FACETS
a multidisciplinary open access science journal

RESEARCH ARTICLE

Measuring statistical evidence and multiple testing

Michael Evans[a]* and Jabed Tomal[b]

[a]Department of Statistical Sciences, University of Toronto, Toronto, ON M5S 3G3, Canada; [b]Department of Computer and Mathematical Sciences, University of Toronto Scarborough, 1265 Military Trail, Toronto, ON M1C 1A4, Canada

*mevans@utstat.utoronto.ca

Abstract

The measurement of statistical evidence is of considerable current interest in fields where statistical criteria are used to determine knowledge. The most commonly used approach to measuring such evidence is through the use of *p*-values, even though these are known to possess a number of properties that lead to doubts concerning their validity as measures of evidence. It is less well known that there are alternatives with the desired properties of a measure of statistical evidence. The measure of evidence given by the relative belief ratio is employed in this paper. A relative belief multiple testing algorithm was developed to control for false positives and false negatives through bounds on the evidence determined by measures of bias. The relative belief multiple testing algorithm was shown to be consistent and to possess an optimal property when considering the testing of a hypothesis randomly chosen from the collection of considered hypotheses. The relative belief multiple testing algorithm was applied to the problem of inducing sparsity. Priors were chosen via elicitation, and sparsity was induced only when justified by the evidence and there was no dependence on any particular form of a prior for this purpose.

Key words: multiple testing, sparsity, statistical evidence, relative belief ratios, priors, checking for prior–data conflict, relative belief multiple testing algorithm, testing for sparsity

Citation: Evans M and Tomal J. 2018. Measuring statistical evidence and multiple testing. FACETS 3: 563–583. doi:10.1139/facets-2017-0121

Handling Editor: Patrick Ingram

Received: November 20, 2017

Accepted: February 6, 2018

Published: May 25, 2018

Published by: Canadian Science Publishing

Introduction

The need for the measurement of statistical evidence arises as an issue in science as follows. The scientific problem under consideration concerns some quantity of interest for which an investigator either wants to know its value or has a hypothesis that the quantity takes a specific value and wants to know if this is true or false. To answer such a question data x are collected. It is rare that the data provide a definitive answer but it is believed that the data contain evidence concerning this. The purpose of statistical reasoning or inference is to use this evidence to estimate the quantity of interest and provide an assessment of the accuracy of the estimate or indicate whether there is evidence either in favor of or against the hypothesized value, and provide an assessment of the strength of this evidence.

To implement statistical inference, additional ingredients are required. First, it is presumed that the data x can be thought of as having arisen from a probability distribution as represented by the density f. Provided the data were collected properly, this assumption is reasonable and this is assumed here. A consequence of this is that the data can be thought of as being objective in the sense that f fully describes how the data were produced from the set $\mathcal{X}$ of possible data values. Of course, f is generally not known so it is assumed that $f \in \{f_\theta : \theta \in \Theta\}$, a family of probability densities on $\mathcal{X}$ referred to as the

statistical model. Here θ is called the parameter and Θ the parameter space of the model. The quantity of interest is then represented as $\psi = \Psi(\theta)$, where $\Psi{:}\Theta \to \Psi$, and we don't distinguish between the function and its range to save notation.

A natural approach to constructing a theory of inference is to determine a measure of the evidence in the data x that ψ is the true value for each $\psi \in \Psi$. The value $\psi(x) \in \Psi$ that maximizes this measure of evidence is, then, the obvious estimate and a subset $C(x) \subset \Psi$ of values with evidence measures above some threshold would, through a measure of its size, serve to give an assessment of the accuracy of $\psi(x)$. For a null hypothesis $H_0 : \Psi(\theta) = \psi_0$ the measure of evidence at ψ_0 indicates whether there is evidence in favor of or against H_0 and a measure of the strength of this evidence is then obtained by comparing the evidence at ψ_0 with the evidence at each of the other possible values for ψ. A theory that accomplishes this, based on the relative belief ratio $RB(\psi \mid x)$ as the measure of evidence, is described by Evans (2015) and outlined in the section "Statistical analysis based on relative belief".

Even though p-values are commonly used to measure evidence, it has long been recognized that there are serious issues associated with their use (for example, see Royall (1997)). This can be readily observed by considering what the cutoff is to determine when there is evidence against or for a hypothesis. Cutoffs like 5% are not only arbitrary, many treatments of p-values insist that it is not possible for a p-value to give evidence in favor of a null hypothesis. Although that is a perfectly valid statement, it seems like a significant weak point for a supposed measure of evidence. Even when a very small p-value is observed this does not mean that a result of scientific interest has been obtained. For, given the finite accuracy with which measurements are taken, it is rarely the case that the truth of H_0 practically corresponds to an exact value ψ_0. Rather, there is a region about ψ_0 such that if the true value lies in this region, for all practical purposes, H_0 is true. Using relative belief ratios evidence can be obtained either for or against H_0, there is a clear measure of the strength of the evidence, and the essential discreteness involved in assessing H_0 is easily handled.

