Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2012)
74, *Part* 3, *pp.* 515–541

Adjusted Bayesian inference for selected parameters

Daniel Yekutieli

Tel Aviv University, Israel

[Received May 2010. Final revision September 2011]

Summary. We address the problem of providing inference from a Bayesian perspective for parameters selected after viewing the data. We present a Bayesian framework for providing inference for selected parameters, based on the observation that providing Bayesian inference for selected parameters is a truncated data problem. We show that if the prior for the parameter is non-informative, or if the parameter is a 'fixed' unknown constant, then it is necessary to adjust the Bayesian inference for selection. Our second contribution is the introduction of Bayesian false discovery rate controlling methodology, which generalizes existing Bayesian false discovery rate methods that are only defined in the two-group mixture model. We illustrate our results by applying them to simulated data and data from a microarray experiment.

Keywords: Bayesian false discovery rate; Directional decisions; False discovery rate; Selection bias; Selective inference

1. Introduction

We discuss providing Bayesian inference for parameters selected after viewing the data. Current thought is that selection has no effect on the inference of parameters from a Bayesian perspective. We show that this is not necessarily so. Consider generating a sample from a Bayesian framework by randomly generating the parameter and conditionally on the parameter data are generated. In one case, selection is applied to samples of the parameter and the data, and in the other case the parameter is sampled and then selection is applied to data samples. The example below shows that selection matters in the latter case, but not in the former case.

1.1. Example 1

Let θ denote students' true academic ability. The marginal density of θ in the population of high school students is $N(0, 1)$. The observed academic ability of students in high school is $Y \sim N(\theta, 1)$, and students with $0 < Y$ are admitted to college. We wish to predict a student's true academic ability from his observed academic ability—but only if the student is admitted to college. We shall show that the Bayesian inference is different for a random high school student from for a random college student.

We first consider the case of a college professor predicting θ for a student in his class. The joint distribution of (θ, Y) for a random college student can be generated by generating (θ, y) for a random high school student and selecting (θ, y) only if $0 < y$. Thus the joint density of (θ, y) that is used for predicting θ is

Address for correspondence: Daniel Yekutieli, Department of Statistics and Operations Research, Tel Aviv University, Tel Aviv 69978, Israel.
E-mail: yekutiel@post.tau.ac.il

 1369–7412/12/74515

$$f_S(\theta, y) \propto \exp\left(-\frac{\theta^2}{2}\right) \exp\left\{-\frac{(\theta-y)^2}{2}\right\} \Big/ \Pr(Y>0) \propto \exp\left\{-\frac{(\theta-y/2)^2}{2(\frac{1}{2})}\right\}, \tag{1}$$

and the conditional distribution of θ given $Y=y$ is $N(y/2, \frac{1}{2})$. The predicted academic ability for a student with $y=1$ is $E(\theta|y=1)=0.5$.

For the case of the high school teacher predicting θ for a student in his class, we assume that there is a high school regulation instructing teachers to predict academic ability only for students who can be admitted to college. This means that, for any true academic ability θ, the values of Y that are used to predict θ are drawn from the $N(\theta, 1)$ density truncated by the event $0<Y$. Since θ for a random student is $N(0, 1)$, the joint density of (θ, y) that is used for predicting θ is

$$f_S(\theta, y) \propto \exp\left(-\frac{\theta^2}{2}\right) \exp\left\{-\frac{(\theta-y)^2}{2}\right\} \Big/ \Pr(Y>0|\theta). \tag{2}$$

In this case there is no closed expression for the conditional distribution of θ given $Y=y$, but since $\Pr(Y>0|\theta)$ decreases in θ then it is stochastically smaller than $N(y/2, \frac{1}{2})$, and the predicted academic ability for a student with $y=1$ is $E(\theta|y=1)=0.10$.

In this paper, we address selection that arises in the statistical analysis of large data sets in which the aim is to find interesting parameters and then provide inferences for these selected parameters. Throughout the paper we use the following simulated example to illustrate the discussion. One can consider it as an example of a microarray experiment in which θ_i is the log-fold change in expression of gene i and Y_i is the observed log-expression-ratio. We shall now show that, even when the selection is applied to the parameter and the data, it is necessary to correct Bayesian inference for selection if the prior on the parameter is non-informative.

1.2. Example 2

The simulation includes 10^5 independent identically distributed (IID) samples of (θ_i, Y_i). To generate θ_i, we first sample λ_i from $\{10, 1\}$ with probabilities 0.90 and 0.10, and then draw θ_i from the Laplace distribution, $\pi_1(\theta_i|\lambda_i) = \lambda_i \exp(-\lambda_i|\theta_i|)/2$. Thus the marginal distribution of θ_i is

$$\pi(\theta_i) = 0.9\,\pi_1(\theta_i|\lambda_i=10) + 0.1\,\pi_1(\theta_i|\lambda_i=1). \tag{3}$$

$Y_i = \theta_i + \varepsilon_i$, with ε_i independent $N(0, 1)$.

In our analysis we apply the level $q=0.2$ Benjamini and Hochberg (1995) false discovery rate FDR controlling procedure to the two sided p-values, $p_i = 2\{1-\Phi(|Y_i|)\}$, to find interesting θ_i, and then construct 0.95 credible intervals for each interesting θ_i. The Benjamini and Hochberg (BH) procedure yielded $R=932$ discoveries ($p_{(932)} = 0.001862 < 0.001864 = 0.2 \times 932/10^5$): the set of θ_i with $|Y_i| > 3.111$. The 932 selected (θ_i, Y_i) are displayed in Fig. 1.

We use two prior models for constructing credible intervals for θ_i. In the first model the prior distribution for θ_i is $\pi(\theta_i)$ in equation (3). In this case the posterior distribution of θ_i (we derive it later in the paper) is the conditional distribution of θ_i given Y_i. Thus the probability that θ_i is in the 0.95 credible interval constructed for it is, per definition, 0.95. Furthermore, since selection is applied to θ_i and Y_i, selection should have no effect on the Bayesian inference. And, indeed, 0.953 of the selected θ_i (888 out of 932) are covered by their respective 0.95 credible intervals.

In the second model we assume that the marginal distribution of θ_i is unknown and we

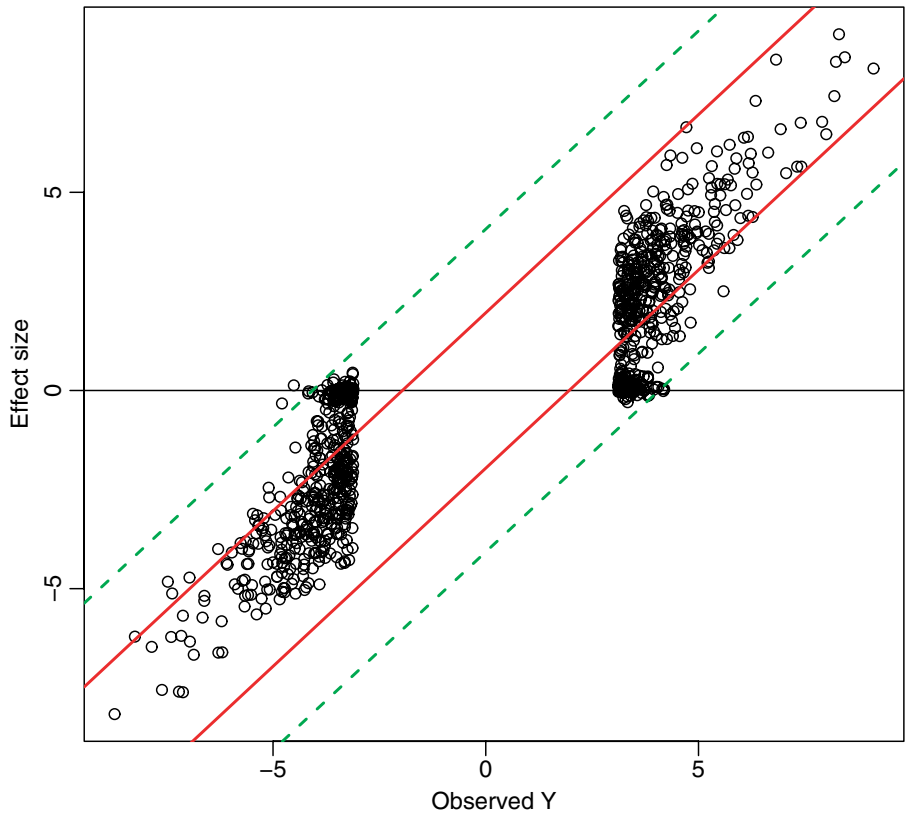

Fig. 1. Simulated example—scatter plot of $|Y_i| > 3.111$ (Y_i-values are drawn on the abscissa of the plot; the ordinates are θ_i-values): ———, marginal 0.95 confidence intervals; – – –, 0.05 false coverage statement rate adjusted confidence intervals

replace it with the non-informative prior $\pi(\theta_i) = 1$. The posterior distribution of θ_i for this prior distribution is $N(Y_i, 1)$. Thus $Y_i \pm Z_{1-0.05/2}$ is a 0.95 credible interval for θ_i (these are the full lines in Fig. 1). Even though the posterior distribution for non-informative priors is not the conditional distribution of the parameters given the data, these are equal tail credible intervals based on minimally informative priors that are known to provide good frequentist performance (Carlin and Louis (1996), section 4.3) that are expected to cover approximately 0.95 of the θ_i. These credible intervals cover 0.951 of all 100000 θ_i, but only 0.654 of the selected θ_i (610 out of 932).

Before presenting our inferential framework in Section 1.9, we review a frequentist approach for discovering interesting parameters and providing inferences for these discoveries in Section 1.3. In Section 1.5 we further motivate the importance of our problem by reviewing literature on providing inference for interesting parameters in genomic studies. In Sections 1.6–1.8 several aspects of Bayesian analysis that are relevant to our work are reviewed.

1.3. Control over the false coverage statement rate

Soric (1989) asserted that the goal of many scientific experiments is to discover non-zero effects and as a result made the important observation that it is mainly the discoveries that are reported

and included in science, and warned that unless the proportion of false discoveries in the set of declared discoveries is kept small there is danger that a large part of science is untrue.

Benjamini and Hochberg (1995) considered the problem of testing m null hypotheses $H_1 \ldots H_m$, of which m_0 are true null hypotheses. They referred to the rejection of a null hypothesis as a discovery and the rejection of a true null hypothesis as a false discovery. To limit the occurrence of false discoveries when testing multiple null hypotheses they introduced the false discovery rate $\mathrm{FDR} = E\{V/\max(R, 1)\}$, where R is the number of discoveries and V is the number of false discoveries, and introduced the BH multiple-testing procedure that controls FDR at a nominal level q.

Benjamini and Yekutieli (2005) generalized the BH testing framework. In their selective inference framework there are m parameters $\theta_1 \ldots \theta_m$, with corresponding estimators $T_1 \ldots T_m$, and the goal is to construct valid marginal confidence intervals (CIs) for the subset of parameters that are selected by a given selection rule $\mathcal{S}(t_1 \ldots t_m) \subseteq \{1 \ldots m\}$. They showed that CIs constructed for selected parameters no longer ensure the nominal coverage probability and suggested the false coverage statement rate FCR as the appropriate criterion to capture the error for CIs constructed for selected parameters. FCR is also defined by $E\{V/\max(R, 1)\}$; however, R is the number of CIs constructed and V is the number of non-covering CIs. Benjamini and Yekutieli (2005) introduced a method of ensuring $\mathrm{FCR} \leqslant q$ for independent $T_1 \ldots T_m$ and any selection criterion: construct marginal $1 - Rq/m$ CIs for each of the R selected parameters. In cases where each θ_i can be associated with a null value θ_i^0 and the selection criteria are multiple-testing procedures that test $\theta_i = \theta_i^0$ *versus* $\theta_i \neq \theta_i^0$, Benjamini and Yekutieli (2005) showed that the level q BH procedure can be expressed as the least conservative multiple-testing procedure that ensures that all level q FCR-adjusted CIs for θ_i, for which the null hypothesis is rejected, will not cover the respective θ_i^0. Furthermore, they showed that for independent $T_1 \ldots T_m$ if all $\theta_i \neq \theta_i^0$ then applying the level q BH procedure to select the parameters and declaring each selected θ_i greater than θ_i^0 if $T_i > \theta_i^0$ and smaller than θ_i^0 if $T_i < \theta_i^0$ controls the directional FDR (the expected proportion of selected parameters assigned the wrong sign) at level $q/2$.

1.4. Example 3

In example 2 all $\theta_i \neq 0$; thus for any multiple-testing procedure $\mathrm{FDR} \equiv 0$. However, declaring θ_i positive for the BH discoveries with $Y_i > 0$ and negative for the BH discoveries with $Y_i < 0$ ensures directional FDR less than 0.1. The number of simulated positive selected θ_i with negative Y_i and negative selected θ_i with positive Y_i is 56; thus the observed directional FDR is 0.060.

The full lines in Fig. 1 are two-sided normal 0.95 CIs: $Y_i \pm Z_{1-0.05/2}$ (recall that these are also the non-informative prior 0.95 credible intervals from example 2). These 0.95 CIs cover 95089 of the 100000 simulated θ_i, but only 610 of the 932 selected θ_i; thus the observed FCR is 0.346. The broken lines are 0.05-FCR-adjusted CIs: $Y_i \pm Z_{1-0.05\times 932/(2\times 10^5)}$. The observed FCR for the FCR-adjusted CIs is 0.046.

