_{(i)}\right) \]
where the definition of f depends on the statistic in the family. The $\phi$-divergence statistics (Jager and Wellner 2007) for 2-sided tests are

$$f^\phi_s(x, y) = \frac{1 - x^s y^{1-s} - (1-x)^s (1-y)^{1-s}}{s(1-s)} \quad s \neq 0, 1$$

$$f^\phi_0(x, y) = x \log(x/y) + (1-x)\log(1-x)/(1-y)$$

$$f^\phi_1(x, y) = y \log(y/x) + (1-y)\log(1-y)/(1-x)$$

In particular,

$$BJ = \sup_{0 \leq s \leq 1} f_1(F(x), F_0(x))$$ and reverse $$BJ = \sup_{0 \leq s \leq 1} f_0(F(x), F_0(x))$$

The modified HC statistics described by Donoho and Jin are given by

$$MHC_{2004} \text{ (Donoho and Jin 2004)} = \max_{R} f_2(F(x), F_0(x))$$

and

$$MHC_{2008} \text{ (Donoho and Jin 2008)} = \max_{R} f_{-1}(F(x), F_0(x))$$

One-sided tests are more appropriate for signal detection,

$$f_s(x, y) = \begin{cases} 
\sqrt{2Nf^\phi_s(x, y)} & y \leq x \\
-\sqrt{2Nf^\phi_s(x, y)} & y > x 
\end{cases}$$

Zhang et al. (2017) provide exact and approximate expressions for the cdf of $S_{n,R}$.

1.2. Bayesian screening methods

Bayesian feature selection methods have been described in the literature, primarily aimed at identifying genes whose expression levels differ among different experimental conditions, either in terms of feature expression intensity or gene counts. With few exceptions, these methods use MCMC calculations, which may limit their practicability for very large feature sets.

Do et al. (2005) describe a Dirichlet process mixture model for the probability that a gene is expressed differently under alternative conditions. Posterior inference for the model is carried out using MCMC simulation.

Lewin et al. (2006) use an ANOVA model to model the logarithm of the expression of genes evaluated using microarrays under alternative experimental conditions as normally distributed with a mean that depends on an overall effect for each gene, a gene-specific differential effect for the condition, and an effect corresponding to the array that is a quadratic spline function of the expression level. The gene and condition-specific variances of the expression levels are assumed to be exchangeable. MCMC calculations are used to obtain posterior distributions of the parameters. Differential gene expressions are identified by sufficiently large values of the posterior probability that the gene-specific differential condition exceeds a predefined critical value.

In a later paper, Lewin et al. (2007) use a three-component mixture at the parameter level to model separately the populations of over-expressed, under-expressed, and non-differentially expressed genes. The choice of the mixture prior can have a large effect on the classification of genes. Simulations cannot provide general guidance as to the best model so that model checking methods are needed. This approach also requires MCMC calculations, for which software is available.

Yu et al. (2008) use a hierarchical random effect model and posterior distributions of Bayes factors to select differentially expressed genes without assuming exchangeability of Bayes factor distributions across genes. The method is based on the computation for each gene of the relative probability of
concluding incorrectly the null hypothesis of no differential effect for the gene. Conjugate prior distributions for the model parameters that may include historical information are used to construct gene-specific posterior distributions for the Bayes factor and the calculation of critical values for the probability that there is no differential effect. In general, determination of the critical value will require MCMC calculations.

Yu et al. (2015) subsequently proposed “confident difference criterion” Bayesian gene selection algorithms, one based on the standardized differences between two mean expression values among genes and the other incorporating as well the differences between two variances. These methods evaluate the posterior probability of a gene having different gene expressions between alternative conditions and declare a gene to be differentially expressed if this posterior probability is large. Gibbs sampling is needed to obtain samples from the posterior distributions of the parameters under hierarchical normal models with conjugate priors.

Van de Wiel et al. (2013) focus on negative binomial count models in a Bayesian generalized linear model setting and propose a hybrid full Bayes – empirical Bayes type approach in which priors of crucial parameters are estimated and arbitrary parametric and non-parametric priors are allowed for other parameters. The calculations are based on Integrated Nested Laplace Approximations (INLAs) for latent Gaussian models (Rue et al. 2009) that avoid the need for MCMC calculations.

Dadaneh et al. (2018) describe methods for detecting differentially expressed genes that model high-throughput sequencing count data using a gamma-negative binomial process, or a beta-negative binomial process (BNBP). Posterior distributions of the model parameters are obtained using Gibbs sampling. Detection of differentially expressed genes is based on tests of a null hypothesis that the posterior distributions of the gene-specific regularized (by gamma or beta processes) negative binomial shape parameters are the same across different experimental conditions. Instead of standard hypothesis testing methods, their approach measures the distances between gene-specific posterior distributions using symmetric Kullback-Leibler divergence.

Fisher and Mehta (2015a,b) describe an approach for feature selection, “Bayesian Ising Approach” (BIA) that can be used to obtain posterior probabilities corresponding to L2 penalized regression models without the need for MCMC calculations. Their method relates an n-dimensional vector \( y \) of responses to an \( n \times p \) matrix of features \( X \) (\( p \gg n \)) by \( y = X\beta + \eta \), with the aim of identifying the nonzero elements of \( \beta \) using L2 penalized regression, i.e., by minimizing \( U(\beta) = (y-X\beta)'(y-X\beta) + \lambda\beta'\beta \) where \( \lambda \) is a penalty factor. The metric used is the expected value of the posterior probability that \( \beta \neq 0 \) for each element of \( \beta \). Instead of MCMC calculations that a conventional Bayesian approach would use, Fisher and Mehta propose a non-MCMC alternative that implements statistical physics ideas for studying high-dimensional regression based on “magnetizations of an Ising model with weak couplings”. A key quantity for this purpose is the value of a natural scale parameter \( \lambda^* \) such that the method will identify nonzero \( \beta \) values accurately when \( \lambda \ll \lambda^* \), but will be unreliable when \( \lambda \gg \lambda^* \). A plot of the estimated values of the expected probability that \( \beta \neq 0 \) against a range of values of \( \lambda^*/\lambda \) for each feature, referred to as a feature selection path, provides a way to identify relevant features. The purpose is to screen variables (features) rapidly to identify variables with low correlation to condition states (e.g., treatment) that can be removed before undertaking more intensive calculations. Rate limiting factors include the need to compute high-dimensional correlation matrices and the fact that correlations can strongly affect the posterior probabilities.

### 1.3. Proposed approach

The method presented here is motivated by previous work on drug safety for Bayesian screening of adverse effects (Gould 2008, 2013, 2018) based on random mixtures of distributions of event counts. The features in safety applications typically are counts of events generated by single-parameter binomial, or Poisson probability functions recorded in parallel test-control arms. The same method also applies to screening based on gene counts if binomial likelihoods are replaced with negative binomial likelihoods. The present paper extends the application scope of the approach to multi-
parameter distributions of continuous feature values with particular emphasis on normally distributed features. The method retains the useful statistical properties and computational efficiency previously demonstrated for the Bayesian paradigm, particularly with regard to inferences about the mixture parameters and effect sizes. We also describe the diagnostic properties of the method.

Although the mathematical models used here are superficially similar, the present approach has some fundamental differences from previous approaches. First, the focus is not on detecting features that differ between experimental conditions with control of the false discovery rate (FDR), as is the case for most current methods. Instead, the focus is on providing a statistically valid tool that subject matter experts can use to identify and, if appropriate, remove from further consideration features that are unlikely to differ materially between experimental conditions, so that the missed discovery rate (MDR) becomes the important metric. Reducing the dimensionality of the problem of identifying important features improves the efficiency and effectiveness of methods specifically aimed at identifying the important genes/features, including those described briefly above (Fisher and Mehta 2015a). Secondly, when the experimental conditions can be identified formally as “control” or “test”, the posterior distributions of metrics expressing differences between the conditions can be conditional on the ‘control’ feature response because the issue addressed is whether the process generating the expected response of a feature to the “test” condition is the same as (or differs from) the process generating the expected feature response to the ‘control’ condition. Thirdly, the strategy for determining the prior distributions of key model parameters takes the apparently novel approach of explicitly incorporating clinical and regulatory requirements about acceptable error rates. Finally, the calculations are exact (non-asymptotic) and rapid because they do not require MCMC calculations.

The paper is structured as follows: We first introduce the method and define the underlying generative model. We then describe the Bayesian screening calculations and provide simulated and real-life examples including comparisons with HC-based approaches (Donoho and Jin 2015; Walther 2013; Zhou et al. 2019) and an FDR-focused screening approach (Storey 2002, 2003, 2007; Storey et al. 2019, 2004; Storey and Tibshirani 2003). A comparison is also provided for the real-life example using a Bayesian approach with conveniently usable software that does not require MCMC computations (Fisher and Mehta 2015a). The comparisons are by no means exhaustive and, in the interest of maintaining focus and computational practicability, do not include other Bayesian methods described above or other methods that have been described in the literature (Fan et al. 2014; Fan and Lv 2008, 2010; Fan and Song 2010). Finally, we discuss the benefits and limitations of the method and opportunities for future research. An annotated example of the use of the R code and a listing of the R code is provided in the Supplemental information.

2. Method

2.1. Data
The data consist of collections of observations on $N_{\text{feat}}$ features reflecting the responses to a control or a test treatment. $Y_C = \{Y_{Ci}, i = 1, \ldots, N_{\text{feat}}\}$ denotes the feature-specific means of the observations on the control treatment, $Y_T = \{Y_{Ti}, i = 1, \ldots, N_{\text{feat}}\}$ denotes the feature-specific means of the observations on the test treatment, $s_{C_i}^2$ and $s_{T_i}^2$ denote estimates of the within-feature variance for each treatment that are independent of each other and from the observed responses, and $u_{Ti} = m_{Ti} s_{T_i}^2$ and $u_{Ci} = m_{Ci} s_{C_i}^2$ denote the sums of the residual squared errors for the responses to feature $i$ on the test and control treatments, respectively, where $m_{Ti}$ and $m_{Ci}$ denote the corresponding degrees of freedom. The pooled residual sum of squares and degrees of freedom for feature $i$ are $u_i = u_{Ti} + u_{Ci}$ and $m_{Ti} + m_{Ci}$. It is convenient to think of $C$ and $T$ as denoting control and test treatments, respectively, but this distinction is not essential. $C$ and $T$ could refer to two phenotypes, and the objective could be to identify genes that are expressed differently between the two phenotypes (section 3.2 provides an example).
2.2. Model

The calculations are described in terms of normally distributed outcomes for computational convenience, although, as the Appendix makes clear, the assumption of normality is not essential. The observed mean feature values $Y_{Ci}$ and $Y_{Ti}$ are assumed to have normal likelihoods with respective means $\mu_{Ci}$ and $\mu_{Ti}$ and precisions $a_{Ci}\eta_i$ and $a_{Ti}\eta_i$ ($\eta_i = 1/\sigma_i^2$ where $\sigma_i^2$ denotes the feature-specific residual variance), respectively, $i = 1, \ldots, N_{\text{feat}}$. Let $a_C = \{a_{Ci}, i = 1, \ldots, N_{\text{feat}}$ and $a_T = \{a_{Ti}, i = 1, \ldots, N_{\text{feat}}\}$ denote the vectors of precision multipliers.