The theory of relative belief requires an additional ingredient, namely a prior probability distribution π must be specified on Θ that reflects the beliefs concerning what values of θ are more or less likely. The prior is determined by an elicitation algorithm that is an argument as to why the prior in question is to be considered suitable. The prior π is subjective in nature and that seems contrary to the dictates of science, which properly has objectivity as the goal. Although it doesn't justify the use of priors, it is to be noted that the model $\{f_\theta : \theta \in \Theta\}$ is also subjective as it is chosen by the investigator. One could argue in favor of this subjectivity, however, particularly when the choices are being made by an expert, as informed input should result in a better analysis, but doubts linger. Part of our approach to dealing with this concern is to check that any ingredient chosen is not contradicted by the objective data. Therefore, model checking and checking for prior–data conflict are necessary. Also, it is possible to choose a prior such that a desired result is obtained but such bias can be measured and controlled a priori by design. Some discussion on assessing prior–data conflict and bias is provided in the section "Statistical analysis based on relative belief".

The focus of this paper is the following problem. Suppose Ψ is an open subset of R^k and we wish to assess the individual hypotheses $H_{0i} = \{\theta : \Psi_i(\theta) = \psi_{0i}\}$, namely H_{0i} is the hypothesis that the i-th coordinate of ψ equals ψ_{0i}. Considering these hypotheses separately is the multiple testing problem and the concern is to ensure that while controlling the individual error rate, the overall error rate is not too large. An error means either the acceptance of H_{0i} when it is false (a false negative) or the rejection of H_{0i} when it is true (a false positive). One approach is to make an inference about the number of H_{0i} that are true (or false) and then use this to control the number of H_{0i} that are accepted (or rejected). In the section "Inferences for multiple tests", this is shown to work for small k but to fail for large k. As a remedy for this, a relative belief multiple testing algorithm is developed that controls for false positives and false negatives through the use of bounds on the evidence that are determined

by the measurement of bias. This approach is shown to be consistent and to possess an optimal property when considering the assessment of a randomly selected hypothesis from the set of hypotheses.

In the section "Applications", an application is made of the relative belief multiple testing algorithm to the problem of inducing sparsity. If it is known that $\psi_i = \Psi_i(\theta) = \psi_{0i}$, then the effective dimension of the quantity of interest is $k - 1$, which is a simplification of the model. Sometimes there is a belief that many of the hypotheses H_{0i} are true, but there is little prior knowledge about which are true and it is, therefore, not clear how to choose a prior that reflects this belief. A common approach is to use a prior that, together with a particular estimation procedure, forces many of the ψ_i to take the corresponding value ψ_{0i}. For example, the use of a Laplace prior together with using the maximum value of the posterior as the estimate, known as maximum a posteriori (MAP) estimation, is known to accomplish this for certain problems. Problems with this approach include the possibility that such an assignment is simply an artifact of the prior and the estimation procedure and that sparsity requires an overly concentrated prior that leads to prior–data conflict with the coordinates for which H_{0i} is rejected. It would be preferable to have a procedure that was not dependent on a specific form for the prior, avoided prior–data conflict, and was based on the statistical evidence contained in the data, and this is the approach taken here. Practical applications are presented, with special emphasis on regression problems including the situation where the number of predictors exceeds the number of observations.

Evans (2015) noted that there are connections between relative belief and the pure likelihood approach to inference, as both consider statistical evidence as the core concept. This is also reflected in the approach to multiple testing developed in the current paper and that discussed by Strug and Hodge (2006a, 2006b). There have been several priors proposed for the sparsity problem through MAP estimation; for example, the spike-and-slab prior discussed by George and McCulloch (1993) and Rockova and George (2014), the Laplace prior discussed by Park and Casella (2008), and the horseshoe prior of Carvalho et al. (2009). Any prior can be used with the approach taken here, but logically an elicited prior is preferred over one possessing certain properties.