1.5. Selective inference in genomic association studies

The need to correct inference for selection is widely recognized in genomewide association studies, which typically test association between a disease and hundreds of thousands of markers located throughout the human genome, often expressed as an odds ratio of manifesting the disease in carriers of a risk allele. Only multiplicity-adjusted significant findings are reported. This limits the occurrence of false positive results; however, it introduces bias into the odds ratio estimates. Analysing 301 published studies covering 25 different reported associations, Lohmueller *et al.* (2003) found that for 24 associations the odds ratio in the first positive report exceeded the genetic effect that is estimated by meta-analysis of the remaining studies. Zollner

and Pritchard (2007) suggested correcting for the selection bias by providing point estimates and CIs based on the likelihood conditionally on having observed a significant association. Zhong and Prentice (2008) further assumed that in the absence of selection the log-odds-ratio estimator is normally distributed. Similarly to our Bayesian analysis of the simulated example, they based their inference on a truncated normal conditional likelihood.

1.6. Parameter selection in Bayesian analysis

Berry and Hochberg (1999) commented that the Bayesian treatment of the multiplicity problem also includes decision analysis, rather than just finding posterior distributions.

Scott and Berger (2006) discussed Bayesian analysis of microarray data. The prior model for θ_i, the expectation of the log-fold change in expression of gene i, is that $\theta_i = 0$ with probability p and $\theta_i \sim N(0, V)$ with probability $1 - p$. The decision analysis that was performed in Scott and Berger (2006) is the discovery of the subset of active genes. Scott and Berger (2006) declared a gene active ($\theta_i \neq 0$) if the posterior expected loss of this action is smaller than the posterior expected loss of declaring the gene inactive ($\theta_i = 0$). The loss function for deciding that $\theta_i = 0$ is proportional to $|\theta_i|$, and the loss for erroneously deciding that $\theta_i \neq 0$ is the fixed cost of doing a targeted experiment to verify that the gene is in fact active.

The decision analysis in Bayesian FDR-analysis of microarray data is also deciding which genes are active. In Efron *et al.* (2001), θ_i is selected if its local FDR, which is the posterior probability given y_i that $\theta_i = 0$, is less than a nominal value q. Storey (2002, 2003) suggested specifying selection rules for which the positive FDR, pFDR, which is defined as the conditional probability that $\theta_i = 0$ given that θ_i is selected, is less than q. In the optimal discovery procedure that was suggested in Storey (2007), the statistic that was used for specifying the selection rule is a plug-in estimator of the local FDR. Storey (2007) showed that the optimal discovery procedure provides the maximal probability of selecting θ_i among all selection rules with the same pFDR-level.

1.7. Selection bias in Bayesian analysis

Dawid (1994) explained why selection should have no effect on Bayesian inference:

> 'Since Bayesian posterior distributions are already fully conditioned on the data, the posterior distribution of any quantity is the same, whether it was chosen in advance or selected in the light of the data'.

Senn (2008) reviewed the disparity between Bayesian and frequentist approaches regarding selection. He considered the example of providing inference for θ_{i^*}, which is the effect of the pharmaceutical associated with the largest sample mean y_{i^*}, among a class of m compounds with $Y_i \sim N(\theta_i, 4)$. He first showed that if θ_i are IID $N(0, 1)$ the posterior distribution of θ_{i^*} is $N(y_{i^*}/5, 4/5)$. He then assumed a hierarchical model in which the treatments form a compound class. The class effect is $\lambda \sim N(0, 1 - \gamma^2)$ and θ_i are IID $N(\lambda, \gamma^2)$. In this case he showed that the posterior distribution of θ_{i^*} depends on the number of other compounds and their overall mean; however, it is unaffected by the fact that θ_{i^*} was selected because it corresponds to the largest sample mean.

The observation that Bayesian inference may be affected by selection was already made in Mandel and Rinott (2007, 2009). Mandel and Rinott (2007) considered the scenario of providing inference for p, which is the probability of success in a binomial experiment, conditionally on observing two or more successes. Similarly to example 1, they distinguished between the case that in each binomial experiment p is drawn independently from its prior distribution and the

case that the value of p is the same in all binomial experiments, and they showed that in the second case the Bayesian inference is affected by selection.

1.8. Fixed and random effects in Bayesian analysis

In the Bayesian framework there can be no fixed effects since the parameters are regarded as having probability distributions. However, discussing one-way classification Box and Tiao (1992), section 7.2, used the sampling theory terminology of fixed and random effects to distinguish between situations in which the individual means can be regarded as distinct values that are expected to bear no strong relationship to each other that can take on values anywhere within a wide range, and situations in which the individual means can be regarded as draws from a distribution. Box and Tiao illustrated this distinction with the example of one-way classification of several groups of laboratory yields. In the first case the groups correspond to different methods of making a particular chemical product, whereas in the second case the groups correspond to different batches made by the same method. The distinction only carries through to the prior model that is elicited for the group means. In the first case the group means are elicited flat non-informative priors. They called this model the fixed effect model. In the second case the group means are IID $N(\lambda, \sigma^2)$. This model is called the random-effect model.

1.9. Preliminary definitions

Let θ denote the parameter, Y denote the data and Ω is the sample space of Y. $\pi(\theta)$ is the prior distribution of θ, and $f(y|\theta)$ is the likelihood function. The multiple parameters for which inference may or may not be provided are actually multiple functions of θ: $h_1(\theta), h_2(\theta), \ldots$. In selective inference for each $h_i(\theta)$ there is a subset $S^i_\Omega \subseteq \Omega$, such that inference is provided for $h_i(\theta)$ only if $y \in S^i_\Omega$ is observed. For example, in our analysis of microarray data in Section 6, Y is the entire set of observed gene expression levels; $\theta = (\sigma^2, \mu)$ consists of the variances and expectations of the log-expression levels for all the genes in the array, and inference is provided for $h_g(\theta) = \mu_g$, the expectation of the log-fold change in expression of gene $g = 1, \ldots, G$, only if gene g is declared differentially expressed by the BH procedure.

Control over FCR is a frequentist mechanism for providing selective inference. Note that in example 2 a randomly selected θ_i is covered by its FCR-adjusted CI with probability 0.95 or greater. But this frequentist selective inference mechanism suffers from several intrinsic limitations: it is impossible to incorporate prior information on the parameters; it does not provide selection-adjusted point estimates or selection-adjusted inference for functions of the parameters; the selection adjustment is the same regardless of the selection criterion applied and the value of the estimator. Fig. 1 suggests that the selection adjustment needed shrinks the CIs towards 0, rather than just widening the CIs, and the larger $|Y_i|$ the smaller selection adjustment is needed for θ_i.

In selective inference the entire data set $Y = y$ is observed. However, as inference is provided for $h_i(\theta)$ only if $y \in S^i_\Omega$, then $Y = y$ used for providing selective inference for $h_i(\theta)$ is actually a realization of the joint distribution of (θ, Y), truncated by the event that $y \in S^i_\Omega$ (describing Bayesian selective inference a truncation problem was suggested by Bradley Efron in private communication; for a discussion on truncation see Mandel (2007) and Gelman *et al.* (2004), section 7.8). Thus to provide Bayesian selective inference for $h_i(\theta)$ we introduce a framework for providing Bayesian inference based on the truncated distribution of (θ, Y). We call this inference selection-adjusted Bayesian inference.

Predicting true academic ability from observed academic ability for a high school student and for a college student, which was discussed in example 1, are Bayesian selective inference

problems in which inference is provided for $h(\theta) = \theta$ only if $S_\Omega = \{y : 0 < y\}$ occurs. Even though the selection mechanism is different, in both cases, (θ, y) for which θ is predicted from y are truncated samples from the distribution of (θ, Y) in the population of all high school students.

1.10. Outline of the paper

In Section 2 we discuss modelling selection-adjusted Bayesian inference: we provide an operative definition for the joint truncated distribution of (θ, Y); we distinguish between parameters according to the way that their distribution is affected by selection and derive the joint truncated distribution of (θ, Y) in either case; for either case, and also for parameters with non-informative priors, we define the components (i.e. prior, likelihood and posterior) of selection-adjusted Bayes inference; we then specifically derive these components for (θ, Y) that correspond to Box and Tiao's random-effect model and fixed effect model. In Section 3 we define selection-adjusted Bayes inference as the Bayes rules in Bayesian selective inference. We also present a Bayesian FCR for the random-effect model and explain the relationship between selection-adjusted Bayes inference and providing FCR-control.

In Section 4 we present Bayesian FDR controlling methodology for specifying selection rules in the random-effects model for cases in which selection is used for making statistical discoveries. We also provide an empirical Bayes algorithm for applying this methodology in cases that correspond to Box and Tiao's fixed effect model. In Section 5 we explain the relationship between the Bayesian FDR methods that are presented in Section 4 and existing Bayesian FDR methods, and describe how to provide selection-adjusted Bayes inference in the two-group mixture model.

In Section 6 we analyse microarray data. The goal of the analysis is to find overexpressed and underexpressed genes while controlling directional FDR $\leqslant 0.05$, and to provide inference for the change in expression for these selected genes. The level 0.10 BH procedure applied to t-statistic p-values fails to discover any differentially expressed genes. Applying the level 0.10 BH procedure to p-values corresponding to hybrid classical–Bayes moderated t-statistics yields 245 discoveries; however, it is not clear how to provide frequentist selective inference for these discoveries. For comparison, our level 0.05 Bayesian FDR selection rule based on the moderated t-statistic yields 1124 discoveries, and the level 0.05 Bayesian FDR selection rule based on the optimal statistic yields 1271 discoveries. In the second part of the analysis, we provide Bayesian selective inference for the expected base 2 log-fold change in expression for a differentially expressed gene.

The paper concludes with a discussion of the conceptual and methodological contributions of this paper.

2. Modelling selection-adjusted Bayesian inference

The primary problem in modelling selection-adjusted Bayes inference is specifying the joint truncated distribution of (θ, Y), which we denote $f_S(\theta, y)$. It is important to note that $f_S(\theta, y)$ is the joint distribution of (θ, Y) according to which selective inference is provided for $h(\theta)$, and not the joint distribution of (θ, Y), $f(\theta, Y) = \pi(\theta)\, f(y|\theta)$. We use this characterization for defining $f_S(\theta, y)$.

Definition 1. Assume that selective inference for $h(\theta)$ involves an action $\delta(Y)$ that is associated with a loss function $L\{h(\theta), \delta\}$. $f_S(\theta, y)$ is defined as the distribution over which the expected loss

$$r_S(\delta) = \int_\theta \int_{y \in S_\Omega} f_S(\theta, y)\, L\{h(\theta), \delta(y)\}\, \mathrm{d}y\, \mathrm{d}\theta \tag{4}$$

is the average risk incurred in selective inference for $h(\theta)$.

2.1. 'Fixed', 'random' and 'mixed' parameters in Bayesian selective inference

Example 1 illustrated that $f_S(\theta, y)$ is determined by the way that selection acts on θ. Unlike Box and Tiao who used the terms fixed and random effects to describe the type of prior distribution elicited for θ, we use the terms fixed, random and mixed parameters to describe the way that the distribution of θ is affected by selection. For each parameter type, we derive $f_S(\theta, y)$, $\pi_S(\theta)$, the marginal truncated distribution of θ, and $f_S(y|\theta)$, the truncated conditional distribution of $Y|\theta$.

2.1.1. *Fixed parameter truncated sampling model*

We call θ a fixed parameter if its distribution is unaffected by selection and selection is applied to the conditional distribution of Y given θ. Fixed parameters are unknown constants whose values are assumed to be sampled from $\pi(\theta)$ and remain unchanged. Thus, for each value of θ, the risk that is incurred in providing selective inference for $h(\theta)$ is the expected loss over the truncated conditional distribution of $Y|\theta$

$$\int_{y \in S_\Omega} f(y|\theta)/\Pr(S_\Omega|\theta)\, L\{h(\theta), \delta(y)\}\, \mathrm{d}y,$$

for $\Pr(S_\Omega|\theta) = \int_{y \in S_\Omega} f(y|\theta)\, \mathrm{d}y$, and the average risk is its expectation over the marginal density of θ,

$$r_S(\delta) = \int_\theta \int_{y \in S_\Omega} \pi(\theta)\, f(y|\theta)/\Pr(S_\Omega|\theta)\, L\{h(\theta), \delta(y)\}\, \mathrm{d}y\, \mathrm{d}\theta. \tag{5}$$

Thus in this case the joint truncated distribution of (θ, Y) is

$$f_S(\theta, y) = I_{S_\Omega}(y)\, \pi(\theta)\, f(y|\theta)/\Pr(S_\Omega|\theta), \tag{6}$$

the marginal truncated density of θ is

$$\pi_S(\theta) = \pi(\theta) \tag{7}$$

and the truncated conditional distribution of $Y|\theta$ is

$$f_S(y|\theta) = I_{S_\Omega}(y)\, f(y|\theta)/\Pr(S_\Omega|\theta). \tag{8}$$

2.1.2. *Random-parameter truncated sampling model*

We call θ a random parameter in cases where selection is applied to the joint distribution of (θ, Y). In this case θ is drawn from $\pi(\theta)$ and Y is drawn from $f(y|\theta)$, but inference is provided for $h(\theta)$ only for (θ, y) with $y \in S_\Omega$. Thus the average risk incurred in providing selective inference $h(\theta)$ is

$$r_S(\delta) = \int_\theta \int_{y \in S_\Omega} \pi(\theta)\, f(y|\theta)/\Pr(S_\Omega)\, L\{h(\theta), \delta(y)\}\, \mathrm{d}y\, \mathrm{d}\theta, \tag{9}$$

for $\Pr(S_\Omega) = \int_\theta \int_{y \in S_\Omega} \pi(\theta)\, f(y|\theta)\, \mathrm{d}y$. Thus the truncated distribution of (θ, Y) is

$$f_S(\theta, y) = I_{S_\Omega}(y)\, \pi(\theta)\, f(y|\theta)/\Pr(S_\Omega). \tag{10}$$

Integrating out y yields the marginal truncated distribution of θ

$$\pi_S(\theta) = \pi(\theta)\Pr(S_\Omega|\theta)/\Pr(S_\Omega). \tag{11}$$

Dividing equation (10) by equation (11) reveals that in this case the truncated distribution of $Y|\theta$ is also the conditional likelihood in equation (8).