The means $\mu_{Ci}$ and $\mu_{Ti}$ have normal prior distributions with respective expected values $\theta_{Ci} = \theta_{Ci}$ and $\theta_{Ti}$, and precision $b_0\eta_i$, and the precision parameters $\eta_i$ are assumed to have a gamma prior distribution with parameters $\zeta_i/2$ and $\zeta_2/2$. The value of $\theta_{Ti}$ is a random mixture $\theta_{Ti} = (1-\gamma_i)\theta_{Ti}^{(C)} + \gamma_i\theta_{Ti}^{(T)}$ of the values that would apply under two conditions: that the same process generated the expected response of feature $i$ to the test and control treatments ($\theta_{Ti} = \theta_{Ti}^{(C)}$) or that different processes generated the expected responses ($\theta_{Ti} = \theta_{Ti}^{(T)} \neq \theta_{Ti}^{(C)}$). The mixture indicator variable has a Bernoulli prior distribution with parameter $\pi_0$, which is generated from a beta ($\xi_1, \xi_2$) prior density. The quantity $\pi_0$ is a parameter in the model for which a posterior distribution will be generated based on the observations, as opposed to conventional methods, which assume that $\pi_0$ is a fixed quantity whose value must be specified or estimated.

The parameters of the distribution of the difference or ratio metric corresponding to any feature depend on the process parameters (the values of and $\theta_{Ti}^{(T)}$ and $\theta_{Ti}^{(C)}$) and on scientific, clinical, or regulatory requirements. The approach extends a previously described method for count data (Gould 2018) to realizations of arbitrary likelihoods.

The value of $\theta_{Ti}^{(C)}$, the a priori mean of the expected response of feature $i$ to the test treatment when the same process generates the expected responses to the control and test treatments given the responses of feature $i$ to the control treatment, is assumed to be the same as $\theta_{Ci}$, the mean of the posterior density of $\mu_{Ci}$ that incorporates the observed response $Y_{Ci}$ and the mean $\theta_{Ci}$ of the prior density of $\mu_{Ci}$. This assumption is tenable because equality of processes means that $\theta_{Ti} = \theta_{Ci}$ and, in the absence of information about $\theta_{Ti}$, the best a priori estimate of $\theta_{Ti}$ for feature $i$ given the control treatment responses are $\theta_{Ci}$ when the same processes generate the expected responses to the two treatments.

2.3. Prior distribution parameters

A metric $M_i = M_i \left( \mu_{Ti}^{(T)}, \mu_{Ti}^{(C)} \right)$ such as the arithmetic difference $D_i = \mu_{Ti}^{(T)} - \mu_{Ti}^{(C)}$ or the ratio $R_i = \mu_{Ti}^{(T)} / \mu_{Ti}^{(C)}$ expresses the effect of the test treatment on the response of feature $i$. A key aspect of the method is that regulatory and clinical requirements, and the response of feature $i$ to the control treatment, determine the prior distributions of $\mu_{Ti}^{(T)}$ and $\mu_{Ti}^{(C)}$. The prior or posterior distribution of $M_i$ depends on the parameters of the corresponding distributions of $\mu_{Ti}^{(T)}$ and $\mu_{Ti}^{(C)}$.

If the processes are the same, then $\theta_{Ti}^{(T)} = \theta_{Ti}^{(C)}$ and $M_i$ should be unlikely to exceed a critical value $M_{\text{crit}}$ so that a priori,

$$ F_M(M_{\text{crit}}; \theta_{Ti}^{(C)}, \theta_{Ti}^{(C)}) = P(M_i \leq M_{\text{crit}}; \theta_{Ti}^{(C)}, \theta_{Ti}^{(C)}) \geq \phi_0, $$

(e.g., $M_{\text{crit}} = 4$ and $\phi_0 = 0.95$), where $F_M$ denotes the cdf of $M$. Likewise, if the processes differ, then $M_i$ should be unlikely to fall below $M_{\text{crit}}$ so that a priori,

$$ F_M(M_{\text{crit}}; \theta_{Ti}^{(T)}, \theta_{Ti}^{(C)}) = P(M_i \leq M_{\text{crit}}; \theta_{Ti}^{(T)}, \theta_{Ti}^{(C)}) \leq \phi_1. $$
(e.g., $\phi_1 = 0.1$ or 0.2). The biological and regulatory requirements for the screening application are expressed by (1) and (2). Figure 1 illustrates the stochastic ordering relationship. Section D of the Appendix describes how the requirements expressed by (1) and (2) determine the parameters of the prior distribution of $M$.

The focus of what follows is on the arithmetic difference between the expected responses of a feature to the control and test treatments so that henceforth the text often refers to $D_i$ and $d_{\text{crit}}$ instead of $M_i$ and $M_{\text{crit}}$.

### 2.4. Screening calculations

The calculations produce three key results: (a) posterior probability values that the same process generated the C and T expectations for feature i, $P_{\text{Spi}} = P(\gamma_i = 0)$ [“SP” means “same process”]. (b) values of the posterior cdf of $M_i$, for assessing the magnitude of the effect of the test treatment relative to the control, and (c) the posterior distribution of the proportion $\pi_0$ of the features where the same processes generated the expected values for the control and test treatments.

The $P_{\text{Spi}}$ values express the strength of the current evidence as to whether the same process generated the expected responses to the control and test treatments for feature i. A $P_{\text{Spi}}$ value less than a predefined constant, say $\omega$, would imply that different processes generated the expected feature responses, i.e. that the i-th feature is potentially informative. When there are two conditions (A and B, say) and there is no inherent reason to regard one as a control and one as a test, then the screening process can be run twice, with A regarded as the test and B as the control, and then with B regarded as the test and A as the control (see section 3.2 as an example).

### 2.5. Diagnostic properties

The practical utility of the method depends on its specificity, sensitivity, and positive and negative predictive values. These values are determined by assumptions about the true distributions of expected values of the feature responses to the control and test treatments and the number of replicated observations on the test and control treatments for each feature. The values of the parameters of the marginal densities of the observed responses of the features to the test and control treatments can be specified subjectively, reflecting a desire to evaluate the robustness of the method, e.g., by considering 'worst cases', or can be based explicitly on the findings from the current (or another completed) experiment.
Suppose that the posterior density of the mixing parameter $\pi_0$ determined from the current trial also describes the predictive distribution of this parameter in the new trial, provided by (A6). The diagnostic properties can be calculated for each combination of control parameters ($\Phi_0, \Phi_1, M_{crit}$) used for the analysis of the current trial (each of these combinations will provide values of the posterior density of $\pi_0$) and for various sample sizes for the future trial.

Simulation of the screening process as it applies to a current experiment proceeds as follows. The current experiment provides vectors of observed feature mean responses $Y_C$ and $Y_T$, values of the pooled within-feature residual squared errors, $u_C$ and $u_T$ with, respectively, $m_C$ and $m_T$ degrees of freedom (df), and the posterior density of $\pi_0$. The posterior expectations of $\theta_C, \theta_T, \zeta_0$, and $\zeta_1$ based on the responses to the control and test treatments define predictive distributions of future observations. It would be advisable to confirm that the marginal densities of the responses corresponding to the control treatment with the posterior expected parameter values fit the observed data acceptably using, e.g., the Kolmogorov-Smirnov or Cramer-von Mises tests.

$S$ replications of the current experiment are to be done, with $a_{Ci} = a_{Ti} = a, i = 1, \ldots, N_{\text{feat}}$ for each replicate. Each replication starts with $N_{\text{feat}}$ observations $v$ from a beta $(m + \zeta_1, m)$ density that provides $N_{\text{feat}}$ corresponding values of $u = \zeta_2 v/(1-v)$, $N_{\text{feat}}$ observations $Z_C$ from a central $t$ distribution with $(m + \zeta_1)$ df, and $N_{\text{feat}}$ observations $Z_T$ from the same central $t$ distribution. The potential future treatment-related outcomes $X_C^{(s)}$ and $X_T^{(s)}$ for the $s$-th replicate of the experiment are

$$X_C^{(s)} = \frac{u_i + \zeta_2}{a (m + \zeta_1)} Z_C + \theta_C^0 \text{ and } X_T^{(s)} = \frac{u_i + \zeta_2}{a (m + \zeta_1)} Z_T + \theta_T^0$$

$N_{\text{feat}}$ random values of $\pi_{0i}, i = 1, \ldots, N$, are drawn from the posterior density of $\pi_0$ based on the current experiment, and corresponding realizations $\bar{Y}_i$ are drawn from Bernoulli distributions with parameters $\pi_{0i}, i = 1, \ldots, N_{\text{feat}}$. Separate values are obtained for each replicate. The simulated feature responses for the $s$-th replicate are

$$\text{control: } Y_{C_i}^{(s)} = X_C^{(s)} \text{ and test: } Y_{T_i}^{(s)} = (1 - \bar{Y}_i) X_C^{(s)} + \bar{Y}_i X_T^{(s)}$$

The values of $p_{\text{Spi}}^{(s)}$, $s = 1, \ldots, S$, [Appendix eqn (A8)] are the probabilities that the same process generates the expected responses of feature $i$ to the control and test treatments. Asserting that a feature may be affected by the test treatment differently from the control treatment depends on the magnitude of this probability. Whether this identification is correct or not depends on the corresponding value of $\bar{Y}_i$. Let $d_i^{(s)} = 0$ for the $s$-th replicate if $p_{\text{Spi}}^{(s)} > \omega(0 < \omega < 1)$ and 1 otherwise; $d_i^{(s)} = 0$ asserts that the same process generates the expected responses for feature $i$ to the control and test treatments; $\omega$ is a threshold probability value for determining whether the control and test treatments affect a feature differently or not. The value of $d_i^{(s)}$ will be correct if $d_i^{(s)} = \bar{Y}_i$. For the $s$-th replicate, let

$$h_{00}^{(s)} = N^{-1} \sum_{i=1}^{N_{\text{feat}}} I(d_i^{(s)} = \bar{Y}_i = 0) \quad h_{01}^{(s)} = N^{-1} \sum_{i=1}^{N_{\text{feat}}} I(\bar{Y}_i = 0, d_i^{(s)} = 1)$$

$$h_{10}^{(s)} = N^{-1} \sum_{i=1}^{N_{\text{feat}}} I(\bar{Y}_i = 1, d_i^{(s)} = 0) \quad h_{11}^{(s)} = N^{-1} \sum_{i=1}^{N_{\text{feat}}} I(d_i^{(s)} = \bar{Y}_i = 1)$$

where $I(x)$ is 1 if $x$ is true, or 0 if $x$ is false. The $h_{ab}^{(s)}$ values estimate the probability that the screening process correctly identifies the features that are and are not affected by the test treatment. The averages of these values over $S$ replicates provide the basis for evaluating the diagnostic properties of the screening method. Let $H$ denote the $2 \times 2$ matrix consisting of these averages,

$$H = \begin{bmatrix} h_{00} = S^{-1} \sum_{s=1}^{S} h_{00}^{(s)} & h_{01} = S^{-1} \sum_{s=1}^{S} h_{01}^{(s)} \\ h_{10} = S^{-1} \sum_{s=1}^{S} h_{10}^{(s)} & h_{11} = S^{-1} \sum_{s=1}^{S} h_{11}^{(s)} \end{bmatrix}$$
The diagnostic properties of the screening procedure, considered as a function of the specifications \( (M_{\text{crit}}, \Phi_0, \Phi_1, \psi) \), are functions of the elements of \( H \),

\[
\text{Specificity} = \frac{\bar{h}_{00}}{\bar{h}_{00} + \bar{h}_{01}} \quad \text{Sensitivity} = \frac{\bar{h}_{11}}{\bar{h}_{10} + \bar{h}_{11}}
\]

\[
\text{FDR} = \frac{\bar{h}_{01}}{\bar{h}_{01} + \bar{h}_{11}} \quad \text{MDR} = \frac{\bar{h}_{10}}{\bar{h}_{00} + \bar{h}_{10}}
\]

Neither the sensitivity nor the specificity depends on the marginal distribution of the comparison metric \( M \) (difference \( D \) or ratio \( R \)). However, the FDR (False Discovery Rate) and MDR (Missed Discovery Rate) can be sensitive to the event prevalence. It is easy to show that

\[
\text{FDR} = (1 + UR)^{-1} \quad \text{MDR} = (1 + V/R)^{-1}
\]

where \( R = P(\text{Different processes})/P(\text{Same process}) \), \( U = \text{Sensitivity}/(1 - \text{Specificity}) \), and \( V = \text{Specificity}/(1 - \text{Sensitivity}) \). The quantities \( U \) and \( V \) are independent of the probability that the same process generates the expected responses, but \( R \) is not. The FDR increases and the MDR decreases as \( R \) decreases, so that the FDR could be large when few features are affected differently by the test and control treatments. Good control of the MDR requires high sensitivity, regardless of the specificity. Figure 2 illustrates the relationships.