Statistical analysis based on relative belief

Suppose that interest is in inference about the quantity $\Psi(\theta) = \psi$. Let Π_Ψ denote the prior measure of ψ, with density π_Ψ, and let $\Pi_\Psi(\cdot \mid x)$ denote the posterior measure of ψ, with density $\pi_\Psi(\cdot \mid x)$. Evidence is measured by change in belief (for example, see Salmon (1973) or Howson and Urbach (2006)), thus if belief in ψ increases there is evidence in favor of this value and evidence against it if belief decreases. Evans (2015) argued for the relative belief ratio $RB_\Psi(\psi \mid x) = \lim_{\delta \to 0} \Pi_\Psi(N_\delta(\psi) \mid x)/\Pi_\Psi(N_\delta(\psi))$ as a measure of evidence, where $N_\delta(\psi)$ is a sequence of neighborhoods of ψ converging (nicely, as defined by Rudin (1974)) to $\{\psi\}$ as $\delta \to 0$. When π_Ψ and $\pi_\Psi(\cdot \mid x)$ are continuous at ψ, then

$$RB_\Psi(\psi \mid x) = \pi_\Psi(\psi \mid x)/\pi_\Psi(\psi) \tag{1}$$

So $RB_\Psi(\psi \mid x) > 1$ indicates evidence in favor of ψ, $RB_\Psi(\psi \mid x) < 1$ indicates evidence against, and $RB_\Psi(\psi \mid x) = 1$ gives no evidence either way. Any 1–1 increasing function of $RB_\Psi(\cdot \mid x)$ is an equivalent measure of evidence and $RB_\Psi(\cdot \mid x)$ is invariant under smooth reparameterizations, thus relative belief inferences are invariant to these choices.

The best estimate of ψ is the value that maximizes the evidence, namely $\psi(x) = \arg\sup RB_\Psi(\psi \mid x)$. Associated with this is a γ-credible region $C_{\Psi,\gamma}(x) = \{\psi : RB_\Psi(\psi \mid x) \geq c_{\Psi,\gamma}(x)\}$ containing those values whose evidence is above the threshold $c_{\Psi,\gamma}(x) = \inf\{k : \Pi_\Psi(RB_\Psi(\psi \mid x) > k \mid x) \leq \gamma\}$. As $\psi(x) \in C_{\Psi,\gamma}(x)$, for every $\gamma \in [0, 1]$, for selected γ, the "size" of $C_{\Psi,\gamma}(x)$ is a measure of the accuracy of $\psi(x)$. A calibration of $RB_\Psi(\psi_0 \mid x)$ is given by the strength

$$\Pi_\Psi(RB_\Psi(\psi \mid x) \leq RB_\Psi(\psi_0 \mid x) \mid x) \tag{2}$$

When $RB_\Psi(\psi_0 \mid x) < 1$, a small value of eq. (2) indicates a large posterior belief that the true value has a relative belief ratio greater than $RB_\Psi(\psi_0 \mid x)$, and therefore there is strong evidence against ψ_0 but only weak evidence against it if eq. (2) is big. If $RB_\Psi(\psi_0 \mid x) > 1$, a large value of eq. (2) indicates a small posterior probability that the true value has a relative belief ratio greater than $RB_\Psi(\psi_0 \mid x)$, and therefore there is strong evidence in favor of ψ_0, whereas a small value of eq. (2) only indicates weak evidence in favor of ψ_0. A variety of optimality and consistency results have been established for these inferences (see Evans (2015)).

When $H_0 : \Psi(\theta) = \psi_0$ is false both eqs. (1) and (2) converge to 0, and when H_0 is true then eq. (1) converges to the maximum possible value, which is always >1. When H_0 is true and there are only a finite number of possible values for ψ then eq. (2) converges to 1, but in the continuous case eq. (2) can converge to a $U(0,1)$ distribution. The view is taken here, however, that any time continuous probability is used this is an approximation to a finite, discrete context. For example, if ψ is a mean and the response measurements are to the nearest centimeter, then of course the true value of ψ cannot be known to an accuracy >0.5 cm, no matter how large the sample is. Furthermore, there are implicit bounds associated with any measurement process. As such, the restriction can be made to discretized parameters that take only a finite number of values. Thus, when ψ is a continuous, real-valued parameter, it is discretized to the intervals $\ldots, (\psi_0 - 3\delta/2, \psi_0 - \delta/2], (\psi_0 - \delta/2, \psi_0 + \delta/2], (\psi_0 + \delta/2, \psi_0 + 3\delta/2], \ldots$ for some choice of $\delta > 0$, and there are only a finite number of such intervals covering the range of possible values. With this discretization, then $H_0 = (\psi_0 - \delta/2, \psi_0 + \delta/2]$ and eq. (2) is consistent. Thus, δ needs to be specified as part of the application, at least when the goal is assessing the evidence concerning H_0. The value of δ is simply the smallest difference from ψ_0 that matters in the application and presumably a knowledgeable scientist knows what this is and designs the measurement process that produces the data accordingly.