2.1.3. Mixed parameter truncated sampling model

We call θ a mixed parameter in cases where selection is applied to the conditional distribution of (θ, Y) given λ, for a hyperparameter $\lambda \sim \pi_2(\lambda)$ with $\theta|\lambda \sim \pi_1(\theta|\lambda)$. Thus conditioning on λ, θ is 'random' and the average risk that is incurred in providing selective inference is

$$\int_\theta \int_{y\in S_\Omega} \pi_1(\theta|\lambda)\, f(y|\theta)/\Pr(S_\Omega|\lambda)\, L\{h(\theta), \delta(y)\}\, \mathrm{d}y\, \mathrm{d}\theta, \tag{12}$$

where $\Pr(S_\Omega|\lambda) = \int_\theta \int_{y\in S_\Omega} \pi_1(\theta|\lambda)\, f(y|\theta)\, \mathrm{d}y\, \mathrm{d}\theta$. Taking expectation over λ yields the average risk

$$r_S(\delta) = \int_\lambda \int_\theta \int_{y\in S_\Omega} \pi_2(\lambda)\, \frac{\pi_1(\theta|\lambda)\, f(y|\theta)}{\Pr(S_\Omega|\lambda)}\, L\{h(\theta), \delta(y)\}\, \mathrm{d}y\, \mathrm{d}\theta\, \mathrm{d}\lambda. \tag{13}$$

Thus the truncated density of (λ, θ, y) is

$$f_S(\lambda, \theta, y) = I_{S_\Omega}(y)\, \pi_2(\lambda)\, \pi_1(\theta|\lambda)\, f(y|\theta)/\Pr(S_\Omega|\lambda). \tag{14}$$

Changing the order of integration in equation (13) we obtain

$$r_S(\delta) = \int_\theta \int_{y\in S_\Omega} \left\{ \int_\lambda \frac{\pi_2(\lambda)\, \pi_1(\theta|\lambda)}{\Pr(S_\Omega|\lambda)}\, \mathrm{d}\lambda \right\} f(y|\theta)\, L\{h(\theta), \delta(y)\}\, \mathrm{d}y\, \mathrm{d}\theta, \tag{15}$$

and thus the truncated density of (θ, y) is

$$f_S(\theta, y) = I_{S_\Omega}(y) f(y|\theta) \int_\lambda \pi_2(\lambda)\, \pi_1(\theta|\lambda)/\Pr(S_\Omega|\lambda)\, \mathrm{d}\lambda. \tag{16}$$

Integrating out y yields the marginal truncated distribution of θ

$$\pi_S(\theta) = \Pr(S_\Omega|\theta) \int_\lambda \frac{\pi_2(\lambda)\, \pi_1(\theta|\lambda)}{\Pr(S_\Omega|\lambda)} \mathrm{d}\lambda. \tag{17}$$

And again, dividing equation (16) by equation (17) reveals that the truncated distribution of $Y|\theta$ is $f_S(y|\theta)$ in equation (8).

Remark 1. It is important to note that classifying θ as a fixed, random or mixed parameter is context dependent and must be done case by case. In example 1, θ is an unknown constant, for both a random college student and a random high school student. However, comparing expressions (1) and (2) with (6) and (10) reveals that θ is a fixed parameter for a random high school student, and a random parameter for a random college student.

Senn's (2008) example of providing inference for the most active compound can be expressed as a selective inference problem in which, for $i = 1, \ldots, m$, inference is provided for $h_i(\theta) = \theta_i$ only if $S^i_\Omega = \{y : y_i = \max(y_1, \ldots, y_m)\}$ occurs. When θ is the vector of treatment effects of m distinct compounds, each component of θ is a distinct unknown constant whose value is sampled from $N(\lambda, \gamma^2)$ and remains unchanged; therefore θ is a fixed parameter. Now suppose that $\theta_i \sim N(\lambda, \gamma_i)$ are batch effects of m batches treated by a single compound, with compound effect $\lambda \sim N(0, 1-\gamma^2)$. In this case, λ is a fixed unknown constant and, conditional on λ, θ is a random batch effect. Thus θ is a mixed parameter.

2.2. Defining the components of Bayesian selective inference

The selection-adjusted prior distribution is, when it is available, the marginal truncated distribution of θ. We have shown that the selection-adjusted prior distribution for fixed, random or mixed θ is $\pi_S(\theta)$ given in equations (7), (11) or (17). To specify the marginal truncated distribution of θ, we need $\pi(\theta)$ to be the marginal distribution of θ and we need to know how selection acts on θ.

An important case in which $\pi(\theta)$ is not the marginal distribution of θ is when $\pi(\theta)$ is a non-informative prior distribution. Non-informative priors are used to allow conditional analysis on θ when no prior information on θ is available (Berger (1985), section 3.3.1). As Y also provides all the information on θ in the truncated data problem, we argue that the prior distribution that is used for selection-adjusted Bayes inference should also be a non-informative prior. We further argue that whereas the lack of prior knowledge on θ may affect our decision to provide selective inference, the opposite is not true—the decision to provide inference for only certain values of Y should have no effect on the non-informative prior that is elicited for θ. We therefore suggest using the same non-informative prior for selection-adjusted Bayes inference, $\pi_S(\theta) = \pi(\theta)$, which means that if the prior for θ is non-informative then it is treated as a fixed parameter.

The selection-adjusted likelihood is $f_S(y|\theta)$ in equation (8), the truncated conditional distribution of Y given θ. Note that conditioning on θ ensures that the selection-adjusted likelihood is the same in the three truncated sampling models and does not depend on the marginal distribution of θ.

The selection-adjusted posterior distribution is defined by

$$\pi_S(\theta|y) = \pi_S(\theta)\, f_S(y|\theta)/m_S(y), \tag{18}$$

for $m_S(y) = \int \pi_S(\theta)\, f_S(y|\theta)\, \mathrm{d}\theta$. For non-informative priors it is generated by updating the non-informative prior according to the selection-adjusted likelihood. For fixed, random or mixed θ it is the truncated conditional distribution of $\theta|Y$. Thus $\pi_S(\theta|y) \propto f_S(\theta, y)$. But note that only for random θ, for which $f_S(\theta, y) \propto f(\theta, y)$, is the selection-adjusted posterior distribution unaffected by selection.

Remark 2. Dawid (1994) argued that selection has no effect on posterior distributions since conditioning on the selection event is made redundant by conditioning on $Y = y$. Note that this applies only to the case of random θ, for which selection can be expressed as conditioning on an event S in the sample space of (θ, Y). Hence, as Dawid argued, for $(\theta, y) \in S$ the truncated posterior distribution is the same as the untruncated posterior distribution:

$$\pi_S(\theta|y) = \pi(\theta|S, Y = y) = \frac{f(\theta, S, Y = y)}{f(S, Y = y)} = \frac{f(\theta, Y = y)}{f(Y = y)} = \pi(\theta|Y = y) = \pi(\theta|y),$$

whereas for fixed and mixed θ, for which selection cannot generally be expressed as conditioning on an event in the sample space of (θ, Y), $\pi_S(\theta|y)$ is generally different from $\pi(\theta|y)$ as demonstrated in example 1 and in example 4. We illustrate how this point applies to our simulated data in example 5 later.

2.3. Example 4

Senn (2008) concluded that selection has no effect on the Bayesian inference because in his analysis θ is a random parameter. In remark 1 we suggest that in this kind of analysis θ will most probably be a fixed or a mixed parameter. We therefore compute the selection-adjusted posterior mean of $h_2(\theta) = \theta_2$ for $m = 2$ and $y = (0, 2)$, for mixed and fixed θ.

However, as $S^2_\Omega = \{(\theta, y) : y_2 \geqslant y_1\}$, then $\Pr(S^2_\Omega|\lambda) \equiv \Pr(S^2_\Omega) = 0.5$, and the mixed parameter model truncated joint density that is defined in equation (16) reduces to the random-parameter joint density in equation (10). Thus in this case, also for mixed θ, the conditional distribution of θ_2 is unaffected by selection. We use expression (4) in Senn (2008) to compute the conditional mean of θ_2. For $\gamma^2 = 1$ it equals 0.4 and for $\gamma^2 = 0.5$ it equals 0.384.

The selection-adjusted joint density of θ for fixed θ is given by

$$\pi_S\{\theta_1, \theta_2|y = (0, 2)\} \propto \exp\left(-\frac{\lambda^2}{2\gamma^2}\right) \exp\left\{-\frac{(\theta_1 - \lambda)^2}{2(1-\gamma^2)}\right\} \exp\left\{-\frac{(\theta_2 - \lambda)^2}{2(1-\gamma^2)}\right\} \times \exp\left\{-\frac{(0-\theta_1)^2}{2\times 4}\right\} \exp\left\{-\frac{(2-\theta_2)^2}{2\times 4}\right\} \Big/ \Pr(Y_2 \geqslant Y_1|\theta_1, \theta_2).$$

In this case the selection adjustment increases the posterior distribution of θ-values with $\theta_2 < \theta_1$, thereby stochastically decreasing the marginal posterior distribution of θ_2. For $\gamma^2 = 1$ the conditional mean of θ_2 is 0.164 and for $\gamma^2 = 0.5$ it is 0.257.

2.4. Modelling Bayesian selective inference in the random-effect model

Using the terminology that was suggested by Box and Tiao, we call the model for $\theta = (\theta_1 \ldots \theta_m)$ and $Y = \{Y_1 \ldots Y_m\}$, where θ_i are IID $\pi(\theta_i)$ and $Y_i|\theta_i$ are independent $f(y_i|\theta_i)$, a random-effect model.

In the random-effect model θ can be a random parameter, a fixed parameter and even a mixed parameter when there are IID fixed λ_i for which $\theta_i|\lambda_i$ are independent random parameters. In any case the joint distribution of (θ, Y) is

$$f(\theta, y) = \pi(\theta)\, f(y|\theta) = \prod_{i=1}^{m} \pi(\theta_i) \prod_{i=1}^{m} f(y_i|\theta_i). \tag{19}$$

In selective inference for $h_i(\theta) = \theta_i$ with $S^i_\Omega = \{y : y_i \in S_{\text{marg}}\}$, incorporating equation (19) into equation (6) yields the fixed θ selection-adjusted joint distribution of (θ, Y)

$$\begin{aligned} f_S(\theta, y) &= I_{S^i_\Omega}(y) \prod_{j=1}^{m} \pi(\theta_j)\, f(y_j|\theta_j) / \Pr(S^i_\Omega|\theta) \\ &= \prod_{j\neq i} \{\pi(\theta_j) f(y_j|\theta_j)\}\, I_{S_{\text{marg}}}(y_i)\, \pi(\theta_i)\, f(y_i|\theta_i) / \Pr(Y_i \in S_{\text{marg}}|\theta_i). \end{aligned} \tag{20}$$

Integrating out $\theta^{(i)}$ and $y^{(i)}$ in equation (20) yields the selection-adjusted distribution of (θ_i, Y_i) for fixed θ

$$f_S(\theta_i, y_i) = I_{S_{\text{marg}}}(y_i)\, \pi(\theta_i)\, f(y_i|\theta_i) / \Pr(Y_i \in S_{\text{marg}}|\theta_i). \tag{21}$$

Similarly, incorporating equation (19) into equation (10) and integrating out $\theta^{(i)}$ and $y^{(i)}$ yields the selection-adjusted joint distribution of (θ_i, Y_i) for random θ

$$f_S(\theta_i, y_i) = I_{S_{\text{marg}}}(y_i)\, \pi(\theta_i)\, f(y_i|\theta_i) / \Pr(Y_i \in S_{\text{marg}}). \tag{22}$$

Incorporating equation (19) into equation (16) and integrating out $\theta^{(i)}$ and $y^{(i)}$ yields the mixed θ selection-adjusted distribution of (θ_i, Y_i)

$$f_S(\theta_i, y_i) = I_{S_{\text{marg}}}(y_i)\, f(y_i|\theta_i) \int \frac{\pi_2(\lambda_i)\, \pi_1(\theta_i|\lambda_i)}{\Pr(Y_i \in S_{\text{marg}}|\lambda_i)}\, \mathrm{d}\lambda_i. \tag{23}$$

2.4.1. Non-exchangeable random-effect model

The non-exchangeable random-effect model is a generalization of the random-effect model for situations in which θ_i are distinct values that are expected to bear no strong relationship one to

each other, i.e. situations for which Box and Tiao would suggest the fixed effect model. In the non-exchangeable random-effect model θ_i are independent but have distinct prior distributions $\pi^i(\theta_i)$, whereas $Y_i|\theta_i$ are still independent $f(y_i|\theta_i)$. Thus the joint distribution of (θ, Y) is

$$f(\theta, y) = \pi(\theta)\, f(y|\theta) = \prod_{i=1}^{m} \pi^i(\theta_i) \prod_{i=1}^{m} f(y_i|\theta_i). \tag{24}$$

The marginal distribution of (θ_i, Y_i) is

$$f(\theta_i, y_i) = \pi^i(\theta_i)\, f(y_i|\theta_i).$$

But, in selective inference for $h_i(\theta) = \theta_i$ with $S^i_\Omega = \{y : y_i \in S_{\text{marg}}\}$, the selection-adjusted joint distribution of (θ_i, Y_i) for fixed θ is

$$f_S(\theta_i, y_i) = I_{S_{\text{marg}}}(y_i)\, \pi^i(\theta_i)\, f(y_i|\theta_i) / \Pr(Y_i \in S_{\text{marg}}|\theta_i). \tag{25}$$