### 2.6. Comparison with conventional screening

Conventional screening approaches such as the pfdr method (Storey 2002, 2003, 2007; Storey et al. 2019, 2004; Storey and Tibshirani 2003) identify potentially important features such as differentially expressed genes by repeated significance testing with corrections to the significance levels to control the effect of multiple testing on the FDR. A feature is “flagged” as potentially important if a corrected significance level (often called a “q-value”) exceeds a specified critical value \( q_{\text{crit}} \). The value of \( q_{\text{crit}} \) is chosen to control the FDR at an acceptable level.

The Bayesian screening method described here characterizes individual features in terms of the posterior probability that the expected response to a nominal test intervention or treatment is generated by the same process that generated the nominal control test or treatment values as opposed to a process that centers the distribution of expected ‘test’ values at a clinically, biologically, or regulatorily important (difference) value \( d_{\text{crit}} \). The posterior probabilities, along with posterior distributions of the test-control differences, provide the strength of evidence for deciding whether a particular feature should be ‘flagged’ for further investigation. For example, the determination of whether a feature should be pursued may depend on the clinical importance of the feature in the context of product safety evaluation.

Since the Bayesian approach is selective with respect to the magnitude of the difference that is important, while the conventional approach is not, the conventional approach can identify more potential candidates for follow-up than the Bayesian approach does. In particular, an observed test-control difference that is smaller than \( (M_{\text{crit}} = d_{\text{crit}} \) and sufficiently precise because it is the mean of many replications may yield a q-value less than \( q_{\text{crit}} \) while the posterior probability that the expected feature response to the ‘test’ treatment is generated by the same process as the expected response to the ‘control’ treatment may not be small. The ‘real data’ example in the next section illustrates this phenomenon.
Figure 2. FDR and MDR as functions of Prevalence, Sensitivity, and Specificity. FDR curves for each Sensitivity value correspond (top to bottom) to Sensitivity = 0.9, 0.95, 0.97, and 0.99. MDR curves for each Specificity value correspond (top to bottom) to Specificity = 0.9, 0.95, 0.97, and 0.99.
3. Examples

3.1. Simulated data

This example illustrates the application of the method to a large dataset with known control and treatment feature identifiers. Two sets of data (S = 1 replication) were generated, each consisting of the average of \( a_{C1} = a_{T1} = a = 5 \) measurements of each of \( N_{\text{feat}} = 100,000 \) features. Each of the pooled residual sums of squares corresponding to the features (the \( u \) values) has \( m = 8 \) degrees of freedom. Setting the prior parameter values to \( \theta_C = 4, b_0 = 2, \zeta_1 = 12, \) and \( \zeta_2 = 10 \) implies that the marginal density of \( v = u/(u + \zeta_2) \) is a beta density with parameters \((4, 6)\). The marginal expected value of the standard deviation, \( \sigma = \sqrt{u/m} \) is

\[
\sqrt{u/m} = \sqrt{\frac{\zeta_1}{m} \frac{m+1}{2} \frac{\zeta_2-1}{2}/B\left(\frac{m}{2}, \frac{\zeta_2}{2}\right)} = 0.945
\]

so that the standard error for any of the mean \( Y \) values is \( \sigma/\sqrt{5} = 0.423 \). \( Y_C \) values for both sets of data were drawn from a \( N(4, 0.423) \) distribution. The values of \( Y_T \) for both sets of data were drawn from a \( N(\theta_T, 0.423) \) distribution where \( \theta_T = \theta_C + \delta \gamma \) with \( \delta = 4 \). The first set of data assumed that the values of \( \gamma \) were generated from a Bernoulli \((0.25)\) distribution \( \varepsilon(\gamma) = 0.25 \), so that \( \theta_T = 8 \) for about a quarter of the features (25,153 in fact), and \( \theta_T = 4 \) for the remaining features. The second set of data, which illustrates a situation in which signals are rare, assumed that the values of \( \gamma \) were generated from a Bernoulli \((0.005)\) distribution \( \varepsilon(\gamma) = 0.005 \), so that \( \theta_T = 4 \) for about...
Table 1. Features for which \( p_{SP} < 0.3 \) and \( p_{PS}(D > 4) < 0.5 \) when \( E(\gamma) = 0.005 \), \( d_{crit} = 4 \), \( \phi_0 = 0.99 \), and \( \phi_1 = 0.2 \). The numbers in the fifth through tenth columns are the \( p_{SP} \) values corresponding to the \((\phi_0, \phi_1)\) combinations.

| Ftr # | Diff | sd(Diff) | \( \phi_0 \) | 0.95 | 0.95 | 0.95 | 0.95 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
|-------|------|----------|---------------|------|------|------|------|------|------|------|------|------|------|------|
|       |      |          |               | 0.1  | 0.2  | 0.3  | 0.1  | 0.2  | 0.3  | 0.3  | 3    | 3    | 3    | 3    | 3    |
| 23746 | 4.28 | 0.38     | 1              | 1    | 1    | 1    | 0.62 | 0.21 | 0.09 | 0.98 | 0.96 | 0.91 | 0.82 | 0.68 |
| 56325 | 4.39 | 0.37     | 1              | 1    | 1    | 1    | 0.54 | 0.17 | 0.07 | 0.99 | 0.97 | 0.94 | 0.87 | 0.75 |
| 57663 | 4.18 | 0.30     | 1              | 1    | 1    | 1    | 0.53 | 0.13 | 0.04 | 0.99 | 0.97 | 0.92 | 0.81 | 0.64 |
| 61807 | 4.24 | 0.33     | 1              | 1    | 1    | 1    | 0.53 | 0.16 | 0.07 | 0.99 | 0.97 | 0.93 | 0.85 | 0.70 |
| 98743 | 4.32 | 0.26     | 1              | 1    | 1    | 1    | 0.33 | 0.05 | 0.01 | 0.99 | 0.98 | 0.96 | 0.89 | 0.74 |

Table 2. Diagnostic properties of the screening process when \( d_{crit} = 4 \), \( p_{SP} < \omega \) implies that different processes generate the expected feature responses to the control and test treatments. Specificity = 1 in all cases.

| \( \omega \) | \( \phi_0 \) | 0.95 | 0.95 | 0.95 | 0.95 | 0.99 | 0.99 | 0.99 | 0.95 | 0.95 | 0.95 | 0.95 | 0.99 | 0.99 |
|-------|---------------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| \( \phi_1 \) | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 |
| Sens  | 0.31 | 0.52 | 0.65 | 0.67 | 0.78 | 0.84 | 0 | 0 | 0 | 0.21 | 0.29 | 0.36 |
| FDR   | 0   | 0   | 0.01 | 0.01 | 0.02 | 0.03 | – | – | – | – | – | – | – | – |
| MDR   | 0.19 | 0.14 | 0.11 | 0.1 | 0.07 | 0.05 | 0.01 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 |
| \( \phi_0 \) | 0.95 | 0.95 | 0.95 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| \( \phi_1 \) | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 |
| Sens  | 0.31 | 0.52 | 0.65 | 0.67 | 0.78 | 0.84 | 0 | 0 | 0 | 0.24 | 0.33 | 0.41 |
| FDR   | 0   | 0   | 0.01 | 0.01 | 0.02 | 0.03 | – | – | – | – | – | – | – | – |
| MDR   | 0.19 | 0.14 | 0.11 | 0.1 | 0.07 | 0.05 | 0.01 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 |
| \( \phi_0 \) | 0.95 | 0.95 | 0.95 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| \( \phi_1 \) | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 |
| Sens  | 0.43 | 0.66 | 0.78 | 0.75 | 0.84 | 0.89 | 0 | 0 | 0 | 0.27 | 0.38 | 0.44 |
| FDR   | 0   | 0.01 | 0.03 | 0.01 | 0.03 | 0.05 | – | – | – | – | – | – | – | – |
| MDR   | 0.16 | 0.1 | 0.07 | 0.08 | 0.05 | 0.04 | 0.01 | 0.01 | 0.01 | 0 | 0 | 0 | 0 | 0 |
| \( \phi_0 \) | 0.95 | 0.95 | 0.95 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| \( \phi_1 \) | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 |
| Sens  | 0.51 | 0.74 | 0.85 | 0.79 | 0.88 | 0.92 | 0 | 0 | 0 | 0.31 | 0.43 | 0.5 |
| FDR   | 0   | 0.02 | 0.04 | 0.02 | 0.04 | 0.07 | – | – | – | – | – | – | – | – |
| MDR   | 0.14 | 0.08 | 0.05 | 0.07 | 0.04 | 0.03 | 0.01 | 0.01 | 0.01 | 0 | 0 | 0 | 0 | 0 |

500 features (517 in fact), and \( \theta_1 = 4 \) for the remaining features. These prior parameter values were used only to generate the simulated data and did not affect any of the subsequent calculations. The simulated data provide the ‘truth’ about the processes that generated the expected responses to the ‘control’ and ‘test’ interventions.

The values of \( b_0 \) and \( \delta \) required to satisfy (1) and (2) were determined as described in section D of the Appendix. The value of \( \bar{\xi}_1 \) was set to be equal to 2 for the screening calculations. The computations took about 4 minutes for each case on a laptop computer.

Figure 3 displays the posterior densities of \( \pi_0 \) as functions of \( \phi_0 \) and \( \phi_1 \) when \( d_{crit} = 4 \) for both examples. The range of values of \( \pi_0 \) with nontrivial posterior density values clearly depends on the expected value of \( \gamma \) that was used for the simulations. These ranges depend on the values of \( \phi_0 \) and \( \phi_1 \) for both examples. The probability that the same process generated the feature responses to the control and test treatments decreased as \( \phi_0 \) increased for each fixed value of \( \phi_1 \), and as \( \phi_1 \) increased for each fixed value of \( \phi_0 \). Figure 3 suggests that setting \( \phi_0 \) to a relatively high value, e.g., 0.99 or greater, and setting \( \phi_1 \) to 0.3 or greater, would lead to a posterior density for \( \pi_0 \) centered close to the true (usually unknown) value.

Low values of the posterior probability \( p_{SP} \) that the same process generated the control and test expected feature values do not necessarily imply that large difference values will have high posterior probabilities. Table 1 illustrates this point by displaying the values of \( p_{SP} \) that are less than 0.3 when \( E(\gamma) = 0.005 \), \( d_{crit} = 4 \), \( \phi_0 = 0.99 \), and \( \phi_1 = 0.2 \) for the features for which the posterior probability that
Table 3. Diagnostic properties of conventional screening procedures.