Let $A \subset \mathcal{X}$ be such that H_0 is accepted whenever $x \in A$, thus, with $M(\cdot \mid H_0)$ denoting the prior predictive measure given that H_0 is true, $M(A \mid H_0)$ is the prior probability of accepting H_0 when it is true. The relative belief acceptance region is $A_{rb}(\psi_0) = \{x : RB_\Psi(\psi_0 \mid x) > 1\}$. Let $R \subset \mathcal{X}$ be such that H_0 is rejected whenever $x \in R$ and the relative belief rejection region is $R_{rb}(\psi_0) = \{x : RB_\Psi(\psi_0 \mid x) < 1\}$. Letting M denote the unconditional prior predictive measure the following result was proved by Evans (2015).

Theorem 1: (i) The acceptance region $A_{rb}(\psi_0)$ minimizes $M(A)$ among all acceptance regions A, satisfying $M(A \mid H_0) \geq M(A_{rb}(\psi_0) \mid H_0)$. (ii) The rejection region $R_{rb}(\psi_0)$ maximizes $M(R)$ among all rejection regions R, satisfying $M(R \mid H_0) \leq M(R_{rb}(\psi_0) \mid H_0)$.

The implication of this is that, when $\Pi_\Psi(\{\psi_0\}) = 0$, then $A_{rb}(\psi_0)$ minimizes the prior probability that H_0 is accepted given that it is false among all acceptance regions A satisfying the condition in (i) and $R_{rb}(\psi_0)$ maximizes the prior probability that H_0 is rejected given that it is false among all rejection regions R satisfying the condition in (ii). The same result holds for the case when $\Pi_\Psi(\{\psi_0\}) > 0$ with the inequalities in (i) and (ii) replaced by equalities. Under independent identically distributed (IID) sampling, $M(A_{rb}(\psi_0) \mid H_0) \to 1$ and $M(R_{rb}(\psi_0) \mid H_0) \to 0$ as sample size increases, so these quantities can be controlled by design. Theorem 1 can be generalized to obtain optimality results for the acceptance region $A_{rb,\,q}(\psi_0) = \{x : RB_\Psi(\psi_0 \mid x) > q\}$ and the rejection region $R_{rb,\,q}(\psi_0) = \{x : RB_\Psi(\psi_0 \mid x) < q\}$. The following inequality is useful in the section "Inferences for multiple tests" in controlling error rates.

Theorem 2: $M(R_{rb,\,q}(\psi_0) \mid \psi_0) \leq q$.

Proof: By the Savage–Dickey result (see proposition 4.2.7 in Evans (2015)), $RB_\Psi(\psi_0 \mid x) = m(x \mid \psi_0)/m(x)$. Now $E_{M(\cdot \mid \psi_0)}(m(x)/m(x \mid \psi_0)) = 1$, and therefore, by Markov's inequality, $M(R_{rb,q}(\psi_0) \mid \psi_0) = M(m(x)/m(x \mid \psi_0) > 1/q \mid \psi_0) \leq q$.

One of the key concerns with Bayesian inference methods is that the prior can bias the analysis. Given a measure of evidence, however, it is possible to measure and control bias. The bias against H_0 is given by $M(RB_\Psi(\psi_0 \mid x) \leq 1 \mid \psi_0) = 1 - M(A_{rb}(\psi_0) \mid \psi_0)$ as this is the prior probability that evidence will be obtained against H_0 when it is true. If the bias against H_0 is large, subsequently reporting, after seeing the data, that there is evidence against H_0 is not convincing. The bias in favor of H_0 is given by $M_T(RB_\Psi(\psi_0 \mid x) \geq 1 \mid \psi_0')$ for values $\psi_0' \neq \psi_0$ such that the difference between ψ_0 and ψ_0' represents the smallest difference of practical importance; note that this tends to decrease as ψ_0' moves farther away from ψ_0. When the bias in favor is large, subsequently reporting, after seeing the data, that there is evidence in favor of H_0 is not convincing. For a fixed prior, both biases decrease with sample size and thus, in design situations, they can be used to set sample size and thereby control bias.