2.5. Example 5

Note that (θ, Y) in example 2 are generated by the random-effect model, that the components of $\theta = (\theta_1 \ldots \theta_{100000})$ are independently drawn from $\pi(\theta_i)$ in equation (3) and that $Y_i|\theta_i$ are independent $f(y_i|\theta_i) = \phi(y_i - \theta_i)$. Fig. 1 is a scatter plot of 932 (θ_i, y_i) with $|y_i| > 3.111$; Fig. 2. displays the 470 components with $y_i > 3.111$. For comparison, in the comparable non-exchangeable

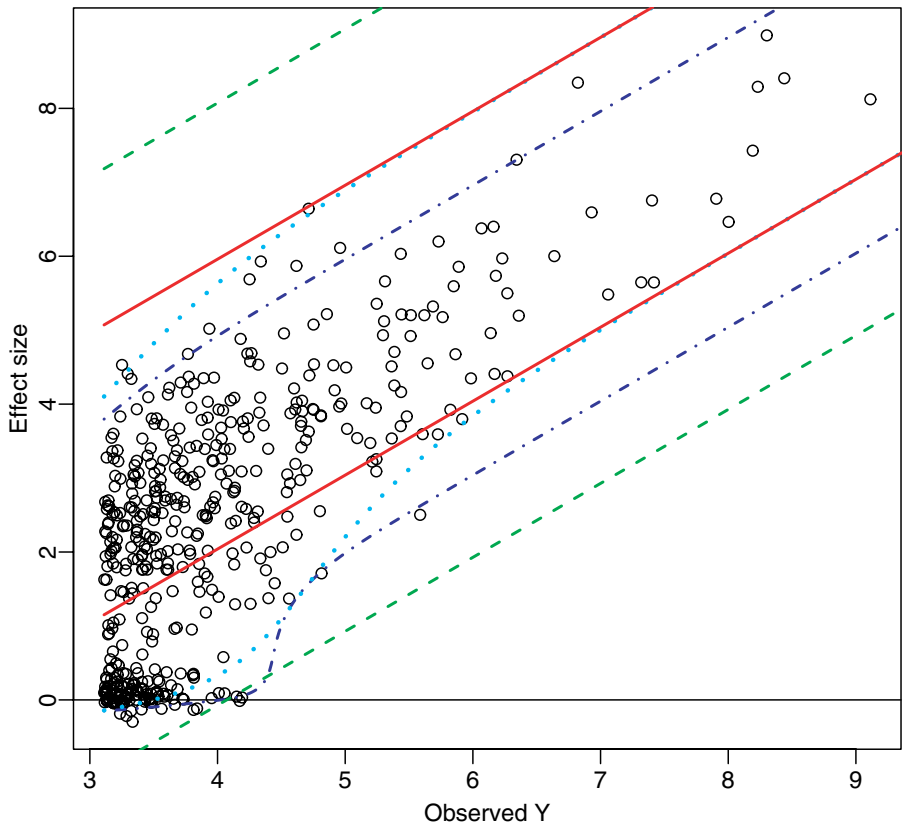

Fig. 2. Simulated example—scatter plot of $Y_1 > 3.111$ components: ——, – – –, CIs from Fig. 1; – · – · –, random-parameter model selection-adjusted Bayes 0.95 credible intervals; · · · · ·, non-informative prior selection-adjusted Bayes 0.95 credible intervals

random-effect model: for $i = 1, \ldots, 90000$, $\theta_i \sim \pi_1(\theta_i|\lambda_i = 10)$ and, for $i = 90001, \ldots, 100000$, $\theta_i \sim \pi_1(\theta_i|\lambda_i = 1)$.

It is important to note that in example 2 we draw a single realization from the joint untruncated distribution of (θ, Y). To observe the difference between random, fixed and mixed θ we conduct another set of simulations, in which we sample 1000 realizations of (θ, Y) from its truncated distributions for $h_1(\theta) = \theta_1$ with $S_\Omega^1 = \{y : |y_1| > 3.111\}$ for random, fixed and mixed θ. Each realization from the random θ truncated distribution is generated by repeatedly sampling (θ, Y) from its untruncated distribution, keeping the first (θ, y) for which $|y_1| > 3.111$. To generate each realization from the fixed θ truncated distribution, we sample θ from $\pi(\theta)$ and then repeatedly sample Y, keeping the first y with $|y_1| > 3.111$. As the components of (θ, Y) are independent the distribution of $(\theta_2, \ldots, \theta_{100000}, Y_2, \ldots, Y_{100000})$ is the same in the three truncation models.

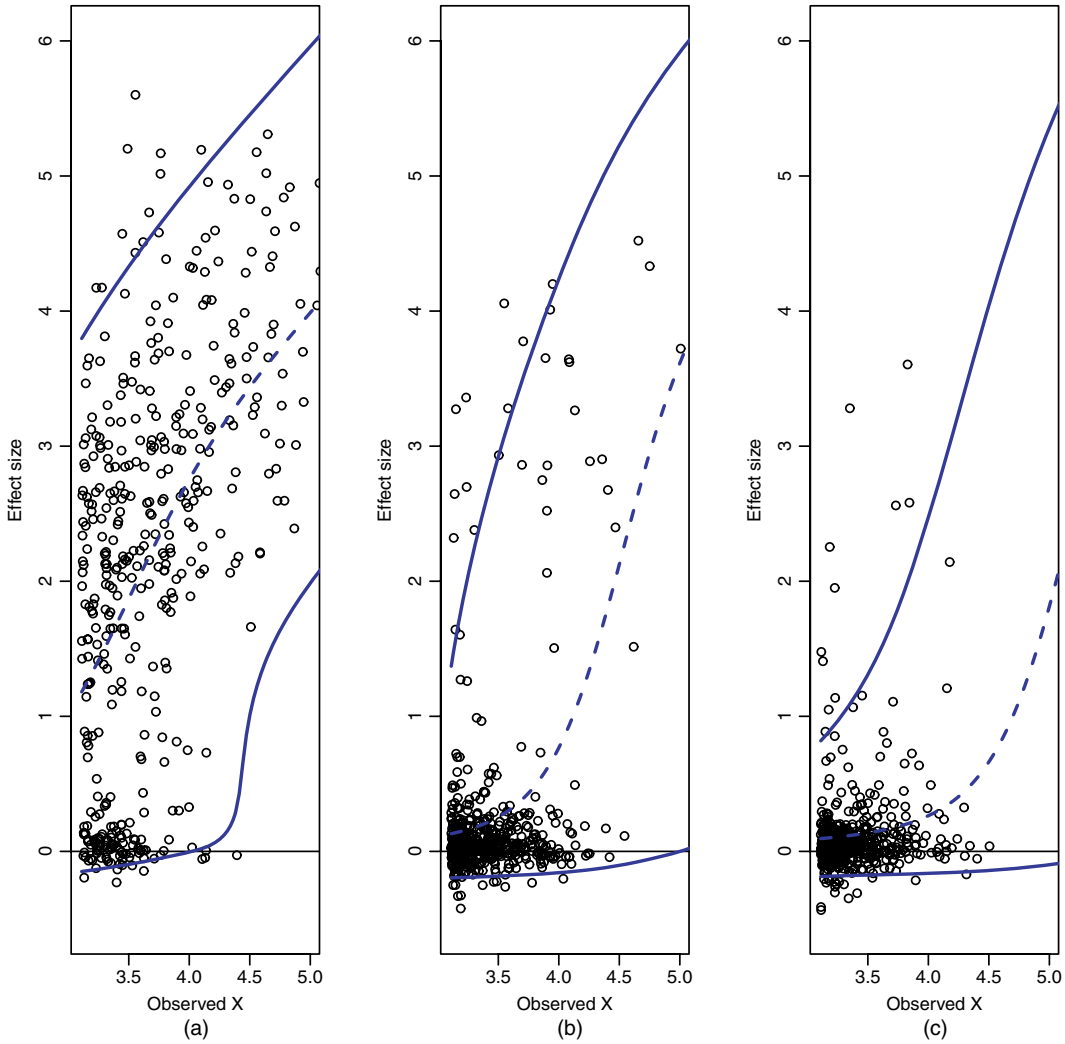

Fig. 3. Simulated example—scatter plot of $Y_1 > 3.111$ realizations of (θ_1, Y_1) in (a) the random-parameter truncated sampling model (466 observations), (b) the mixed parameter truncated sampling model (498 observations) and (c) the fixed parameter truncated sampling model (501 observations): ——, selection-adjusted 0.95 posterior credible intervals for θ_1; – – –, selection-adjusted posterior means

Fig. 3 displays the scatter plots of the $Y_1 > 3.111$ realizations of (θ_1, Y_1) for each truncation model. Fig. 3(a) is the scatter plot for the random-θ model. In this case the joint density of (θ_1, Y_1), which is given in equation (22), is

$$\pi(\theta_1)\,\phi(y_1 - \theta_1),$$

and it is identical to the joint density of (θ_i, Y_i) that is displayed in Fig. 2 and the distribution of (θ_i, Y_i) for $Y_i > 3.111$ in Fig. 1. Fig. 3(c) is the scatter plot for the fixed θ model. In this case the joint density of (θ_1, Y_1), which is given in equation (21), is

$$\pi(\theta_1)\,\phi(y_1 - \theta_1)/\Pr(|Y_1| > 3.111\,|\theta_1).$$

Comparing Figs 3(a) and 3(c) reveals that in this model, for each value of Y_1, the conditional distribution θ_1 is shrunk towards 0. To generate each realization from the mixed θ truncated distribution, for $i = 1, \ldots, 100\,000$ we independently sample λ_i from $\{10, 1\}$, with probabilities 0.90 and 0.10, and then we repeatedly sample (θ, Y), $\theta_i \sim \pi_1(\theta_i|\lambda_i)$ and $Y_i \sim \phi(\theta_i)$, keeping the first (θ, y) for which $|y_1| > 3.111$. The joint density of (θ_1, Y_1) that is given in equation (23) is

$$\left\{\frac{0.9\,\pi_1(\theta_1|\lambda_1 = 10)}{\Pr(|Y_1| > 3.111|\lambda_1 = 10)} + \frac{0.1\,\pi_1(\theta_1|\lambda_1 = 1)}{\Pr(|Y_1| > 3.111|\lambda_1 = 1)}\right\}\phi(y_1 - \theta_1).$$

Comparing Figs 3(a)–3(c) reveals that in this model the shrinking of the distribution of $\theta_1|Y_1 = y_1$ towards 0 is weaker than in the fixed θ model.

3. Selection-adjusted Bayesian inference

To define selection-adjusted Bayes inference, we express the average risk that is incurred by providing selective inference for $h(\theta)$

$$\begin{aligned} r_S(\delta) &= \int_\theta \int_{y \in S_\Omega} L\{h(\theta), \delta(y)\}\,\pi_S(\theta)\,f_S(y|\theta)\,\mathrm{d}y\,\mathrm{d}\theta \\ &= \int_{y \in S_\Omega} \left[\int_\theta L\{h(\theta), \delta(y)\}\,\pi_S(\theta|y)\,\mathrm{d}\theta\right] m_S(y)\,\mathrm{d}y. \end{aligned} \tag{26}$$

Thus the Bayes rules in selective inference are the actions minimizing the selection-adjusted posterior expected loss

$$\rho_S(\delta, y) = \int L\{h(\theta), \delta(y)\}\,\pi_S(\theta|y)\,\mathrm{d}\theta,$$

and in general Bayesian selective inference should be based on the selection-adjusted posterior distribution of $h(\theta)$, $\pi_S\{h(\theta)|y\}$. Selection-adjusted $1-\alpha$ credible intervals for $h(\theta)$ are subsets A for which $\Pr_{\pi_S\{h(\theta)|y\}}\{h(\theta) \in A\} = 1 - \alpha$, and the posterior mean or mode of $\pi_S\{h(\theta)|y\}$ can serve as selection-adjusted point estimators for $h(\theta)$.

3.1. Example 6

We provide selection-adjusted Bayes inference for the data that were simulated in example 2 for two selected parameters: $h_{12647}(\theta) = \theta_{12647}$ with $S_\Omega^{12647} = \{y : |y_{12647}| > 3.111\}$ and $h_{90543}(\theta) = \theta_{90543}$ with $S_\Omega^{90543} = \{y : |y_{90543}| > 3.111\}$. Since selection is applied to (θ, Y) then θ is a random parameter. Recall that we use two prior models for θ in our analysis. In the first model we assume that (θ, Y) was generated by a random-effect model with $\pi(\theta_i)$ in equation (3). In this model the selection-adjusted Bayes posterior distribution of θ_i is proportional to the distribution of (θ_i, Y_i) in equation (22):

$$\pi_S(\theta_i|y_i) \propto \pi(\theta_i)\,\phi(y_i - \theta_i). \tag{27}$$

In the second model (θ, Y) is generated by a non-exchangeable random-effect model with unknown $\pi^i(\theta_i)$ (note that if it were assumed that θ was generated by a random-effect model then empirical Bayes methods could be used to estimate $\pi(\theta_i)$). Thus, following Box and Tiao, we use the flat non-informative prior $\pi^i(\theta_i) = 1$ in our analysis. The flat prior unadjusted posterior distribution of θ_i is

$$\pi(\theta_i|y_i) \propto \phi(y_i - \theta_i). \tag{28}$$

The non-informative prior selection-adjusted Bayes posterior distribution of θ_i is proportional to the distribution of (θ_i, Y_i) for fixed θ in equation (21)

$$\pi_S(\theta_i|y_i) \propto \phi(y_i - \theta_i)/\Pr(S_{\text{marg}}|\theta_i), \tag{29}$$

with $\Pr(S_{\text{marg}}|\theta_i) = \Phi(-3.111 - \theta_i) + 1 - \Phi(3.111 - \theta_i)$.