A. Positive fdr criterion (qvalue) (Storey et al. 2019)

| E(γ) | q_{crit} | Sensitivity | Specificity | FDR | MDR |
|------|----------|-------------|-------------|-----|-----|
| 0.25 | Min      | 0.000001    | 0.023       | 0.843 | 0.009 | 0.033 |
|      | Max      | 0.01        | 0.913       | 1    | 0.341 | 0.245 |
| 0.005| Min      | 0.000001    | 0           | 0.938 | 0    | 0.001 |
|      | Max      | 0.01        | 0.794       | 1    | 0.938 | 0.005 |

B. Empirical Bayes criterion (ashr)LFSR = Local false sign rate (Stephens et al, 2018)

| E(γ) | LFSR_{crit} | Sensitivity | Specificity | FDR | MDR |
|------|-------------|-------------|-------------|-----|-----|
| 0.25 | Min         | 0.000001    | 0.722       | 0.755 | 0.174 | 0.021 |
|      | Max         | 0.01        | 0.951       | 0.949 | 0.437 | 0.089 |
| 0.005| Min         | 0.000001    | 0.698       | 0.753 | 0.932 | 0.000 |
|      | Max         | 0.01        | 0.938       | 0.951 | 0.981 | 0.002 |

C. Higher Criticism criterion (SetTest) (Zhang and Wu 2018)

| E(γ) | Sensitivity | Specificity | FDR | MDR |
|------|-------------|-------------|-----|-----|
| 0.25 | 0.448       | 0.994       | 0.041 | 0.156 |
| 0.005| 0.899       | 0.937       | 0.931 | 0.001 |

(M_i = D_i \geq 4 \leq 0.50, and also displays the corresponding posterior probabilities that the true expected difference D_i exceeds various specified values. Table 1 (and Figure 3) also demonstrates the profound effect of the value of \( \phi_0 \) on the value of \( \min p_{SP} \).

Table 2 summarizes the diagnostic properties of the screening process for both simulation cases as a function of the design parameters; Specificity = 1 for all cases. Increasing \( \phi_0, \phi_1, \) and \( \omega \) clearly increased the sensitivity and decreased the MDR. There was not much effect on the FDR.

For purposes of comparison with conventional frequentist-based methods, the well-known positive fdr (pfdr) method implemented by the qvalue function (Storey 2003; Storey et al. 2019, 2004; Storey and Tibshirani 2003), an empirical Bayes method implemented using the ashr function (Stephens et al. 2019), and the Higher Criticism approach implemented using the SetTest function (Zhang and Wu 2018) also were applied to the simulated data. Table 3 provides values of the sensitivity, specificity, FDR, and MDR for each of these procedures. The diagnostic properties are functions of the critical criterion for the pfdr and empirical Bayes approaches, so they are reported corresponding to the minimum and maximum of a set of parameters.

Table 4. Diagnostic properties of the Bayesian screening approach when the variance of the simulated observations is multiplied by 10. (All specificities were essentially 1.0).

E(γ) = 0.25

| \( \omega = 0.5 \) | \( \sigma^2 = 0.89 \) | \( \sigma^2 = 9.0 \) | \( \sigma^2 = 0.89 \) | \( \sigma^2 = 9.0 \) |
|-------------------|-----------------|-----------------|-----------------|-----------------|
| \( \phi_0 \)      | 0.99            | 0.99            | 0.99            | 0.99            |
| \( \phi_1 \)      | 0.2             | 0.3             | 0.2             | 0.3             |
| Sens              | 0.78            | 0.84            | 0.57            | 0.63            |
| FDR               | 0.02            | 0.03            | 0.03            | 0.04            |
| MDR               | 0.07            | 0.05            | 0.13            | 0.11            |
| \( \omega = 0.6 \) | 0.99            | 0.99            | 0.99            | 0.99            |
| \( \phi_0 \)      | 0.2             | 0.3             | 0.3             | 0.3             |
| \( \phi_1 \)      | 0.81            | 0.86            | 0.63            | 0.69            |
| Sens              | 0.02            | 0.04            | 0.04            | 0.05            |
| MDR               | 0.06            | 0.04            | 0.11            | 0.09            |
| \( \omega = 0.7 \) | 0.99            | 0.99            | 0.99            | 0.99            |
| \( \phi_0 \)      | 0.84            | 0.89            | 0.76            | 0.33            |
| \( \phi_1 \)      | 0.03            | 0.05            | 0.05            | 0.07            |
| Sens              | 0.05            | 0.04            | 0.09            | 0.07            |
| MDR               | 0.05            | 0.04            | 0.09            | 0.07            |
potential critical values. Two-tail test values are considered for the qvalue method in Table 3 because the assumption of an essentially constant density of empirical p-values towards the upper end of the range (values near 1) required by the method does not appear to apply for the 1-tail significance levels.

Table 3 suggests that the FDR and MDR values obtained using the pfdr or empirical Bayes approaches can be sensitive to the true value of $\pi_0$ and to the critical value. The findings from Table 2 suggest that the Bayesian screening approach may be relatively insensitive to the true value of $\pi_0$ and to the values of the tuning parameters $\phi_0$, $\phi_1$, and $\omega$. This may not be surprising because the conventional and Bayesian approaches address different questions, as pointed out in Section 2.6. In addition, the findings in Table 3 would not be unexpected in view of the relationship between FDR, MDR, and prevalence displayed in Figure 2.

A question arose in the review of the manuscript about the effect of the variability of the responses on the diagnostic properties of the method. We addressed this question by repeating the simulations just described with the variance increased by a factor of 10. Table 4 displays the effect on the diagnostic properties. There appears to be little effect on the FDR when the prevalence of features that are affected by the test treatment is not rare ($E(\gamma) = 0.25$) although, consistently with the effect on power in conventional frequentist hypothesis tests, the sensitivity and MDR do decrease. When few features are affected by the test treatment ($E(\gamma) = 0.005$), the MDR and sensitivity do not appear to be affected greatly, but the FDR does increase substantially.

### 3.2. Real data

This example uses the well-known data regarding gene expression differences between two forms of leukemia [ALL = Acute Lymphoblastic Leukemia and AML = Acute Myeloid Leukemia] (Golub et al. 1999). The expression values were obtained from 27 samples of childhood ALL from the Dana Farber Cancer Institute leukemia data bank and 11 adult AML samples from the Cancer and Leukemia Group B leukemia cell bank plus an additional independent set of 20 ALL and 14 AML adult and childhood leukemias obtained from various sources. The unknown transformations of the original expression values for the 7128 genes provided by the 72 patients were downloaded from http://web.stanford.edu/~hastie/CASI_files/DATA/leukemia_big.csv. The analysis of the data can be, and has been, approached in many ways. The analyses described here use the average of the 72 expression values for each gene and each type of leukemia along with the usual within-gene residual sum of squares to illustrate an application of the screening approach for identifying genes that seem to be expressed differently in ALL and AML. Although the present analyses do not do so, typical data mining methods would be applied in practice with part of the data used to identify genes with potentially different expressions, and the remainder to validate the choice. This could be done with different splits of the data to provide some assurance about the robustness of the chosen set of genes. Another issue of interest, not addressed here, would be whether the ALL-AML gene differences were expressed differently for adult and childhood leukemias.

Examination of the distributions of the mean differences between the feature responses for the two forms of leukemia difference suggests that true mean differences larger than a critical value ($d_{\text{crit}}$) of 0.5 or 0.6 are highly unlikely; it is not clear that true differences less than 0.5 would be meaningful. Consequently, values of $d_{\text{crit}}$ of 0.5 and 0.6 were used for this illustrative example. Values of $\phi_0$ of 0.95 and 0.99, and values of $\phi_1$ of 0.1, 0.2, and 0.3 were chosen to define the clinical/regulatory requirements.

Figure 4 displays posterior densities of $\pi_0$ for the differences ALL-AML and AML-ALL when $d_{\text{crit}} = 0.5$. The displays for $d_{\text{crit}} = 0.6$ (not shown) are similar. Differential gene expression seems unlikely for either difference. The value of $\phi_0$ strongly affected the posterior density of $\pi_0$. These are subtle effects because most of the probability content of the distribution of $\pi_0$ is very close to 1, so that very few genes are likely to be expressed differently in ALL and AML regardless of the values of $\phi_0$, $\phi_1$, and $d_{\text{crit}}$. The finding here of a small number of genes that appear to be expressed differently between the ALL and AML types is consistent with the results obtained by the higher criticism approach reported by Donoho and Jin (2009a).
The findings included in Table 5 show that this is what happened. Table 5 displays the genes whose $p_{SP}$ values exceeded $\omega = 0.6$ for any of the differences ALL-AML and AML-ALL when $\phi_0 = 0.95$ or 0.99 and $\phi_1 = 0.1$, 0.2, and 0.3, and also displays the posterior probabilities that the true difference exceeds specified values. These posterior probabilities are important because observed differences with small $p_{SP}$ values do not necessarily lead to non-negligible posterior probabilities of meaningful differences. The values of $p_{SP}$ generally were near 1 (>0.99) for almost all of the genes when $\phi_0 \leq 0.95$. In other words, almost no genes would have been detected by the screening process if $\phi_0 \leq 0.95$.

Application of the pfdr method yielded different results primarily because the frequentist and Bayesian approaches address different questions, as noted in section 2.6. Table 6 displays the number of genes flagged by the qvalue program as a function of the $q_{crit}$ values.

Figure 5 displays posterior probability trajectories produced by the Bayesian approach described by Fisher and Mehta (2015a) as a function of the ratio $\lambda^*/\lambda$ for the first “best” 500 features, the “best” 20 features, and the “best” 12 features. The evidence for selecting any feature as “relevant” is fairly weak, yet a few features appear to be a bit less weak than most.

Table 7 lists the summary statistics for genes flagged by the Bayes screening approach described here with $d_{crit} = 0.5$, $\phi_0 = 0.99$, $\phi_1 = 0.2$, and $\omega = 0.6$, the pfdr approach with $q_{crit} = 1 \times 10^{-8}$, the Berk-Jones statistic implemented in an early version of the SetTest program (Zhang and Wu 2018), and the 20 ‘best’ features identified using the Fisher and Mehta (2015a) Bayesian approach. Of the 30 features in this table, 12 were flagged by all of the methods, 3 were flagged by 3, 7 by 2, and 8 features were flagged by only one method. The qvalue approach flagged 25 features, HC flagged 17, Fisher and Mehta flagged 18, and the Bayesian screening method flagged 19.
4. Discussion

This paper describes Bayesian screening of continuous-valued features using a Gaussian mixture model that extends a Bayesian screening method for adverse event counts. The computations produce three key quantities: the probability for each feature that the same process generates the expected responses to the C and T interventions/categories, the posterior distribution of the true T-C difference for each feature, and the posterior distribution of the proportion of features (the value of $n_0$) for which the C and T expected responses are generated by the same process. Between-group effect magnitudes can be calculated as differences or ratios. The assumption of normality is algebraically and computationally convenient, but not necessary. Similar calculations can be carried out with other distributional models, although possibly with a greater computational burden.

Bayesian screening was evaluated and contrasted in two scenarios that may be encountered in practice, namely when “treatment effects” on the feature values are “common”, and when they are “rare”, with $y_i$ are drawn from a Bernoulli (0.25) and Bernoulli (0.005), respectively. For example, the “common” scenario may correspond to expression data from DNA-based studies, whereas the “rare” scenario is more germane to expression data from RNA-based studies, which enables identification of rare or weakly expressed genes (Liu Y, et al. 2015).