Fig. 4 displays the posterior distributions of θ_{12647} (Fig. 4(a)) and θ_{90543} (Fig. 4(b)). The flat prior unadjusted posterior mean and mode of θ_{12647} equal $Y_{12647} = 3.40$, and the 0.95 credible interval is [1.44, 5.36]. The selection-adjusted Bayes posterior distribution of θ_{12647} is shrunk towards 0. The random-θ selection-adjusted Bayes posterior distribution of θ_{12647} is bimodal with a spike at 0 and a mode at 2.40, the posterior mean is 1.68 and the 0.95 credible interval

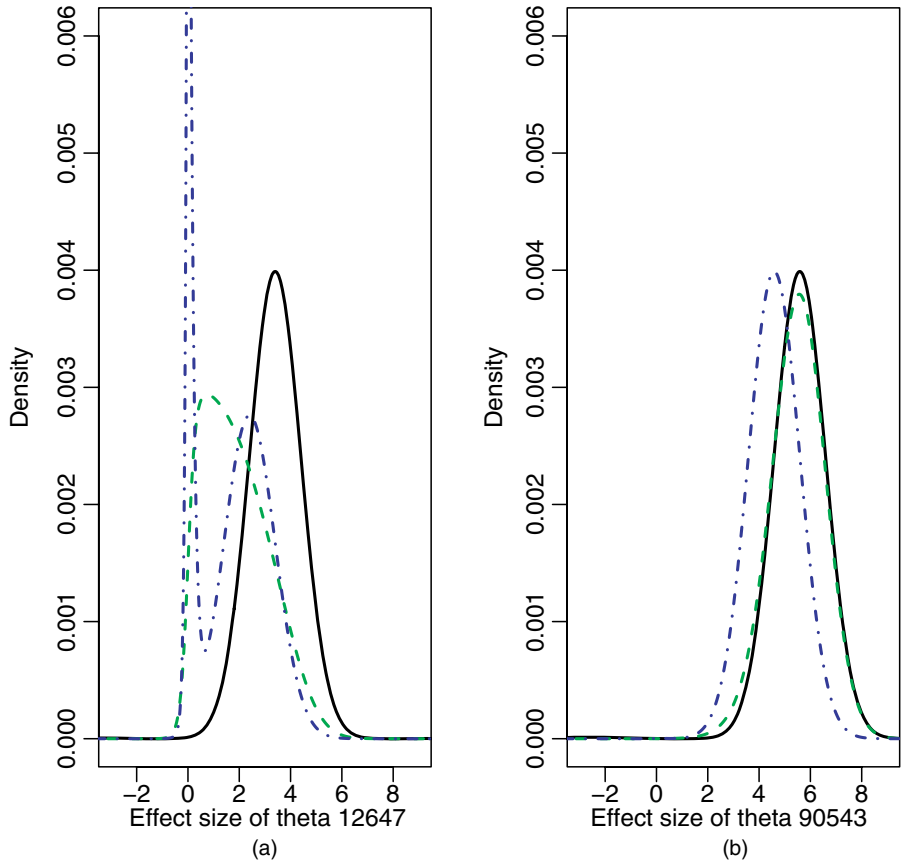

Fig. 4. Simulated example—selection-adjusted Bayes posterior distributions for (a) θ_{12647} and (b) θ_{90543}: ———, unadjusted posteriors; – · – · –, random-parameter model selection-adjusted Bayes posteriors; – – –, non-informative-prior selection-adjusted Bayes posteriors

is $[-0.11, 4.20]$. The flat prior selection-adjusted Bayes posterior mode of θ_{12647} is 0.74, the posterior mean is 1.88 and the 0.95 credible interval is $[-0.04, 4.64]$.

The flat prior unadjusted posterior mean and mode of θ_{90543} equal $Y_{90543} = 5.59$, and the 0.95 credible interval is $[3.63, 7.55]$. The much larger Y_{90543} produces a non-negligible likelihood only for θ_i-values that correspond to almost certain selection. Thus in this case the selection adjustment is small: the flat prior selection-adjusted Bayes posterior mode is 5.57, the posterior mean is 5.48 and the 0.95 credible interval is $[3.26, 7.52]$. The shrinking towards 0 in the random-θ model posterior is stronger: the posterior mean and mode are 4.59 and the 0.95 credible interval is $[2.62, 6.55]$.

Remark 3. It is important to note that, as extremely unlikely values of θ with an extremely small selection probability can have a large selection-adjusted likelihood, the selection adjustment posterior distribution can be very different from the unadjusted posterior distribution. The selection-adjusted likelihood can even be non-informative and improper—if the selection rule includes only the observed value $Y = y$ then the selection-adjusted likelihood is constant for all parameter values. Example 7 illustrates this and shows how it is affected by the choice of the selection rule and that it is not unique to Bayesian selective inference. In this paper we employ selection rules whose selection probability is minimized at $\theta = 0$ and approaches 1 for large $|\theta|$; thus the selection adjustments shrink the likelihood towards 0.

3.2. Example 7

We derive the non-informative prior selection-adjusted Bayes posterior distribution of θ_{12647}, which is given in equation (29), for an alternative one-sided selection rule $S_\Omega^{12647} = \{y : y_{12647} > 3.111\}$. In this case the selection-adjusted posterior is stochastically smaller and much more diffuse. The selection-adjusted posterior mode is 0.19 and the selection-adjusted posterior mean is -2.87; the 0.95 selection-adjusted credible interval is $[-15.41, 3.91]$. An unlikely value $\theta_{12647} = -5.87$, with unadjusted likelihood $\phi(-5.87 - 3.40) = 8.73 \times 10^{-20}$ and selection probability $\Phi(-5.87 - 3.111) = 1.34 \times 10^{-19}$, has the same selection-adjusted posterior density as the unadjusted posterior mode $\theta_{12647} = 3.40$, i.e. $\pi_S(\theta_{12647} = -5.87 | Y_{12647} = 3.40) = \pi_S(\theta_{12647} = 3.40 | Y_{12647} = 3.40)$.

We now show that frequentist selection-adjusted inference can also be very different from the unadjusted frequentist inference and highly dependent on the type of selection rule that is used. The flat prior unadjusted 0.95 credible interval for θ_{12647}, $[1.44, 5.36]$, is also a 0.95 frequentist CI for θ_{12647}. To construct selection-adjusted frequentist 0.95 CIs for θ_{12647} we begin by testing, at level 0.05 and for each value of θ_0, the null hypothesis that $\theta_{12647} = \theta_0$. The sampling distribution of $Y_{12647} | \theta_{12647} = \theta_0$ is $f_S(y_{12647} | \theta_{12647})$ in equation (8) with $\theta_{12647} = \theta_0$. Thus we reject the null hypothesis that $\theta_{12647} = \theta_0$ if y_{12647} is smaller than the 0.025-quantile or larger than the 0.975-quantile of $f_S(y_{12647} | \theta_0)$, and the 0.95 CI for θ_{12647} is the set of θ_0-values for which the null hypothesis that $\theta_{12647} = \theta_0$ is not rejected for $y_{12647} = 3.40$. For the selection rule $S_\Omega^{12647} = \{y : |y_{12647}| > 3.111\}$ the 0.95 CI for θ_{12647} is $[-0.37, 5.03]$, whereas for $S_\Omega^{12647} = \{y : y_{12647} > 3.111\}$ the 0.95 CI for θ_{12647} is $[-9.44, 5.03]$.

3.3. False coverage statement proportion control in the random-effect model

We define FCR for (θ, Y) generated by the random-effect model. The initial set of parameters is $\theta_1 \dots \theta_m$. The subset of selected parameters is $\{\theta_i : y_i \in S_{\text{marg}}\}$, and a marginal CI $A_{\text{marg}}(y_i)$ is constructed for each selected θ_i. For $i = 1, \dots, m$, let $R_i = I(Y_i \in S_{\text{marg}})$ and $V_i = I\{Y_i \in S_{\text{marg}}, \theta_i \notin A_{\text{marg}}(Y_i)\}$. $R = \Sigma R_i$ is the number of selected parameters, $V = \Sigma V_i$ is the number of non-covering CIs and $\text{FCP} = V / \max(1, R)$ is the false coverage statement

proportion. In Benjamini and Yekutieli (2005) FCR refers to a frequentist FCR that corresponds to $E_{Y|\theta}(\text{FCP})$ for (θ, Y) generated by a random-effect model. In this paper FCR is a Bayesian FCR, which is defined by $E_{\theta,Y}(\text{FCP})$. We also consider the positive FCR, $\text{pFCR} = E_{\theta,Y}(\text{FCP}|R>0)$.

3.3.1. Relationship between false coverage statment proportion control and Bayesian selective inference

Note that, for $i=1,\ldots,m$, the indicators R_i and V_i are defined for the joint (untruncated) distribution of (θ, Y). The event $R_i=1$ is given by $\{(\theta, y): y_i \in S_{\text{marg}}\}$. The conditional distribution of (θ, Y) given $R_i=1$ is

$$f(\theta, y|R_i=1) = I_{S_{\text{marg}}}(y_i) \prod_{j=1}^{m} \pi(\theta_j)\, f(y_j|\theta_j)/\Pr(Y_i \in S_{\text{marg}}), \tag{30}$$

and integrating out $\theta^{(i)}$ and $y^{(i)}$ yields the conditional distribution of (θ_i, Y_i) given $R_i=1$ to be

$$f(\theta_i, y_i|R_i=1) = I_{S_{\text{marg}}}(y_i)\, \pi(\theta_i)\, f(y_i|\theta_i)/\Pr(Y_i \in S_{\text{marg}}). \tag{31}$$

This is the same as the random-parameter selection-adjusted distribution of (θ_i, Y_i) that is given in equation (22). This implies that the conditional probability that the CI constructed for θ_i fails to cover θ_i, given that θ_i is selected, can be expressed as the average risk that is incurred in selective inference for $h_i(\theta)=\theta_i$ with $S_\Omega^i = \{y: y_i \in S_{\text{marg}}\}$ and with θ being a random parameter, for the loss function $L\{\theta_i, A_i(y)\} = I\{\theta_i \notin A_{\text{marg}}(y_i)\}$:

$$\Pr(V_i=1|R_i=1) = \int_{\theta_i} \int_{y_i \in S_{\text{marg}}} \frac{\pi(\theta_i)\, f(y_i|\theta_i)\, I\{\theta_i \notin A_{\text{marg}}(y_i)\}}{\Pr(Y_i \in S_{\text{marg}})}\, \mathrm{d}y_i\, \mathrm{d}\theta_i = r_{\text{S}}. \tag{32}$$

$\Pr(V_i=1|R_i=1, Y_i=y_i)$ is equal to the random-θ selection-adjusted posterior expected loss

$$\rho(y_i) = \int I\{\theta_i \notin A_{\text{marg}}(y_i)\}\, \pi_{\text{S}}(\theta_i|y_i)\, \mathrm{d}\theta_i, \tag{33}$$

for $\pi_{\text{S}}(\theta_i|y_i) \propto \pi(\theta_i)\, f(y_i|\theta_i)$ the random-θ selection-adjusted posterior distribution.

Proposition 1. pFCR and $E(V)$ or $E(R)$ are equal to the random-θ average risk in equation (32). If $A_{\text{marg}}(y_i)$ are $1-\alpha$ credible intervals for θ_i based on the random-θ selection-adjusted posterior distribution then $\text{pFCR} = \alpha$.

Proof. In the random-effect model $\{V_i : R_i = 1\}$ are mutually independent with $\Pr(V_i = 1|R_i=1) = r_{\text{S}}$. Thus for each value of $R=k$, $V \sim \text{Binom}(k, r_{\text{S}})$, and conditioning on $R>0$ yields $\text{pFCR} = r_{\text{S}}$. Note that the numerator and denominator in equation (32) equal $E(V_i)$ and $E(R_i)$. Thus $E(V)/E(R) = E(V_i)/E(R_i)$ is also r_{S}. Lastly, for $1-\alpha$ selection-adjusted credible intervals based on $\pi_{\text{S}}(\theta_i|y_i)$, $r_{\text{S}} = \rho(y_i) \equiv \alpha$.

Remark 4. We have shown that in the random-effect model, regardless of whether θ is random fixed or mixed, pFCR equals the random-θ selection-adjusted average risk. As $\text{pFCR} \geqslant$ Bayesian-FCR the random θ average risk can serve as a conservative estimate for Bayesian-FCR. In particular, for large R the sampling dispersion of FCP and of $V/E(R)$ is small; thus FCP, Bayesian-FCR, frequentist-FCR and pFCR that equals $E(V)/E(R)$, which we discuss in the context of specifying selection rules in the non-exchangeable random-effect model, are almost the same.

Remark 5. Recall that if $\pi(\theta_i)$ is a non-informative prior then the selection-adjusted posterior distribution for random θ is defined as

$$\pi_S(\theta_i|y_i) \propto \pi(\theta_i)\, f(y_i|\theta_i)/\Pr(S_{\text{marg}}|\theta_i). \tag{34}$$

As credible intervals based on non-informative priors are expected to provide approximate coverage probability, when $\pi(\theta_i)$ is a non-informative prior then $1-\alpha$ credible intervals based on $\pi_S(\theta_i|y_i)$ in expression (34) yield $\rho(y_i) \approx \alpha$. Thus proposition 1 implies that for non-informative priors the fixed θ marginal $1-\alpha$ credible intervals yield approximate level α FCR-control.

3.4. Example 8

Fig. 2 displays (θ_i, y_i) generated in example 2 with $y_i > 3.111$. The full and broken curves are the 0.95 CIs from Fig. 1. The full curves also correspond to the 0.95 credible intervals for θ_i for the flat prior unadjusted posterior (28). The chain curves are the 0.95 selection-adjusted Bayes credible intervals for the flat prior selection-adjusted posterior in equation (29), and the dotted curves are the 0.95 selection-adjusted Bayes credible intervals for the random-θ selection-adjusted posterior in equation (27).

According to proposition 1 pFCR for random-θ 0.95 selection-adjusted Bayes credible intervals constructed for selected (θ_i, y_i) is 0.05. In example 2 we have seen that FCP for these credible intervals for the 932 selected θ_i was 0.047. As the flat prior unadjusted credible intervals are 0.95 frequentist CIs, we expect the coverage proportion for all 100000 θ_i to be close to 0.95. We have seen that these CIs cover 95089 of the 100000 θ_i, but that FCP for the 932 selected parameters is 0.346. Benjamini and Yekutieli (2005) explained this phenomenon from a frequentist perspective. Remark 5 offers a Bayesian explanation: to provide approximate FCR-control for non-informative priors the credible intervals should be based on the fixed θ selection-adjusted posterior in equation (29), rather than the random-θ selection-adjusted posterior in equation (28). And indeed, FCP of the credible intervals based on equation (29) is 0.040.

4. Specifying false discovery rate controlling selection rules in the random-effect model

We shall now present Bayesian methodology for specifying selection rules in the random-effect model and the non-exchangeable random-effect model for cases in which selection is applied for making statistical discoveries. Similarly to the BH false discovery rate controlling approach, we seek to control the proportion of false discoveries committed. Unlike Benjamini and Hochberg (1995) in which discoveries refer to rejection of null hypotheses and the statistics that were used for specifying the selection rule are p-values testing these null hypotheses, in our approach any event in the parameter space can be considered a discovery and any statistic may be used for specifying the selection rule. But, as suggested in Storey (2007), we shall show that for any given discovery the optimal statistic is the posterior probability that the discovery is false.

As in Section 3.1, we assume that (θ, Y) are generated by the random-effect model; θ_i is selected if $y_i \in S_{\text{marg}}$; and the inference that is provided for θ_i if it is selected is declaring that $\theta_i \in A_{\text{marg}}(y_i)$. However, now $A_{\text{marg}}(y_i)$ is an event that corresponds to making a statistical discovery regarding θ_i. For example, in the microarray analysis in Section 6, in which the discovery is declaring a gene either overexpressed or underexpressed, for $y_i > 0$ the discovery event is $A_{\text{marg}}(y_i) = \{\theta_i : \theta_i > 0\}$.

Once declaring $\theta_i \in A_{\text{marg}}(y_i)$ corresponds to making a statistical discovery, R becomes the number of discoveries, V becomes the number of false discoveries, $V/\max(1, R) = \text{FDP}$ is the false discovery proportion and FCR = FDR. Thus proposition 1 yields the following result.

Corollary 1. In the random-effect model pFDR equals r_S in equation (32), which is the conditional probability given that θ_i is selected that the discovery regarding θ_i is false, and $\rho(y_i)$ in equation (33) is the conditional probability given selection and given $Y_i = y_i$ that the discovery is false.

Thus to ensure level q FDR-control, when considering selection rules of the form $S_{\text{marg}} = \{y_i : T(y_i) \leqslant s\}$, we suggest choosing s for which r_S in equation (32) is less than or equal to q. Furthermore, re-expressing r_S,

$$r_S = \frac{\displaystyle\int_{y_i \in S_{\text{marg}}} m(y_i) \int_{\theta_i} \pi_S(\theta_i|y_i)\, I\{\theta_i \notin A_{\text{marg}}(y_i)\}\, \mathrm{d}\theta_i\, \mathrm{d}y_i}{\Pr(Y_i \in S_{\text{marg}})}$$
$$= \frac{\displaystyle\int_{y_i \in S_{\text{marg}}} m(y_i)\, \rho(y_i)\, \mathrm{d}y_i}{\displaystyle\int_{y_i \in S_{\text{marg}}} m(y_i)\, \mathrm{d}y_i}, \tag{35}$$

where $m(y_i) = \int \pi(\theta_i)\, f(y_i|\theta_i)\, \mathrm{d}\theta_i$, yields the following Neyman–Pearson lemma type of result, presented in Storey (2007).

Corollary 2. The selection rule of the form $S_{\text{marg}} = \{y_i : \rho(y_i) \leqslant s\}$ has the largest selection probability of all selection rules with the same pFDR.

Another option is to use $\rho(y_i)$ to specify the selection rule directly, by defining

$$S_{\text{marg}} = \{y_i : \rho(y_i) \leqslant q\}. \tag{36}$$

Unlike the continuum of possible credible intervals that can be constructed for θ_i, the number of possible discoveries that can be made regarding θ_i is usually finite, e.g. discovering that θ_i is either negative or positive or discovering that θ_i is the largest component in θ. In particular, when there is only a single possible discovery for all selected values of y_i, i.e. $A_{\text{marg}}(y_i) \equiv A_{\text{marg}}$, then, expressing the random-θ average risk corresponding to this discovery

$$r_S = \int\int_{y_i \in S_{\text{marg}}} I(\theta_i \notin A_{\text{marg}})\, \frac{\pi(\theta_i)\, f(y_i|\theta_i)}{\Pr(Y_i \in S_{\text{marg}})}\, \mathrm{d}y_i\, \mathrm{d}\theta_i$$
$$= \int I(\theta_i \notin A_{\text{marg}})\, \frac{\pi(\theta_i) \Pr(Y_i \in S_{\text{marg}}|\theta_i)}{\Pr(Y_i \in S_{\text{marg}})}\, \mathrm{d}\theta_i$$
$$= \int I(\theta_i \notin A_{\text{marg}})\, \pi_S(\theta_i)\, \mathrm{d}\theta_i, \tag{37}$$

for $\pi_S(\theta_i) = \pi(\theta_i) \Pr(S_{\text{marg}}|\theta_i)/\Pr(S_{\text{marg}})$ the random-θ selection-adjusted prior density derived in equation (11), yields the following result.

Corollary 3. If $A_{\text{marg}}(y_i) \equiv A_{\text{marg}}$ then pFDR is equal to the random-θ selection-adjusted prior probability that $\theta_i \notin A_{\text{marg}}$.

4.1. Specifying false discovery rate controlling selection rules in the non-exchangeable random-effect model

In this subsection, (θ, Y) is generated by the non-exchangeable random-effect model, θ_i is selected if $y_i \in S_{\text{marg}}$ and the inference that is provided for selected θ_i is the discovery that $\theta_i \in A_{\text{marg}}(y_i)$. Let $A^1_{\text{marg}} \ldots A^D_{\text{marg}}$ denote the D possible discoveries that can be made on θ_i. For $d = 1, \ldots, D$, let

R^d denote the number of discoveries of A^d_{marg} and let V^d denote the number of false discoveries of A^d_{marg}. The results in this section are derived under the assumption that $A_{\text{marg}}(y_i) \equiv A_{\text{marg}}$. However, as $E(R) = E(R^1) + \ldots + E(R^D)$ and $E(V) = E(V^1) + \ldots + E(V^D)$, they can be easily extended for the case of $D > 1$.

To derive the results in this section, we assume that there also exists $(\tilde{\theta}, \tilde{Y})$, generated by the random-parameter model that $\tilde{\theta}_i$ are IID $\tilde{\pi}(\theta_i) = \Sigma_{i=1}^m \pi^i(\theta_i)/m$, and $\tilde{Y}_i|\tilde{\theta}_i$ are independent $f(\tilde{y}_i|\tilde{\theta}_i)$.

Lemma 1. For any subset B, $W_i = I(y_i \in S_{\text{marg}}, \theta_i \notin B)$, and $\tilde{W}_i = I(\tilde{y}_i \in S_{\text{marg}}, \tilde{\theta}_i \notin B)$

$$E\left(\sum_{i=1}^m W_i\right) = E\left(\sum_{i=1}^m \tilde{W}_i\right).$$

Proof.

$$\begin{aligned}
E\left(\sum_{i=1}^m W_i\right) &= \sum_{i=1}^m \Pr(Y_i \in S_{\text{marg}}, \theta_i \notin B) \\
&= \sum_{i=1}^m \int_{\theta_i \notin B} \int_{y_i \in S_{\text{marg}}} \pi^i(\theta_i)\, f(y_i|\theta_i)\, \mathrm{d}y_i\, \mathrm{d}\theta_i \\
&= \sum_{i=1}^m \int_{\theta_1 \notin B} \int_{y_1 \in S_{\text{marg}}} \pi^i(\theta_1)\, f(y_1|\theta_1)\, \mathrm{d}y_1\, \mathrm{d}\theta_1 \\
&= m \int_{\theta_1 \notin B} \int_{y_1 \in S_{\text{marg}}} \sum_{i=1}^m \pi^i(\theta_1)/m\, f(y_1|\theta_1)\, \mathrm{d}y_1\, \mathrm{d}\theta_1 \\
&= m \int_{\theta_1 \notin B} \int_{y_1 \in S_{\text{marg}}} \tilde{\pi}(\theta_1)\, f(y_1|\theta_1)\, \mathrm{d}y_1\, \mathrm{d}\theta_1 = E\left(\sum_{i=1}^m \tilde{W}_i\right).
\end{aligned}$$

For $B = \emptyset$, $\Sigma_{i=1}^m W_i$ is the number of discoveries R, whereas, for $B = A_{\text{marg}}$, $\Sigma_{i=1}^m W_i$ is the number of false discoveries. Therefore lemma 1 implies that $E(V)$, $E(R)$, and thus also $\text{pFDR} = E(V)/E(R)$, for (θ, Y) and for $(\tilde{\theta}, \tilde{Y})$ are the same. According to corollary 1 pFDR for $(\tilde{\theta}, \tilde{Y})$ is the corresponding random-θ average risk, which we denote $\tilde{r}_S$. Thus since $\text{FDR} \leqslant \text{pFDR}$, and pFDR is the same for (θ, Y) and for $(\tilde{\theta}, \tilde{Y})$, we obtain the following result.

Corollary 4. In the non-exchangeable random-parameter model selecting θ_i if $y_i \in S_{\text{marg}}$ yields level $\tilde{r}_S$ FDR-control.

To define a general method for specifying false discovery rate controlling selection rules for (θ, Y) generated by the non-exchangeable random-effect model with unknown marginal priors, applying empirical Bayes methods to $y_1 \ldots y_m$ actually estimates $\tilde{\pi}(\theta_i)$, the mixture of the (unknown) marginal densities of $\theta_1 \ldots \theta_m$. Combining this with corollary 4 implies that FDR of any selection rule can be approximated by $\tilde{r}_S$ computed by treating (θ, Y) as if it was generated by the random-effect model and using the empirical Bayes estimate of $\tilde{\pi}(\theta_i)$. Furthermore, as $E(R) = E(\tilde{R})$ and $E(\tilde{R}) = m \Pr(\tilde{Y}_i \in S_{\text{marg}})$, then also in the non-exchangeable random-effect model the selection rule $S_{\text{marg}} = \{y_i : \tilde{\rho}(y_i) \leqslant s\}$, where $\tilde{\rho}(y_i)$ is the posterior expected loss in equation (33) computed for $(\tilde{Y}, \tilde{\theta})$, yields the maximal $E(R)$ among all S_{marg} with the same $\tilde{r}_S$.

Definition 2. An algorithm for specifying level q FDR controlling selection rules in the non-exchangeable random-effect model is as follows.

Step 1: apply empirical Bayes methods to $y_1 \ldots y_m$ to produce $\tilde{\pi}(\theta_i)$.
Step 2: use $\tilde{\pi}(\theta_i)$ to compute $\tilde{r}_S$ for any given selection rule.

Step 3:

(a) to specify a level q false discovery rate controlling selection rule of the form $S_{\text{marg}} = \{y : T(y_i) \leqslant s\}$, for a given statistic $T(y_i)$, find s for which $\tilde{r}_S = q$;
(b) the level q false discovery rate controlling selection rule yielding the maximal expected number of discoveries is $S_{\text{marg}} = \{y : \tilde{\rho}(y_i) \leqslant s\}$ with s, for which $\tilde{r}_S = q$.

4.2. Example 9

In example 2 selection is associated with $D = 2$ directional discoveries. According to corollary 1 pFDR for the selection rule $|y_i| \geqslant s$ is equal to the random-θ average risk for the loss function $I\{\text{sgn}(\theta_i) \neq \text{sgn}(y_i)\}$

$$E_{m_S(y)}\{I(y < -a)\Pr_{\pi_S(\theta|y)}(\theta > 0) + I(y > a)\Pr_{\pi_S(\theta|y)}(\theta < 0)\}. \tag{38}$$

Recall that $|y_i| > 3.111$ was used to ensure that the directional FDR is less than 0.1. For $s = 3.111$ the average risk (38) is 0.070, whereas setting $s = 2.915$ yields the selection criterion for which the average risk is 0.10. The posterior expected loss corresponding to the directional FDR is

$$\rho(y_i) = \Pr_{\pi(\theta|y)}\{\text{sgn}(\theta_i) \neq \text{sgn}(y_i)\}.$$

Note that in this example $\rho(y_i)$ increases in $|y_i|$; thus $|y_i| \geqslant 2.915$ is the $r_S = 0.10$ selection rule yielding the maximal expected number of discoveries. For $y_i \geqslant 0$, $\rho(y_i)$ is the conditional probability given y_i that $\theta_i < 0$. $\rho(0) = 0.5$, $\rho(3.111) = 0.176$ and $\rho(3.472) = 0.10$. Thus $|y_i| \geqslant 3.472$ is the selection criterion that is suggested in equation (36) for $q = 0.10$.