Table 5. Values of $p_{crit}$ (as functions of $\phi_0$ and $\phi_1$) and posterior probabilities (as functions of $d$) that differences ($D_0$) exceed specified values when $d_{crit} = 0.5$. The probabilities for $D_0$ correspond to $\phi_0 = 0.99$ and $\phi_1 = 0.2$.

| Gene # |YT  |YC | Diff| $a_{crit}$| $\phi_0$| $\phi_1$| $\phi_0$| $\phi_1$| $\phi_0$| $\phi_1$| $\phi_0$| $\phi_1$| $\phi_0$| $\phi_1$| $\phi_0$| $\phi_1$| $\phi_0$| $\phi_1$| $\phi_0$| $\phi_1$|
|-------|----|----|-----|---------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
| 804   | 0.80 | 0.12 | 0.57 | 0.05 | 0.99 | 0.99 | 0.98 | 0.48 | 0.16 | 0.16 | 0.09 | 0.09 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 1144  | 0.96 | 0.05 | 0.85 | 0.11 | 0.99 | 0.98 | 0.95 | 0.44 | 0.42 | 0.42 | 0.05 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 1635  | 1.06 | 0.19 | 0.86 | 0.11 | 0.99 | 0.99 | 0.94 | 0.40 | 0.42 | 0.42 | 0.07 | 0.07 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 1685  | 1.56 | 0.05 | 1.61 | 0.18 | 0.99 | 0.99 | 0.52 | 0.52 | 0.42 | 0.42 | 0.07 | 0.07 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 1909  | 0.08 | 0.08 | 0.80 | 0.11 | 0.99 | 0.99 | 0.48 | 0.48 | 0.84 | 0.84 | 0.17 | 0.17 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 2354  | 0.81 | 1.07 | 0.74 | 0.07 | 0.98 | 0.98 | 0.98 | 0.98 | 0.48 | 0.48 | 0.07 | 0.07 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 2642  | 0.65 | 0.44 | 1.21 | 0.15 | 0.99 | 0.99 | 0.62 | 0.62 | 1.00 | 1.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 3722  | 0.12 | 0.74 | 0.46 | 0.06 | 1.00 | 1.00 | 0.59 | 0.59 | 0.99 | 0.99 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 4167  | 0.90 | 0.46 | 0.06 | 0.81 | 0.56 | 0.56 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 4328  | 1.18 | 0.13 | 0.56 | 0.18 | 0.99 | 0.99 | 0.27 | 0.27 | 0.24 | 0.24 | 0.16 | 0.16 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 5772  | 1.74 | 1.04 | 0.70 | 0.09 | 0.99 | 0.99 | 0.48 | 0.48 | 0.47 | 0.47 | 0.25 | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 6225  | 0.60 | 1.74 | 0.76 | 0.06 | 0.44 | 0.44 | 0.43 | 0.43 | 0.09 | 0.09 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 6685  | 0.12 | 0.43 | 0.88 | 0.07 | 0.74 | 0.74 | 0.80 | 0.80 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| 7118  | 0.12 | 0.62 | 0.58 | 0.08 | 0.58 | 0.46 | 0.46 | 0.09 | 0.09 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Table 6. The number of genes flagged by the qvalue program as a function of the value of $q_{crit}$.

| $q_{crit}$ | Flagged | $q_{crit}$ | Flagged | $q_{crit}$ | Flagged | $q_{crit}$ | Flagged | $q_{crit}$ | Flagged |
|-----------|--------|-----------|--------|-----------|--------|-----------|--------|-----------|--------|
| $1 \times 10^{-8}$ | 28 | $1 \times 10^{-7}$ | 49 | $1 \times 10^{-6}$ | 87 | $1 \times 10^{-5}$ | 139 |
| $2 \times 10^{-8}$ | 35 | $2 \times 10^{-7}$ | 58 | $2 \times 10^{-6}$ | 99 | $2 \times 10^{-5}$ | 177 |
| $3 \times 10^{-8}$ | 40 | $3 \times 10^{-7}$ | 64 | $3 \times 10^{-6}$ | 111 | $3 \times 10^{-5}$ | 199 |
| $4 \times 10^{-8}$ | 41 | $4 \times 10^{-7}$ | 66 | $4 \times 10^{-6}$ | 117 | $4 \times 10^{-5}$ | 216 |
| $5 \times 10^{-8}$ | 46 | $5 \times 10^{-7}$ | 73 | $5 \times 10^{-6}$ | 123 | $5 \times 10^{-5}$ | 222 |
The performance of the Bayesian method was also compared with the performance of a conventional screening method based on a multiplicity-adjusted hypothesis testing paradigm, a conventional screening method based on the “Higher Criticism” principle, and a Bayesian method. The diagnostic properties were similar, but not identical, although with a suitable choice of design parameters and critical values, the approaches identified nearly identical features as being differentially expressed. This also underscores the potential operating range of Bayesian screening, not primarily developed to control the FDR.

The practical applicability of any screening method depends on its flexibility and computational efficiency. Although Bayesian methods can be computationally demanding, the calculations required by the procedure described here turn out to be very rapid when the assumption of normality is tenable, even for large datasets. The analyses carried out on a simulated set of 100,000 features took about 4 minutes and the analyses carried out on a gene microarray data set with 7128 features took about 30 seconds, both on a standard laptop computer. All of the necessary quantities are calculated from closed-form expressions so that exact solutions are obtained when the assumptions about the data generating model are correct.
The method relies on design parameters that are chosen to incorporate explicitly knowledge and requirements specified by scientists, clinicians, and regulators. Evaluation of suitable ranges of these parameters provides insights into the diagnostic properties, such as sensitivity, MDR and FDR of feature detection for a problem at hand. In this way, the application requirements determine the diagnostic properties in terms that are meaningful to the users. This makes the method flexible and data adaptable and this information can be further leveraged in prospective study planning.

The main objective of feature screening is the identification of potentially informative features as measured by the MDR. Therefore, it can be applied in learn-and-confirm situations. The learning step separates non-informative from potentially informative features. Subsequent analyses, and perhaps trials or experiments, are needed to confirm that these features are truly informative.

The calculations described here have assumed that the feature expressions are independent so that no initial reduction was performed to remove redundancies. However, independence often does not apply, and correlations can affect the sensitivity for detecting important features. Various approaches have been described for incorporating correlations into the analyses, including some of the Bayesian approaches described above. A recently described, computationally convenient deterministic dimension reduction approach (Millstein et al. 2020) implemented in the R package partition (Barrett and Millstein 2020) could be used initially to identify subsets of features that can be combined into new composite features. These possibilities are not pursued further here, since consideration of correlation reduction methods is outside the scope of this paper.

There are several potential future developments of our current work. Developing feature screening with respect to a time-to-event endpoint would be a natural next step because of a strong focus on survival analysis in oncology (Cristescu et al. 2018) and neuroscience biomarker research (Li et al. 2017). Another interesting area to explore would be to evaluate the utility of the Bayesian screening as a step prior to the development of a prediction algorithm (classifier). This would be useful, e.g. in responder/no-responder identification from high-dimensional biomedical data.

Disclosure statement
No potential conflict of interest was reported by the author(s).

Funding
The author(s) reported that there is no funding associated with the work featured in this article.

ORCID
A. Lawrence Gould http://orcid.org/0000-0002-9661-4474

References
Ahdesmäki, M., and K. Strimmer. 2010. Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. Annals of Applied Statistics 4 (1):503–519. doi:10.1214/09-AOAS277.
Arnold, T. B., and J. W. Emerson. 2011. Nonparametric goodness-of-fit tests for discrete null distributions. The R Journal 3 (2):34–39. doi:10.32614/RJ-2011-016.
Barrett, I., R. Mukherjee, and X. Lin. 2017. The generalized higher criticism for testing SNP-set effects in genetic association studies. Journal of the American Statistical Association 112 (517):64–76. doi:10.1080/01621459.2016.1192039.
Barrett, M., and J. Millstein. 2020. partition: A fast and flexible framework for data reduction in R. Journal of Open Source Software 5 (47):1991. doi:10.21105/joss.01991.
Berk, R. H., and D. H. Jones. 1979. Goodness-of-fit test statistics that dominate the Kolmogorov statistics. Zeitschrift Fur Wahrscheinlichkeitstheorie Und Verwandte Gebiete 47 (1):47–59. doi:10.1007/BF00533250.
Cai, T. T., J. S. Jin, and M. G. Low. 2007. Estimation and confidence sets for sparse normal mixtures. *Annals of Statistics* 35 (6):2421–2449. doi:10.1214/009053607000000334.

Cai, T. T., and W. G. Sun. 2017a. Large-scale global and simultaneous inference: Estimation and testing in very high dimensions. *Annual review of economics* 9: 411–439.

Cai, T. T., and W. Sun. 2017b. Optimal screening and discovery of sparse signals with applications to multistage high throughput studies. *Journal of the Royal Statistical Society, Series B* 79 (1):197–223. doi:10.1111/rssb.12171.

Cristescu, R., R. Mogg, M. Ayers, A. Albright, E. Murphy, J. Yearley, X. Sher, X. Q. Liu, H. C. Lu, M. Nebozhyn, et al. 2018. Pan-tumor genomic biomarkers for PD-1 checkpoint blockade-based immunotherapy. *Science* 362 (6411):197–202.

Dadaneh, S. Z., X. N. Qian, and M. Y. Zhou. 2018. BNP-Seq: Bayesian nonparametric differential expression analysis of sequencing count data. *Journal of the American Statistical Association* 113 (521):81–94. doi:10.1080/01621459.2017.1328358.

Dickhaus, T., B. Blankertz, and F. C. Meinecke. 2013. Binary classification with pFDR-pFNR losses. *Biometrical Journal* 55 (3):463–477. doi:10.1002/bimj.201200054.

Do, K. A., P. Muller, and F. Tang. 2005. A Bayesian mixture model for differential gene expression. *Journal of the Royal Statistical Society Series C-Applications Statistics* 54 (3):627–644. doi:10.1111/j.1467-9868.2005.00593.x.

Donoho, D., and J. S. Jin. 2004. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics* 32 (3):962–994. doi:10.1214/009053604000000265.

Donoho, D., and J. S. Jin. 2006. Asymptotic minimaxity of false discovery rate thresholding for sparse exponential data. *Annals of Statistics* 34 (6):2980–3018. doi:10.1214/009053605000000920.

Donoho, D., and J. S. Jin. 2008. “Higher criticism thresholding: Optimal feature selection when useful features are rare and weak”, *Proceedings of the National Academy of Sciences of the United States of America*, 105: 14790–14795.

Donoho, D., and J. Jin. 2009a. Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philosophical Transactions of the Royal Society, Series A* 367 (1906):4449–4470. doi:10.1098/rsta.2009.0129.

Donoho, D., and J. S. Jin. 2009b. Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philosophical Transactions of the Royal Society, Series A* 367 (1906):4449–4470.

Donoho, D., and J. S. Jin. 2015. Higher criticism for large-scale inference, especially for rare and weak effects. *Statistical Science* 30 (1):1–25. doi:10.1214/14-STS506.

Efron, B. 2001. Selection criteria for scatterplot smoothers. *The Annals of Statistics* 29 (2):470–504. doi:10.1214/aoas/1009210549.

Fan, J. Q., F. Han, and H. Liu. 2014. Challenges of big data analysis. *National Science Review* 1 (2):293–314. doi:10.1093/nsr/nwt032.

Fan, J. Q., and J. C. Lv. 2008. Sure Independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 70 (5):849–883. doi:10.1111/j.1467-9868.2008.00674.x.

Fan, J. Q., and J. C. Lv. 2010. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 20 (1):101–148.

Fan, J. Q., and R. Song. 2010. Sure Independence screening in generalized linear models with NP-dimensionality. *Annals of Statistics* 38 (6):3567–3604. doi:10.1214/10-AOS798.

Fisher, C. K., and P. Mehta. 2015a. Bayesian feature selection for high-dimensional linear regression via the Ising approximation with applications to genomics. *Bioinformatics* 31 (11):1754–1761. doi:10.1093/bioinformatics/btv037.

Fisher, C. K., and P. Mehta. 2015b. Bayesian feature selection with strongly regularizing priors maps to the Ising model. *Neural Computation* 27 (11):2411–2422. doi:10.1162/NECO_a_00780.

Genovese, C., and L. Wasserman. 2004. A stochastic process approach to false discovery control. *Annals of Statistics* 32 (3):1035–1061. doi:10.1214/009053604000000283.

Gold, R. T., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al. 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286 (5439):531–537. doi:10.1126/science.286.5439.531.

Gould, A. L. 2008. Detecting potential safety issues in clinical trials by Bayesian screening. *Biometrical Journal* 50 (5):837–851. doi:10.1002/bimj.200710469.

Gould, A. L. 2013. Detecting potential safety issues in large clinical or observational trials by Bayesian screening when event counts arise from Poisson distributions. *Journal of Pharmacological Statistics* 23 (4):829–847. doi:10.1080/10543406.2013.789887.

Gould, A. L. 2018. Unified screening for potential elevated adverse event risk and other associations. *Statistics in Medicine* 37 (18):2667–2689. doi:10.1002/sim.7686.

Ingster, Y. I. 1999. Minimax detection of a signal for $l^p$-balls. *Mathematical Methods of Statistics* 7:401–428.

Jager, L., and J. A. Wellner. 2007. Goodness-of-fit tests via phi-divergences. *Annals of Statistics* 35 (5):2018–2053. doi:10.1214/009053607000000244.

Ji, P. S., and J. S. Jin. 2012. UPS delivers optimal phase diagram in high-dimensional variable selection. *Annals of Statistics* 40 (1):73–103. doi:10.1214/11-AOS947.

Jin, J. S., and Z. T. Ke. 2016. Rare and weak effects in large-scale inference: Methods and phase diagrams. *Statistica Sinica* 26:1–34.
Kamerud, D. B., L. W. Deaton, A. J. Bosch, M. Driscoll, D. Young, E. Gbur, C. A. Goodsell, E. Hertz, G. S. Rogers, M. Skalsky, et al. 1978. Random variable X-Y, X, Y normal. *American Mathematical Monthly* 85:206–208.

Klaus, B., and K. Strimmer. 2013. Signal identification for rare and weak features: Higher criticism or false discovery rates? *Biostatistics* 14 (1):129–143. doi:10.1093/biostatistics/kxx030.

Lai, Y. 2017. A statistical method for the conservative adjustment of false discovery rate (q-value). *Bmc Bioinformatics* 18 (Suppl 3):69. doi:10.1186/s12859-017-1474-6.

Langaa, M., and B. H. Lindqvist. 2005. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society Series B* 67 (4):555–572. doi:10.1111/j.1467-9868.2005.00515.x.

Lewin, A., N. Bochkina, and S. Richardson. 2007. Fully Bayesian mixture model for differential gene expression: Simulations and model checks. *Statistical Applications in Genetics and Molecular Biology* 6 (1). doi:10.2202/1544-6115.1314.

Lewin, A., S. Richardson, C. Marshall, A. Glazier, and T. Aitman. 2006. Bayesian modeling of differential gene expression. *Biometrics* 62 (1):1–9. doi:10.1111/j.1541-0420.2005.00394.x.

Li, K., W. Chan, R. S. Doody, J. Quinn, and S. Luo. 2017. Prediction of conversion to Alzheimer’s disease with longitudinal measures and time-to-event data. *Journal of Alzheimer’s Disease* 58:360–370.

Li, J., and D. Siegmund. 2015. Higher criticism: P-values and criticism. *Annals of Statistics* 43 (3):1323–1350. doi:10.1214/15-AOS1312.

Li, W., D. Xiang, F. Tsung, and P. Xiaolong. 2019. A diagnostic procedure for high-dimensional data streams via missed discovery rate control. *Technometrics* 62:1–27.

Liang, K. 2016. False discovery rate estimation for large-scale homogeneous discrete p-values. *Biometrics* 72 (2):639–648. doi:10.1111/biom.12429.

Liu, Y., M. Morley, J. Brandimarto, et al. 2015. RNA-Seq identifies novel myocardial gene expression signatures of heart failure. *Genomics*. 105(2):83–89. doi:10.1016/j.ygeno.2014.12.002.

Lock, E. F., and D. B. Dunson. 2015. Shared kernel Bayesian screening. *Biometrika* 102 (4):829–842. doi:10.1093/biomet/asv032.

Longford, N. T. 2013. Screening as an application of decision theory. *Statistics in Medicine* 32 (5):849–863. doi:10.1002/sim.5554.

Longford, N. T. 2014. A decision-theoretical alternative to testing many hypotheses. *Biostatistics* 15 (1):154–169. doi:10.1093/biostatistics/kxt030.

Millstein, J., F. Battaglin, M. Barrett, S. Cao, W. Zhang, S. Stintzing, V. Heinemann, and H. J. Lenz. 2020. Partition: A surjective mapping approach for dimensionality reduction. *Bioinformatics* 36 (3):676–681. doi:10.1093/bioinformatics/btz661.

Rue, H., S. Martino, and N. Chopin. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 71 (2):319–392. doi:10.1111/j.1467-9868.2008.00700.x.

Sarkar, S. K., T. Zhou, and D. Ghosh. 2008. A general decision theoretic formulation of procedures controlling FDR and FNR from a Bayesian perspective. *Statistica Sinica* 18:925–945.

Stephens, M. 2017. False discovery rates: A new deal. *Biostatistics* 18 (2):275–294. doi:10.1093/biostatistics/kxx041.

Stephens, M., P. Carbonetto, M. Lu, L. Sun, J. Willwerscheid, and N. Xiao. 2019. ashr: Methods for adaptive shrinkage using empirical Bayes. R package version 2.2-39, https://CRAN.R-project.org/package=ashr.

Storey, J. D. 2002. A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* 64 (3):479–498. doi:10.1111/1467-9868.00346.

Storey, J. D. 2003. The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics* 31 (6):2013–2035. doi:10.1214/aoas/1074290335.

Storey, J. D. 2007. The optimal discovery procedure: A new approach to simultaneous significance testing. *Journal of the Royal Statistical Society Series B* 69 (3):347–368. doi:10.1111/j.1467-9868.2007.005592.x.

Storey, J. D., A. J. Bass, A. Dabney, and D. Robinson. 2019. qvalue: Q-value estimation for false discovery rate control. R package version 2.18.0, http://github.com/jdstorey/qvalue.

Storey, J. D., J. E. Taylor, and D. Siegmund. 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society, Series B* 66 (1):187–205. doi:10.1111/j.1467-9868.2004.00439.x.

Storey, J. D., and R. Tibshirani. 2003. “Statistical significance for genome-wide experiments”, *Proceedings of the National Academy of Sciences*, 100: 9440–9445.

Tayob, N., F. Stingo, K. A. Do, A. S. F. Lok, and Z. D. Feng. 2018. A Bayesian screening approach for hepatocellular carcinoma using multiple longitudinal biomarkers. *Biometrics* 74 (1):249–259. doi:10.1111/biom.12717.

Van De Wiel, M. A., G. G. R. Leday, L. Pardo, H. Rue, A. W. Van Der Vaart, and W. N. Van Wieringen. 2013. Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics* 14 (1):113–128. doi:10.1093/biostatistics/kxs031.

Walther, G. 2013. The average likelihood ratio for large-scale multiple testing and detecting sparse matrices. *Probability to Statistics and Back: High-Dimensional Models and Processes* 9:317–326.
Wehrens, R., and P. Franceschi. 2012a. Meta-Statistics for variable selection: The R package biomark. Journal of Statistical Software 51 (10):1–18. doi:10.18637/jss.v051.i10.

Wehrens, R., and P. Franceschi. 2012b. Thresholding for biomarker selection in multivariate data using Higher Criticism. Molecular Biosystems 8 (9):2339–2346. doi:10.1039/c2mb25121c.

Wellner, J. A., and V. Koltchinskii. 2003. A note on the asymptotic distribution of Berk-Jones Type statistics under the null hypothesis. Progress in Probability 55:321–332.

Wen, X. 2017. Robust Bayesian FDR control using Bayes factors, with applications to multi-tissue eQTL discovery. Statistical Biosciences 9 (1):28–49. doi:10.1007/s12561-016-9153-0.

Wu, W. S., and E. A. Pena. 2013. Bayes multiple decision functions. Electronic Journal of Statistics 7:1272–1300. doi:10.1214/13-EJS813.

Xie, J. C., T. T. Cai, and H. Z. Li. 2011. Sample size and power analysis for sparse signal recovery in genome-wide association studies. Biometrika 98 (2):273–290. doi:10.1093/biomet/asr003.

Xu, L., T. Hanson, E. J. Bedrick, and C. Restrepo. 2010. Hypothesis tests on mixture model components with applications in ecology and agriculture. Journal of Agricultural, Biological, and Environmental Statistics 15 (3):308–326. doi:10.1007/s13253-010-0020-z.

Yu, F., M. H. Chen, and L. Ku. 2008. Detecting differentially expressed genes using calibrated Bayes factors. Statistica Sinica 18:783–802.

Yu, F., M. H. Chen, L. Kuo, H. Talbott, and J. S. Davis. 2015. Confident difference criterion: A new Bayesian differentially expressed gene selection algorithm with applications. Bmc Bioinformatics 16 (1). doi:10.1186/s12859-015-0664-3.

Yu, C., and D. Zelterman. 2017. A parametric model to estimate the proportion from true null using a distribution for p-values. Computational Statistics & Data Analysis 114:105–118. doi:10.1016/j.csda.2017.04.008.

Zhang, H., J. Jin, and W. Zheyang. 2017. Distributions and statistical power of optimal signal-detection methods in finite cases. In. arXiv: 1702.07082v1 [math. 1–37.

Zhang, H., and W. Zheyang 2018. “SetTest: Group testing procedures for signal detection and goodness-of-fit. R package version 0.2.0. https://CRAN.R-project.org/package=SetTest.”

Zhou, T. Y., L. P. Zhu, C. Xu, and R. Z. Li. 2019. Model-free forward screening via cumulative divergence. Journal of the American Statistical Association 115:1393-1405.
A. Outline of Method

The general description that follows extends an earlier development (Gould 2018) by incorporating nuisance parameter(s) and allowing the observed values to be continuous rather than discrete. This approach allows for a wide choice of probabilistic models. Although the calculations could in principle be done numerically, as is often the case for methods described in the literature (e.g., Lock and Dunson 2015; Tayob et al. 2018; Xu et al. 2010), this could be computationally costly if there are many features. Appreciable computational efficiency can be realized by introducing distributional assumptions such as normality that permit direct computation. Details of how this can be done for normally distributed feature values are provided following the general description.

Let $f_i(y_{C_i}; \mu_{C_i}, \eta_i)$ and $f_i(y_{O_i}; \mu_{C_i}, \eta_i)$ denote the likelihoods for the feature $i$ values in the T and C groups (corresponding to treatment and control interventions, respectively), $i = 1, \ldots, N$, and assume that these values are independently distributed among the features. The scalar or vector parameters $\mu_{C_i}$ and $\mu_{O_i}$ may differ between the test and control groups, and there may be other parameters ($\eta_i$) that do not differ between the groups; let $f_{\eta_i}(\eta_i; \zeta)$ denote a prior density for $\eta$. As with conventional Bayesian analyses, suppose that $\mu_{C_i}$ and $\mu_{O_i}$ have prior densities $f_{\mu}(\mu_{C_i}; \theta_{C_i}, \eta_i)$ and $f_{\mu}(\mu_{O_i}; \theta_{O_i}, \eta_i)$, respectively, with unknown parameters $\theta_{C_i}$, $\theta_{O_i}$, and $\eta_i$. The nuisance parameters $\mu_{C_i}$, $\mu_{O_i}$, and $\eta_i$ are removed by integrating the product of the likelihood and prior densities to yield the marginal densities of the observations, $g_i(y_{C_i}; \theta_i, \eta_i; \theta_i, \zeta)$.