The random-effect model generated in example 2 is the $(\tilde{\theta}, \tilde{Y})$ that corresponds to the non-exchangeable random-effect model (θ, Y) in example 5. To illustrate our results on the non-exchangeable random-effect model, we evaluated $E(V)$, $E(R)$ and the directional FDR for $n = 10^5$ samples of $(\tilde{\theta}, \tilde{Y})$ and of (θ, Y). In both cases the mean number of discoveries was 919.9 (standard error se < 0.07), the mean number of false discoveries was 64.4 (se < 0.03) and the mean directional FDP was 0.070 (se < 0.00003).

5. Relationship between selection-adjusted Bayes inference and Bayesian false discovery rate methods

The term Bayesian false discovery rate methods refers to the multiple-testing procedures that were presented in Efron *et al.* (2001) and Storey (2002, 2003) for the following two-group mixture model. H_i, $i = 1, \ldots, m$, are IID Bernoulli$(1 - \pi_0)$ random variables. $H_i = 0$ corresponds to a true null hypothesis, whereas $H_i = 1$ corresponds to a false null hypothesis. Given $H_i = j$, Y_i is independently drawn from f_j, for $j = 0, 1$.

pFDR corresponds to a rejection region Γ. It is defined by $E(V/R|R > 0)$ where R is the number of $y_i \in \Gamma$, and V is the number of $y_i \in \Gamma$ with $H_i = 0$. Storey (2002) proved that

$$\text{pFDR}(\Gamma) = \Pr(H_i = 0 | Y_i \in \Gamma) \tag{39}$$

$$= \frac{\pi_0 \Pr(Y_i \in \Gamma | H_i = 0)}{\pi_0 \Pr(Y_i \in \Gamma | Y_i = 0) + (1 - \pi_0)\Pr(Y_i \in \Gamma | H_i = 1)}, \tag{40}$$

with $\Pr(Y_i \in \Gamma | H_i = j) = \int_{y_i \in \Gamma} f_j(y_i)\, \mathrm{d}y_i$. For the multiple-testing procedure each null hypothesis is associated with a rejection region Γ_i, determined by y_i; pFDR corresponding to Γ_i, which is called the q-value, is computed; and the null hypothesis $H_i = 0$ is rejected if the q-value is

less than or equal to q. The local false discovery rate is defined in Efron *et al.* (2001) as the conditional probability given $Y_i = y_i$ that $H_i = 0$

$$\mathrm{fdr}(y_i) = \frac{\pi_0 f_0(y_i)}{\pi_0 f_0(y_i) + (1-\pi_0) f_1(y_i)}.$$

The multiple-testing procedure based on the local false discovery rate is to reject $H_i = 0$ if $\mathrm{fdr}(y_i) \leqslant q$.

Note that Bayesian false discovery rate methods can be expressed as a special case of the false discovery rate controlling selection rules that were presented in the previous section, in which the components of the parameter vector are dichotomous. The parameter is $H = (H_1 \ldots H_m)$, and (H, Y) are generated by a random-effect model: the marginal distribution of H_i is $\pi(H_i = j) = (1-\pi_0)^j \pi_0^{1-j}$, f_j is the likelihood, H_i is selected if $y_i \in \Gamma$ and selection is associated with the discovery that $H_i = 1$. Note also that expression (40) is a special case of expression (37): it is the random-parameter average risk for the loss function $I(H_i = 0)$, expressed as the selection-adjusted prior distribution of making a false discovery

$$\pi_\Gamma(H_i = 0) \propto \pi(H_i = 0) \Pr(Y_i \in \Gamma | H_i = 0).$$

Thus the equality in expression (39) that was proved by Storey is a special case of corollary 3. The local false discovery rate is the random-θ selection-adjusted posterior expected loss; thus the multiple-testing procedure based on the local false discovery rate is a special case, of the selection rule in expression (36). Lastly, the relationship between the local false discovery rate and pFDR, $\mathrm{pFDR} = E_{y \in \Gamma}\{\mathrm{fdr}(y)\}$, follows from the definition of the average risk in equation (26).

Bayesian false discovery rate methods are valid regardless of whether H is a random or fixed parameter. However, in selective inference for $h_i(H) = H_i$, the selection-adjusted posterior probability that $H_i = 0$ for a random H is equal to the local fdr, whereas if H is a fixed parameter, or if π_0 is the non-informative prior probability that $H_i = 0$, then the selection-adjusted posterior distribution that $H_i = 0$ is

$$\frac{\pi_0 f_\Gamma(y_i | H_i = 0)}{\pi_0 f_\Gamma(y_i | H_i = 0) + (1-\pi_0) f_\Gamma(y_i | H_i = 1)},$$

for $f_\Gamma(y_i | H_i = j) = f_j(y_i) / \Pr(y_i \in \Gamma | H_i = j)$ the selection-adjusted likelihood.

6. Analysis of microarray data

We analyse the Dudoit and Yang (2003) swirl data set. The data include four arrays with 8448 genes, comparing ribonucleic acid from zebra fish with the swirl mutation with ribonucleic acid from wild-type fish. For gene g, $g = 1, \ldots, 8448$, the parameters are μ_g, the expected base 2 log-fold change in expression due to the swirl mutation, and σ_g^2, the variance of the base 2 log-fold change in expression.

In our analysis we assume that (θ, Y) are generated by a non-exchangeable random-effect model. Since the measurement error variances are expected to vary from experiment to experiment, σ_g^2 are IID random parameters with scaled inverse χ^2 marginal prior density $\pi(\sigma_g^2)$, whose hyperparameters $s_0^2 = 0.052$ and $\nu_0 = 4.02$ were derived by applying the R LIMMA package (Smyth, 2005) `eBayes` function to the sample variances, whereas μ_g are distinct independent fixed parameters that are elicited flat non-informative priors, $\pi_{ni}(\mu_g) \propto 1$. However, for assessing FDR of the BH procedure and for specifying the Bayesian selection rules we use the emiprical Bayes prior

$$\tilde{\pi}(\mu_g) = 8.5 \exp(-8.5\,|\mu_g|)/2,$$

that provided a good fit to the empirical distribution of $\bar{y}_1 \ldots \bar{y}_{8448}$. Given μ_g and σ_g, s_g^2, the sample variances, are independent $\sigma_g^2 \chi_3^2/3$, and $\bar{y}_g$, the observed mean base 2 log-expression-ratios, are independent $N(\mu_g, \sigma_g^2/4)$. Thus the marginal likelihood is given by

$$f(\bar{y}_g, s_g^2 | \mu_g, \sigma_g^2) \propto \sigma_g^{-4} \exp\left[-\frac{1}{2\sigma_g^2} \{3s_g^2 + 4(\mu_g - \bar{y}_g)^2\} \right]. \tag{41}$$

Our goal in the analysis is to specify a selection rule for which the mean directional error in declaring selected genes with $\bar{y}_g > 0$ overexpressed and declaring selected genes with $\bar{y}_g < 0$ underexpressed is less than 0.05, and to provide inference for the change in expression of selected genes.

6.1. Specifying the selection rules

In the first part of our analysis we apply the level $q = 0.10$ BH procedure to moderated t-statistic p-values to discover differentially expressed genes; assess the directional false discovery rate of the selection rule specified by the BH procedure, and compare its performance with the performance of the level $q = 0.05$ directional false discovery rate controlling selection rules based on moderated t-statistics and on the posterior expected loss.

LIMMA implements a hybrid classical–Bayes approach in which μ_g are assumed to be unknown constants whereas σ_g^2 are IID $\pi(\sigma_g^2)$. The moderated t-statistics are defined as $\tilde{t}_g = \bar{y}_g/(\tilde{s}_g/2)$, for $\tilde{s}_g^2 = (\nu_0 s_0^2 + 3s_g^2)/(\nu_0 + 3)$ the posterior mean of $\sigma_g^2 | s_g^2$. As $\tilde{s}_g^2/\sigma_g^2 \sim \chi_{\nu_0+3}^2/(\nu_0+3)$, $(\bar{y}_g - \mu_g)/(\tilde{s}_g/2)$ are $\nu_0 + 3$ degrees of freedom t random variables. Thus the p-values that LIMMA provides to test a null hypothesis of non-differential expression are $\tilde{p}_g = 2\{1 - F_{\nu_0+3}(|\tilde{t}_g|)\}$, where F_ν is the ν degrees of freedom t cumulative density function. Applied at level $q = 0.10$ to the 8448 p-values the BH procedure yielded 245 discoveries, corresponding to the rejection region $|\tilde{t}_g| > 4.479$. The observed mean base 2 log-expression-ratios and sample standard deviations of the 8448 genes are drawn in Fig. 5. The BH discoveries are the 245 observations beneath the chain curve $|\tilde{t}_g| = 4.479$. To see why this rejection region corresponds to 0.05 directional FDR-control note that, for all μ_g, the probability of a directional error is less than $1 - F_{\nu_0+3}(4.479)$; thus $12.08 = 8448\{1 - F_{\nu_0+3}(4.479)\}$ is a conservative estimate for the number of false directional discoveries, and $0.049 = 12.08/245$ is a conservative estimate for the directional false discovery rate.

For comparison, the frequentist treatment of this problem would be to test the null hypotheses of non-differential expression by 3 degrees of freedom test statistics $t_g = \bar{y}_g/(s_g/2)$. Since the 3 degrees of freedom t-distribution has heavier tails, $F_3^{-1}\{1 - 0.1/(2 \times 8448)\} = 57.10$ whereas $\max(|t_g|)$ is only 27.90. Thus applying the level $q = 0.1$ BH to $p_1 \ldots p_{8448}$, with $p_g = 2\{1 - F_3(|t_g|)\}$, yields no discoveries.

To assess the directional false discovery rate we derive the random-θ selection-adjusted Bayes posterior distribution

$$\tilde{\pi}_S(\mu_g, \sigma_g^2 | \bar{y}_g, s_g) = \frac{I\{(\bar{y}_g, s_g^2) \in S_{\text{marg}}\}\, \tilde{\pi}(\mu_g, \sigma_g^2)\, f(\bar{y}_g, s_g | \mu_g, \sigma_g^2)}{\Pr\{(\bar{y}_g, s_g^2) \in S_{\text{marg}}\}}, \tag{42}$$

for the empirical Bayes prior distribution $\tilde{\pi}(\mu_g, \sigma_g^2) = \tilde{\pi}(\mu_g)\, \pi(\sigma_g^2)$. We then integrate out σ_g^2 in equation (42) to derive $\tilde{\pi}_S(\mu_g | \bar{y}_g, s_g)$, the marginal random-$\theta$ selection-adjusted Bayes posterior distribution of μ_g, and the random-θ posterior expected loss corresponding to directional errors

$$\tilde{\rho}(\bar{y}_g, s_g^2) = \int I\{\mu_g \neq \text{sgn}(\bar{y}_g)\}\, \tilde{\pi}_S(\mu_g | \bar{y}_g, s_g^2)\, \mathrm{d}\mu_g,$$

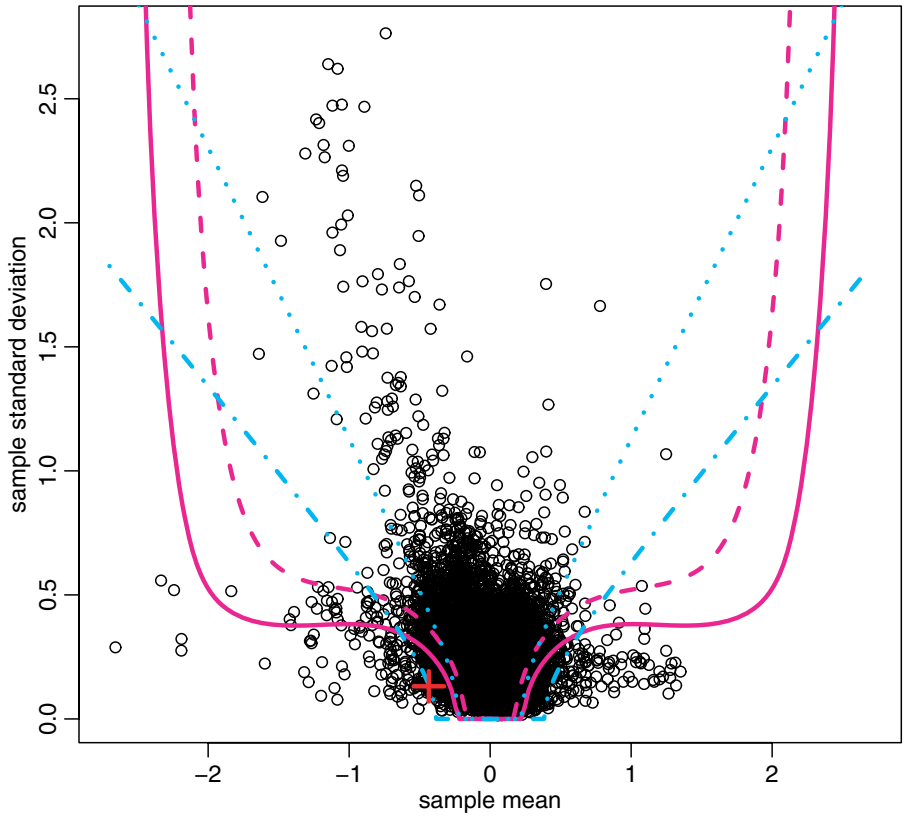

Fig. 5. Swirl data—scatter plot of sample means and standard deviations (the abscissa of the plot is $\bar{y}_g$; the ordinates are s_g): - · - · -, $|\tilde{t}_g| = 4.479$; · · · · ·, $|\tilde{t}_g| = 2.64$; ——, $\tilde{\rho}(\bar{y}_g, s_g) = 0.05$; – – –, $\tilde{\rho}(\bar{y}_g, s_g) = 0.088$; +, $(\bar{y}_{6239}, s_{6239})$

and use it to compute numerically the random-θ average risk corresponding to the false discovery rate

$$\tilde{r}_{\mathrm{S}}(S_{\mathrm{marg}}) = E_{m_{\mathrm{S}}(\bar{y}_g, s_g^2)}\{\tilde{\rho}(\bar{y}_g, s_g^2)\},$$

for

$$m_{\mathrm{S}}(\bar{y}_g, s_g) = \frac{I\{(\bar{y}_g, s_g^2) \in S_{\mathrm{marg}}\}\, \tilde{\pi}(\mu_g, \sigma_g^2)\, f(\bar{y}_g, s_g | \mu_g, \sigma_g)}{\int I\{(\bar{y}_g, s_g^2) \in S_{\mathrm{marg}}\}\, \tilde{\pi}(\mu_g, \sigma_g^2)\, f(\bar{y}_g, s_g | \mu_g, \sigma_g)\, \mathrm{d}\mu_g\, \mathrm{d}\sigma_g}.$$