The testable (Arnold and Emerson 2011) assumption that the feature value distributions for the control group are exchangeable implies that $\theta_{C_i} = \theta_{C}$. The test group parameters $\theta_{O_i}$ are not independent of the features because the features may be affected by the test intervention in different ways. Suppose that $\theta_{O_i}$ can be written as a random mixture

$$\theta_{O_i} = (1 - y_i)\theta_{C_i}^{(0)} + y_i\theta_{C_i}^{(T)}$$

(A1)

The mixture indicators $y_i$ have Bernoulli probability functions with common parameter $\pi_0$,

$$p(y_i; \pi_0) = \pi_0^{1-y_i}(1 - \pi_0)^y_i,$$

and suppose that the prior density of $\pi_0$ is $f_{\pi_0}(\pi_0; \xi)$. Conventional integrations, summations, and algebraic rearrangements provide expressions for the posterior density of $\pi_0$,

$$f_{\pi_0}(\pi_0; \xi, y_{C}, y_{T}, \zeta) = H^{-1}\pi_0^N f_{\pi_0}(\pi_0; \xi) \prod_{i=1}^{N} \left\{ 1 + \frac{1 - \pi_0}{\pi_0} g_i(y_{T_i}; \theta_{C_i}^{(T)}) \right\}$$

(A2)

where

$$H = \prod_{i=1}^{N} g_i(y_{T_i}; \theta_{C_i}^{(T)}; \zeta) \int_{\pi_0} \pi_0^N f_{\pi_0}(\pi_0; \xi) \prod_{i=1}^{N} \left\{ 1 + \frac{1 - \pi_0}{\pi_0} g_i(y_{T_i}; \theta_{C_i}^{(T)}) \right\} d\pi_0$$

The conditional (on $\pi_0$) posterior probability function of $y_i$ is

$$P_{\text{post}}(Y_i; \pi_0, \theta_{C_i}^{(C)}, \theta_{C_i}^{(T)}, y_{C}, y_{T}, \zeta) = \left\{ 1 - y_i + \frac{\pi_0}{1 - \pi_0} g_i(y_{T_i}; \theta_{C_i}^{(T)}) \right\} \frac{1}{1 - \pi_0} \frac{1 - \pi_0}{\pi_0} g_i(y_{T_i}; \theta_{C_i}^{(T)})$$

(A3)

The expectation of (A3) with respect to (A2) provides a corresponding ‘unconditional’ posterior probability. This approach incorporates $\pi_0$ as an explicit model parameter. A number of strategies for determining the value of $\pi_0$ without the assumption of a prior distribution have been described in the hypothesis-testing literature (Genovese and Wasserman 2004; Lai 2017; Langaa and Lindqvist 2005; Liang 2016; Storey 2003; Wen 2017; Yu and Zelterman 2017).

Low values of (A3) [or its unconditional version] are useful for identifying the features that are affected by the test treatment differently from the control. However, they do not provide a perspective on the magnitude of the difference, and it is possible to have a low posterior probability that $y_i = 0$ with only a modest test-control difference. This perspective can be provided by considering the posterior distribution of an appropriate metric (say $M$) for the test-control difference. Suppose that $\tilde{\theta}_i (\tilde{\theta}_{C_i} \text{ or } \tilde{\theta}_{O_i})$ denotes the parameter of the posterior distribution of $\theta_i$, and that $M$ has cdf $F_M$. Then, the conditional posterior cdf of $M$ given $y_i$ and the values of $\tilde{\theta}_{C_i}$ or $\tilde{\theta}_{O_i}$ (which depend on the observed feature values $x_{C_i}$ and $x_{T_i}$) is

$$P_{\text{post}}(M_i \leq m; y_i, \tilde{\theta}_{C_i}, \tilde{\theta}_{O_i}) = (1 - y_i) F_M(m; \tilde{\theta}_{C_i}, \tilde{\theta}_{O_i}) + y_i F_M(m; \tilde{\theta}_{C_i}, \tilde{\theta}_{O_i}),$$
The unconditional posterior cdf of $M$ is

$$
P_{\text{post}}(M_i \leq m; \tilde{\theta}_C, \tilde{\theta}_T) = P_{\text{post}}(y_i = 0; \xi, \theta^{(C)}_T, \theta^{(T)}_T, \mathbf{x}_C, \mathbf{x}_T, \zeta) F_M(m; \tilde{\theta}_C, \tilde{\theta}_T) + P_{\text{post}}(y_i = 1; \xi, \theta^{(C)}_T, \theta^{(T)}_T, \mathbf{x}_C, \mathbf{x}_T, \zeta) F_M(m; \tilde{\theta}_C, \tilde{\theta}_T)$$

### B. Normally Distributed Feature Values

The observed mean feature values $y_{C_i}$ and $y_{T_i}$ are assumed to have normal likelihoods with respective means $\mu_{C_i}$ and $\mu_{T_i}$ and precisions $\kappa_{C_i} \eta_i$ and $\kappa_{T_i} \eta_i$ ($\eta_i = 1/\sigma^2_i$ where $\sigma^2_i$ denotes the feature-specific residual variance). $a_{C_i}$ and $a_{T_i}$ refer (usually) to the sample sizes, $i = 1, \ldots, N_{\text{feat}}$. The precisions may differ among the features (although exchangeability of the precisions among features is assumed), but not between the test and control interventions within a feature. The corresponding observed residual sums of squares are denoted by $u_{C_i}$ and $u_{T_i}$; the quantities $\eta_i u_{C_i}$ and $\eta_i u_{T_i}$ are assumed to have chi-square likelihoods with respective degrees of freedom $m_{C_i}$ and $m_{T_i}$, $i = 1, \ldots, N_{\text{feat}}$. The means $\theta_{C_i}$ and $\theta_{T_i}$ have normal prior distributions with respective expected values $\mu_{C_i} = \mu_{C}$ and $\mu_{T_i}$, common precision $b_0 \eta_0$; and (because of exchangeability) the precision parameters $\eta_i$ are assumed to have a gamma prior distribution with parameters $\zeta_i/2$ and $\zeta_i/2$. Expression (A1) describes the mixture attribute of the process. How the values of $\theta^{(C)}_T$ and $\theta^{(T)}_T$ are determined is key to the screening process. For the moment, these values are taken as given. Posterior inferences about the $\theta_i$ are a main objective of the screening process.

The calculations are based on the product of the likelihoods and prior densities for feature $i$, $i = 1, \ldots, N$

$$f_{\text{norm}}(y_{C_i}; \mu_{C_i}, a_{C_i} \eta_i) f_{\text{norm}}(y_{T_i}; \mu_{T_i}, a_{T_i} \eta_i) f_{\text{norm}}(\mu_{C_i}; \theta_{C}, b_0 \eta_0) f_{\text{norm}}(\mu_{T_i}; \theta_{T}, b_0 \eta_0)$$

$$\times f_{\text{gamma}}\left(u_{C_i} m_{C_i} \eta_i / 2, \frac{\eta_i u_{C_i}}{2} \right) f_{\text{gamma}}\left(u_{T_i} m_{T_i} \eta_i / 2, \frac{\eta_i u_{T_i}}{2} \right) f_{\text{gamma}}\left(\eta_i \zeta_i / 2, \frac{\eta_i \zeta_i}{2} \right)$$

(A4)

where

$$f_{\text{norm}}(y; \mu, \eta) = \frac{1}{\sqrt{2\pi \eta}} e^{-\frac{(y-\mu)^2}{2\eta}}$$

and

$$f_{\text{gamma}}(u; \frac{\alpha}{2}, \frac{\beta}{2}) = \frac{u^{\alpha-1} e^{-\beta u / 2}}{\Gamma(\alpha)}$$

After straightforward algebraic rearrangements and integrations to remove nuisance parameters, (A4) reduces to

$$f\left(y_{C_i}, y_{T_i}, u_i; \theta_{C}, \theta^{(C)}_T, \theta^{(T)}_T, \eta_i, m_i, a_{C_i}, a_{T_i}, b_0, \zeta_i \right)$$

$$= f_{\text{beta}}\left(v_i; \frac{m_i + \zeta_i}{2}, \frac{\zeta_i}{2} \right) f_1(W_i; m_i + \zeta_i) f_t\left(Z^{(C)}_i; m_i + \zeta_i + 1 \right)$$

$$\times \left(1 - y_i \right) + \frac{\eta_i}{\eta_i + 1} f_t\left(Z^{(T)}_i; m_i + \zeta_i + 1 \right)$$

where $f_t(x; v) = v^{-1} B^{-1}(v/2, 1/2)(1 + x^2/v)^{-v/2}$ denotes a central $t$ density and

$$Z^{(C)}_i = \left(\frac{a_{C_i} b_0 (m_i + \zeta_i + 1)}{(a_{C_i} + b_0) (a_{C_i} + m_i + \zeta_i + 1)}\right)^{1/2} \left(y_{C_i} - \theta^{(C)}_C\right) Z^{(T)}_i = \left(\frac{a_{C_i} b_0 (m_i + \zeta_i + 1)}{(a_{C_i} + b_0) (a_{C_i} + m_i + \zeta_i + 1)}\right)^{1/2} \left(y_{T_i} - \theta^{(T)}_T\right)$$

$$W_i = \sqrt{\frac{a_{C_i} b_0 (m_i + \zeta_i)}{(a_{C_i} + b_0) (a_{C_i} + m_i + \zeta_i)}} \left(x_i - \theta_C\right)$$

If $m_i$ is “large” (e.g., $> 30$), then the central $t$ densities in (A5) can be replaced with standard normal densities. The posterior density of $\eta_0$ is given by (A2) which, in this case, has the form

$$f\left(\eta_0; y_{C_i}, y_{T_i}, u_i; \theta_{C}, \theta^{(C)}_T, \theta^{(T)}_T, \zeta_i, \zeta\right)$$

$$= H^{-1} f_{\text{gamma}}(\eta_0; \zeta) \prod_{i=1}^{N_{\text{feat}}} \left\{ 1 + \frac{1 - \eta_0}{\eta_0} f_t\left(Z^{(C)}_i; m_i + \zeta_i + 1\right) \right\}$$

(A6)

where

$$H = f_t^t f_{\text{gamma}}(\eta_0; \zeta) \prod_{i=1}^{N_{\text{feat}}} \left\{ 1 + \frac{1 - \eta_0}{\eta_0} f_t\left(Z^{(C)}_i; m_i + \zeta_i + 1\right) \right\} d\eta_0$$

Consequently, the conditional posterior density (A3) becomes

$$P_{\text{post}}\left(y_i; y_{C_i}, y_{T_i}, u_i; \theta_{C}, \theta^{(C)}_T, \theta^{(T)}_T, \eta_0, \zeta\right)$$

$$= \left\{ 1 + \frac{1 - \eta_0}{\eta_0} f_t\left(Z^{(C)}_i; m_i + \zeta_i + 1\right) \right\}^{-1} \left\{ \frac{\eta_0}{\eta_0} \left(1 - y_i\right) + \frac{\eta_i}{\eta_i + 1} f_t\left(Z^{(C)}_i; m_i + \zeta_i + 1\right) \right\}$$

(A7)
Expression (A7), which is related to the local fdr (Efron 2001), often occurs in the multiple comparison literature as the statistic for testing null hypotheses against alternatives (Cai and Sun 2017b; Dickhaus et al. 2013; He, Sarkar, and Zhou 2015; Sarkar et al. 2008; Storey 2007; Wen 2017). The expectation of (A6) with respect to the posterior density of \( \pi_0 \) provides the unconditional posterior probability function of \( \pi_i \),

\[
\pi_{SP\pi} = P_{\pi_0} \left( \gamma_i = \theta_i \right) = \int_{\pi_0 = 0}^{1} f \left( \gamma_i | y, y_T, u, \theta_C, \theta_T^{(C)}, \theta_T^{(T)}, \xi, \zeta \right) \pi_0 \, d\pi_0
\]

The posterior densities of the \( \gamma_i \) and \( \pi_0 \) parameters depend on the distribution of the \( x_{iY} \) values, and not on the distribution of the \( x_{C} \) values. Consequently, the calculation for any feature can be conditional on the response of feature \( i \) to the control treatment so that, in particular, \( \theta_C \) in all these expressions can be replaced with \( \theta_C \), the vector of posterior expectations of \( \theta_{C_i} \), \( i = 1, \ldots, N \).

None of the calculations just described require simulation. They are either direct evaluation of explicit expressions or low-dimensional numerical integration. Consequently, they are practical to apply when there are many features.

### C. Magnitude of Difference

Low values of (A8) for the \( i \)-th feature are useful for identifying features where the response may be higher on the test treatment than on the control. However, they do not provide a perspective on the magnitude of the response difference (or ratio), and it is possible to have a low posterior probability that \( \gamma_i = 0 \) with only a modest increase in the expected response to the test treatment.

The posterior distribution of the actual or standardized difference between the expected feature responses to a test and a control treatment will often be of particular interest. For normally distributed responses, the difference \( D_i = \mu_{Ti} - \mu_{Ci} \) has a normal distribution whose mean depends, given the observed responses of feature \( i \) to the control treatment, on whether the same process did \( \gamma_i = 0 \) or did not \( \gamma_i = 1 \) generate the expected response for feature \( i \),

\[
\mu_{D_i} = \frac{s_{\gamma_i} \bar{y}_i - b \gamma_i}{a_{\gamma_i}}
\]

and precision \( \bar{a}_{D_i} = \frac{s_{\gamma_i} \bar{a}}{a_{\gamma_i} + s_{\gamma_i} \bar{a}} \). Integration of the product of the density of \( D_i \) and the posterior density of \( \eta_i \) (a gamma density with parameters \( (m_i + \zeta_i)/2 \) \( (\zeta_i + u_i)/2 \)) provides the marginal posterior density of \( D_i \) or, more precisely of the quantity

\[
Z_{D_i} = \sqrt{\frac{\bar{a}_{D_i}}{(m_i + \zeta_i)(\zeta_i + u_i)}} D_i - \mu_{D_i}
\]

which has a central \( t \) distribution with \( m_i + \zeta_i \) degrees of freedom. Consequently,

\[
P(D_i < d | \gamma_i, \delta) = P \left\{ \sqrt{\bar{a}_{D_i}} (D_i - \mu_{D_i}) < \sqrt{\bar{a}_{D_i}} (d - \mu_{D_i}) \right\} = F_{\gamma_i} (\tau_i) (d - \mu_{D_i}) / (m_i + \zeta_i)
\]

where \( \tau_i = \sqrt{\bar{a}_{D_i}} (m_i + \zeta_i) (\zeta_i + u_i) \), expresses the posterior cdf of the difference between \( \mu_{Ti} \) and \( \mu_{Ci} \) given the data and the value of \( \gamma_i \) and \( \delta \). This same expression provides the posterior cdf of the standardized difference between \( \mu_{Ti} \) and \( \mu_{Ci} \) (\( D^* = D_i / s(d_i) \)) when \( d \) in the above expression is replaced with

\[
d^* = d_i
\]

The unconditional posterior cdf of the difference can be obtained easily as the weighted average,

\[
P(D_i < d) = p_{SP\pi} P(D_i < d | \gamma_i = 0, \delta = 0) + (1 - p_{SP\pi}) P(D_i < d | \gamma_i = 1, \delta = 0)
\]

with a similar expression for the standardized difference. If the \( m_i \) values are large, then the values of \( \zeta_i \) and \( \zeta_i \) can be ignored and, the central \( t \) cdf can be replaced with a standard normal cdf.

### D. Prior Distributions for Metrics

The difference between the responses of feature \( i \) to the test and control treatments could be expressed in terms of the arithmetic difference \( D_i = \mu_{Ti} - \mu_{Ci} \) or the ratio \( R_i = \mu_{Ti} / \mu_{Ci} \).
1. Differences

Under the normality assumptions used here, the quantity
\[ Z^{(0)}_D = \sqrt{\frac{b_0}{2\hat{\lambda}}}(D_i - \hat{\delta}) = \hat{\lambda}(D_i - \hat{\delta}_i) \]

has a central t density with \( \hat{\delta}_i \) degrees of freedom. The residual sums of squares (the \( u_{ti} \) or \( u_i \) values) on which the screening process is conditioned provides a particularly convenient way to specify the ratio \( \hat{\delta}/\hat{\delta}_i \), the \textit{a priori} expected value of \( \eta_i \). The expected value of \( u_{ti}^{-1} \) is \( \eta_i/(m_i - 2) \), so that \( E(\eta_i) \) can be estimated by the average of \((m_i - 2)/u_i \) over the features,
\[ \bar{U} = \frac{N_{\text{crit}}}{\sum N_{\text{crit}}}(m_i - 2)/u_i = \bar{E}(\eta_i) \]

The calculations that follow replace \( \lambda \) with \( \hat{\lambda} = \sqrt{b_0 \bar{U}/2} \).

If \( \Psi() \) denotes a standard normal or central t cdf, then requirement (1) in the text, namely that \( P(D \leq d_{\text{crit}} \mid \delta = 0) > \Phi_0 \), determines the value of \( b_0 \) as
\[ b_0 = 2(\Psi^{-1}(\Phi_0)/d_{\text{crit}})^2/\bar{U}. \]

The value of \( \hat{\delta}_i \) is determined by requirement (2) and can be expressed simply as
\[ \hat{\delta}_i = \delta = d_{\text{crit}}\{1 - \Psi^{-1}(\Phi_0)\Psi^{-1}(\Phi_0)\}. \]

Table A1 provides typical values for \( d_{\text{crit}} \times \lambda = \Psi^{-1}(\Phi_0) \), and \( \delta/d_{\text{crit}} \) as functions of \( \Phi_0 \) and \( \Phi_1 \) when \( \Psi \) denotes a normal cdf and when \( \Psi \) denotes the cdf of a t distribution with 8 degrees of freedom.

| \( \Psi^{-1}(\Phi_0) \) | Normal | \( \Psi^{-1}(\Phi_0) \) | \( \Phi_0 \) | \( \Phi_1 \) | \( \Phi_1 \) | \( \Phi_0 \) | \( \Phi_1 \) | \( \Phi_0 \) |
|----------------|--------|-----------------|----------|--------|--------|----------|--------|----------|
| 0.0            | 1.28   | 1.64            | 2.33     | 0.1    | 2.00   | 1.75     | 1.48    | 1.30     | 1.29     | 1.19     |
| 0.2            | 1.66   | 1.51            | 1.36     | 0.2    | 1.64   | 1.48     | 1.31    | 1.39     | 1.29     | 1.19     |
| 0.3            | 1.41   | 1.32            | 1.23     | 0.3    | 1.39   | 1.29     | 1.19    | 1.39     | 1.29     | 1.19     |
| 0.4            | 1.20   | 1.15            | 1.11     | 0.4    | 1.19   | 1.14     | 1.09    | 1.39     | 1.29     | 1.19     |

2. Ratios

When \( \mu_{ti}^{(C)} \) is unlikely to be negative, the statement \( \mu_{ti}^{(T)} / \mu_{ti}^{(C)} < r_{\text{crit}} \)" is essentially the same as the statement \( \mu_{ti}^{(T)} - r_{\text{crit}} \mu_{ti}^{(C)} < 0 \), so that simple expressions can be obtained for probability statements about the ratio \( \mu_{ti}^{(T)}/\mu_{ti}^{(C)} \). The quantity \( W_i = \mu_{ti}^{(T)} - r_{\text{crit}} \mu_{ti}^{(C)} \) has a normal distribution in this case with mean \( \kappa = \mu_{ti}^{(T)} - r_{\text{crit}} \mu_{ti}^{(C)} \) and precision \( \eta_i/(1 + r_{\text{crit}}^2 \eta_i) \). Consequently,
\[ P\left( \frac{\mu_{ti}^{(T)}}{\mu_{ti}^{(C)}} < r_{\text{crit}} \mid \rho_i \right) = \Psi\left( \frac{\kappa}{\eta_i} \sqrt{\frac{r_{\text{crit}}^2}{1 + r_{\text{crit}}^2} (r_{\text{crit}} - \rho_i)} \right) \]

where \( \rho_i = \Theta_{ti}^{(C)}/\Theta_{ti}^{(T)} \) and, here, \( \Psi \) denotes a standard normal cdf. Requirement (1) implies that
\[ \kappa_i = \Theta_{ti}^{(C)} \sqrt{b_0 \eta_i} = \frac{\Psi^{-1}(\Phi_0) \sqrt{1 + r_{\text{crit}}^2} (r_{\text{crit}} - \rho_i)}{r_{\text{crit}} - 1} \]

Requirement (2) implies that
\[ \rho_i = r_{\text{crit}} - (r_{\text{crit}} - 1)\Psi^{-1}(\Phi_1) \Psi^{-1}(\Phi_0) \]

The quantity \( \eta_i \) can be replaced with its \textit{a priori} expectation \( \hat{\eta}_i \), which can be estimated as described above, in which case \( \Psi \) denotes a central t cdf. Table A2 provides typical values for \( k \) and \( \rho \) as functions of \( \Phi_0 \) and \( \Phi_1 \) when \( \Psi \) denotes a normal cdf for \( r_{\text{crit}} = 2 \) and 4.
Table A2.

Values of $k$ and of $\rho$ satisfying requirements (1) and (2) assuming a normal approximation

|       | $r_{\text{crit}} = 2$ |       | $r_{\text{crit}} = 4$ |
|-------|------------------------|-------|------------------------|
| $\phi_0$ | 0.9  | 0.95 | 0.99 | $\phi_0$ | 0.9  | 0.95 | 0.99 |
| $k$ | 2.87 | 3.68 | 5.20 | $k$ | 1.76 | 2.26 | 3.20 |
| 0.01 | 3.00 | 2.78 | 2.55 | 0.01 | 7.00 | 6.34 | 5.65 |
| 0.05 | 2.66 | 2.51 | 2.36 | 0.05 | 5.97 | 5.54 | 5.09 |
| 0.1  | 2.41 | 2.32 | 2.23 | 0.1  | 5.23 | 4.96 | 4.68 |
| 0.2  | 2.20 | 2.15 | 2.11 | 0.2  | 4.59 | 4.46 | 4.33 |

When $\mu_{\text{Ti}}^{(C)}$ could be negative, the probability statements are not simple. Following the development by Kamerud et al. (1978), the density function for the ratio $Z = \mu_{\text{Ti}}^{(T)}/\mu_{\text{Ti}}^{(C)}$ has the form

$$
\begin{align*}

f_Z(z; \mu_{\text{Ti}}, \mu_{\text{C}}, \eta) &= \frac{\eta(z^2)}{\pi(z^2)} e^{-\frac{\eta(z^2)}{\pi(z^2)}} + \sqrt{\pi} e^{-1/2} e^{-\frac{(\mu_{\text{Ti}} - \mu_{\text{C}})^2}{2(z^2)}} \left\{ 1 - 2\Phi(-v) \right\}
\end{align*}
$$

where $v = \sqrt{\frac{\eta}{\pi z^2}}(2\mu_{\text{Ti}} + \mu_{\text{C}})$, and values of the corresponding cdf need to be determined numerically.