$\tilde{r}_{\mathrm{S}}$ for $|\tilde{t}_g| > 4.479$ the $q = 0.10$ BH procedure (the chain curve in Fig. 5) is 0.024, whereas for $|\tilde{t}_g| > 2.64$ (the dotted curve in Fig. 5) is the moderated t selection rule with $\tilde{r}_{\mathrm{S}} = 0.05$. It yields 1124 discoveries. The full and broken curves in Fig. 5 correspond to the selection rules of the form $\tilde{\rho}(\bar{y}_g, s_g^2) < s$. The full curve corresponds to the selection rule with $s = 0.05$, that yields 559 discoveries. The broken curve corresponds to the selection rule with $s = 0.088$, for which $\tilde{r}_{\mathrm{S}} = 0.05$. This is the selection rule that yields the maximal expected number of discoveries among all selection rules with $\tilde{r}_{\mathrm{S}} = 0.05$. In this case it yields 1271 discoveries.

6.2. Providing selection-adjusted Bayes inference

In the second part of our analysis we provide selection-adjusted Bayes inference for μ_{6239}, the expected base 2 log-fold change in expression due to the swirl mutation for gene 6239. The

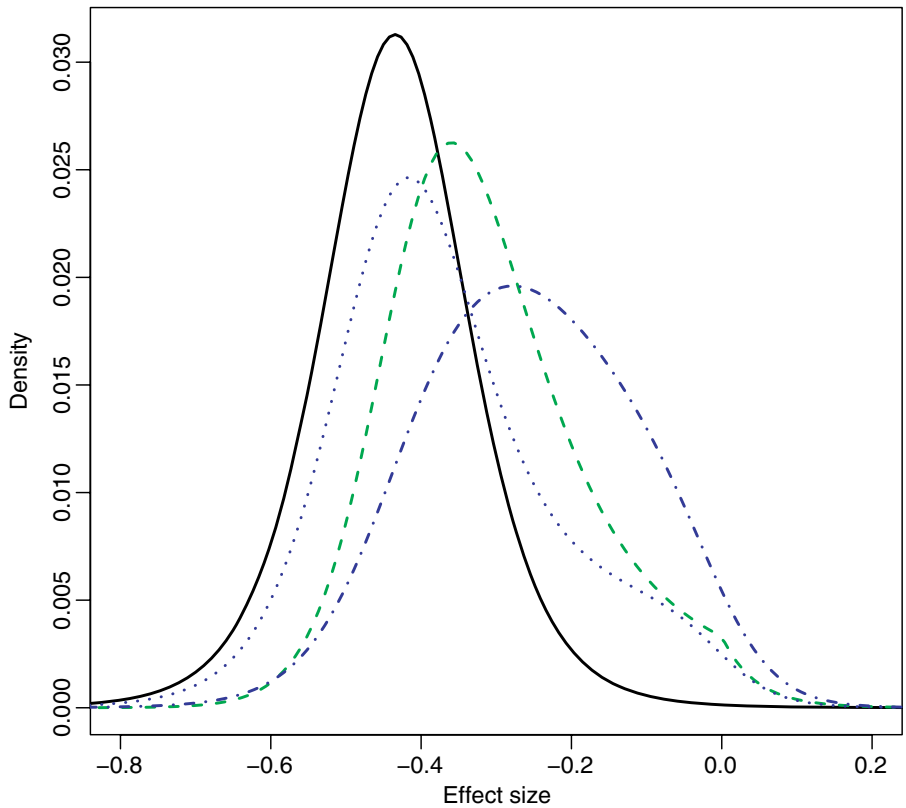

Fig. 6. Swirl data—marginal posterior densities of μ_{6239}: ———, non-informative-prior unadjusted posterior distribution; – – –, empirical Bayes prior posterior distribution; – · – · –, non-informative-prior selection-adjusted Bayes posterior distribution for the selection rule $|\tilde{t}_g| > 4.479$; · · · · · , non-informative-prior selection-adjusted Bayes posterior distribution for the selection rule $|\tilde{t}_g| > 2.64$

statistics for this gene (which are marked by the plus sign in Fig. 5) are $\bar{y}_{6239} = -0.435$ and $s^2_{6239} = 0.0173$; thus $\tilde{t}_{6239} = -4.51$. Note that a frequentist solution to this problem would be to construct an FCR-adjusted, 3 degrees of freedom t-distribution, marginal CI for μ_{6239}.

The marginal posterior distributions of μ_{6239} are drawn in Fig. 6. The full curve corresponds to the non-informative prior unadjusted posterior

$$\pi(\mu_g, \sigma^2_g | \bar{y}_g, s^2_g) \propto \pi_{ni}(\mu_g)\, \pi(\sigma^2_g)\, f(\bar{y}_g, s_g | \mu_g, \sigma^2_g),$$

for which $(\mu_{6239} - \bar{y}_{6239})/(\tilde{s}_{6239}/2) \sim t_{7.02}$. In this case, the posterior mean and mode equal $\bar{y}_{6239} = -0.435$, the 0.95 credible interval for μ_{6239} is $[-0.61, -0.21]$, the posterior probability that $\mu_{6239} > 0$ and a directional error is committed is 0.0014. The broken curve corresponds to $\tilde{\pi}_S(\tilde{\mu}_{6239} | \bar{y}_{6239}, s_{6239})$. Its posterior mode is -0.36, the posterior mean is -0.31, the 0.95 credible interval is $[-0.54, -0.01]$ and the posterior probability that $\mu_{6239} > 0$ is 0.020.

As μ_g is elicited a non-informative prior and σ^2_g is a random parameter, then (μ_g, σ^2_g) is a mixed parameter, and its selection-adjusted posterior distribution is proportional to the joint truncated distribution in equation (14), with μ_g substituting the fixed λ and σ^2_g substituting the random θ,

$$\pi_S(\mu_g, \sigma^2_g | \bar{y}_g, s^2_g) \propto f_S(\mu_g, \sigma^2_g, \bar{y}_g, s^2_g) = \pi(\sigma^2_g)\, \pi_{ni}(\mu_g)\, f(\bar{y}_g, s^2_g | \mu_g, \sigma^2_g) / \Pr(|\tilde{t}_g| > a | \mu_g). \qquad (43)$$

Selection-adjusted Bayes inference for μ_{6239} is based on $\pi_S(\mu_g|\bar{y}_g, s_g)$, the marginal selection-adjusted posterior of μ_{6239}, derived by integrating out σ_g^2 from expression (43). The chain curve is $\pi_S(\mu_g|\bar{y}_g, s_g^2)$ for the selection rule $|\tilde{t}_g| > 4.479$. Its posterior mode is -0.278, the posterior mean is -0.257, the 0.95 credible interval is $[-0.54, 0.02]$ and the posterior probability that $\mu_{6239} > 0$, and thus the gene was erroneously declared underexpressed, is 0.038. The dotted curve corresponds to $|\tilde{t}_g| > 2.64$. In this case the shrinking towards 0 is weaker: the posterior mode is -0.419, the posterior mean is -0.367, the 0.95 credible interval is $[-0.63, -0.02]$ and the posterior probability that $\mu_{6239} > 0$ is 0.017.

7. Discussion

The observation that selection affects Bayesian inference carries the important implication that in Bayesian analysis of large data sets, for each potential parameter, it is necessary to specify explicitly a selection rule that determines when inference is provided for the parameter and to provide inference that is based on the selection-adjusted posterior distribution of the parameter.

Even though specifying a selection rule introduces an arbitrary element to Bayesian analysis, it is important to note that the selection rule is determined before the data have been observed, and once the selection rule has been determined the entire process of providing selection-adjusted Bayes inference is fully specified and is carried out the same way as Bayesian inference. The notable exception is empirical Bayes methods that use the data twice in the analysis: first to elicit the prior distribution and possibly to specify the selection rule, and then to produce posterior distributions.

Our method of controlling the Bayesian false discovery rate corresponds to the fixed rejection region approach that was presented in Yekutieli and Benjamini (1999), that consists of estimating FDR in a series of nested fixed rejection regions and choosing the largest rejection region with estimated FDR less than q. However, as pFDR of any selection rule can be expressed as a selection-adjusted Bayes risk, the problem of controlling the Bayesian false discovery rate in the random-effect and non-exchangeable random-effect models is reduced to a Bayesian decision problem of finding the 'optimal' selection rule with selection-adjusted Bayes risk less than or equal to q. Our Bayesian false discovery rate controlling methods can, in principle, provide tight FDR-control, based on the optimal statistic, for any discovery event, whereas frequentist false discovery rate controlling methods may provide tight FDR-control when the discovery is rejecting a simple null hypothesis but, as illustrated by the performance of the BH procedure in controlling the directional false discovery rate, can only bound the FDR when the discoveries are rejecting composite null hypotheses.

In general, the price that is paid by using stricter selection rules is reduction in the information that the data provide for selective inference. Example 3 suggests that, when specifying selection rules, in addition to the trade-off between allowing too many false (or wasteful) discoveries and failing to make enough discoveries, it may also be advisable to take into account the quality of the inference that is provided for selected parameters.

Acknowledgements

The author acknowledges support from the Israeli Science Foundation and the Wharton School, University of Pennsylvania. The author thanks Yoav Benjamini, Bradley Efron, Edward George and Micha Mandel for interesting discussions and helpful suggestions. The author is grateful to Abba Krieger and the Joint Editor, Associate Editor and two referees for their comments and suggestions that greatly improved the quality of the paper.

References

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2005) False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Am. Statist. Ass.*, **100**, 71–81.

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

Berry, D. A. and Hochberg, Y. (1999) Bayesian perspectives on multiple comparisons. *J. Statist. Plannng Inf.*, **82**, 215–227.

Box, G. E. P. and Tiao, G. C. (1992) *Bayesian Inference in Statistical Analysis*. New York: Wiley.

Carlin, B. P. and Louis, T. A. (1996) *Bayes and Empirical Bayes Methods for Data Analysis*. New York: Chapman and Hall.

Dawid, A. P. (1994) Selection paradoxes of Bayesian inference. In *Multivariate Analysis and Its Applications*. Philadelphia: Institute for Mathematical Statistics.

Dudoit, S. and Yang, Y. H. (2003) Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. In *The Analysis of Gene Expression Data: Methods and Software* (eds G. Parmigiani, E. S. Garrett, R. A. Irizarry and S. L. Zeger), pp. 73–101. New York: Springer.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001) Empirical Bayes analysis of a microarray experiment. *J. Am. Statist. Ass.*, **96**, 1151–1160.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis*. Boca Raton: Chapman and Hall–CRC.

Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. and Hirschhorn, J. N. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.*, **33**, 177–182.

Mandel, M. (2007) Censoring and truncation: highlighting the differences. *Am. Statistn*, **61**, 321–324.

Mandel, M. and Rinott, Y. (2007) On statistical inference under selection bias. *Discussion Paper Series 473*. Center for Rationality and Interactive Decision Theory, Hebrew University, Jerusalem.

Mandel, M. and Rinott, Y. (2009) A selection bias conflict and frequentist versus Bayesian viewpoints. *Am. Statistn*, **64**, 211–217.

Scott, J. G. and Berger, J. O. (2006) An exploration of aspects of Bayesian multiple testing. *J. Statist. Plannng Inf.*, **136**, 2144–2162.

Senn, S. (2008) A note concerning a selection paradox of Dawids. *Am. Statistn*, **62**, 206–210.

Smyth, G. K. (2005) Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions using R and Bioconductor* (eds R. Gentleman, V. Carey, S. Dudoit, R. Irizarry and W. Huber), pp. 397–420. New York: Springer.

Soric, B. (1989) Statistical discoveries and effect-size estimation. *J. Am. Statist. Ass.*, **84**, 608–610.

Storey, J. D. (2002) A direct approach to false discovery rates. *J. R. Statist. Soc.* B, **64**, 479–498.

Storey, J. D. (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.*, **31**, 2013–2035.

Storey, J. D. (2007) The optimal discovery procedure: a new approach to simultaneous significance testing. *J. R. Statist. Soc.* B, **69**, 347–368.

Yekutieli, D. and Benjamini, Y. (1999) A resampling based False Discovery Rate controlling multiple test procedure. *J. Statist. Plannng Inf.*, **82**, 171–196.

Zhong, H. and Prentice, R. L. (2008) Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, **9**, 621–634.

Zollner, S. and Pritchard, J. K. (2007) Overcoming the winners curse: estimating penetrance parameters from case-control data. *Am. J. Hum. Genet.*, **80**, 605–615.