It is never known that the ingredients chosen for a statistical analysis are correct, but hopefully these serve as useful approximations in the sense that inferences drawn from them are reasonably accurate. If x lies in the tails of f_θ for every $\theta \in \Theta$, then it can be concluded that there is a problem with the model and it needs to be modified. It is clear that checking the prior is a meaningless activity if the model is to be discarded, thus model checking is carried out first. If the model passes, then the prior is checked and the approach of Evans and Moshonov (2006) is adopted here. For this let T be a minimal sufficient statistic (MSS) for the model with density m_T, and if the probability $M_T(m_T(t) \leq m_T(T(x)))$ is small, then conclude a prior–data conflict exists as this says that $T(x)$ lies in the tails of the prior-predictive. The consistency of this procedure was established by Evans and Jang (2011a) as, under weak conditions this probability converges to $\Pi(\pi(\theta) \leq \pi(\theta_{true}))$, and a methodology for modifying a prior that fails its checks was developed by Evans and Jang (2011b).

Inferences for multiple tests

Consider now the multiple testing problem. The typical approach to this problem relies on the use of p-values that, for the reasons discussed, are not adopted here. Rather, the relative belief ratio as a valid measure of statistical evidence is used as the basis for all inferences.

To see what the problem is with multiple testing suppose that Ψ_i is finite for each i, perhaps arising via a discretization as discussed in the section "Statistical analysis based on relative belief", and let $\xi = \Xi(\theta) = k^{-1} \sum_{i=1}^{k} I_{H_{0i}}(\Psi_i(\theta))$ be the proportion of the hypotheses H_{0i} that are true. Note that the discreteness is essential, otherwise, under a continuous prior on Ψ, the prior distribution of $\Xi(\theta)$ is degenerate at 0. In an application it is desirable to make inference about the true value of $\xi \in \Xi = \{0, 1/k, 2/k, \ldots, 1\}$ and this is based on the relative belief ratio $RB_\Xi(\xi \mid x) = \Pi(\Xi(\theta) = \xi \mid x)/\Pi(\Xi(\theta) = \xi)$. The appropriate estimate of Ξ is $\xi(x) = \arg\sup_\xi RB_\Xi(\xi \mid x)$ and its accuracy is assessed using the size of $C_{\Xi,\gamma}(x)$ for some choice of $\gamma \in [0, 1]$. Hypotheses such as $H_0 = \{\theta : \Xi(\theta) \in [\xi_0, \xi_1]\}$, namely the proportion true is at least ξ_0 and no greater than ξ_1, is assessed using the relative belief ratio $RB(H_0 \mid x) = \Pi(\xi_0 \leq \Xi(\theta) \leq \xi_1 \mid x)/\Pi(\xi_0 \leq \Xi(\theta) \leq \xi_1)$, which equals $RB_\Xi(\xi_0 \mid x)$ when $\xi_0 = \xi_1$.

The estimate $\xi(x)$ can be used to control how many hypotheses are potentially accepted. For this, select $k\xi(x)$ of the H_{0i} as being true from among those for which $RB_{\Psi_i}(\psi_{0i} \mid x) > 1$. Note that it does not make sense to accept H_{0i} when $RB_{\Psi_i}(\psi_{0i} \mid x) < 1$ as there is evidence against H_{0i}. Thus, if there are fewer than $k\xi(x)$ satisfying $RB_{\Psi_i}(\psi_{0i} \mid x) > 1$, then fewer than this number should be accepted. If there are more than $k\xi(x)$ of the relative belief ratios satisfying $RB_{\Psi_i}(\psi_{0i} \mid x) > 1$, then some method will have to be used to select the $k\xi(x)$ that are potentially accepted. It is clear, however, that the logical way to do this is to order the H_{0i}, for which $RB_{\Psi_i}(\psi_{0i} \mid x) > 1$, based on their strengths $\Pi_\Psi(RB_{\Psi_i}(\psi_{0i} \mid x) \leq RB_{\Psi_i}(\psi_{0i} \mid x) \mid x)$, from largest to smallest, and accept at most the $k\xi(x)$ for which the evidence is strongest. If control is desired of the number of false positives then the relevant parameter of interest is $\upsilon = \Upsilon(\theta) = 1 - \Xi(\theta)$, the proportion of false hypotheses. Note that $\Pi(\Upsilon(\theta) = \upsilon) = \Pi(\Xi(\theta) = 1 - \upsilon)$, and therefore the relative belief estimate of υ satisfies $\upsilon(x) = 1 - \xi(x)$. Following the same procedure, the H_{0i} with

$RB_{\Psi_i}(\psi_{0i} \mid x) < 1$ are ranked via their strengths and at most $k\upsilon(x)$ are rejected. This procedure will be referred to as the multiple testing algorithm.

The consistency of the multiple testing algorithm follows from results proved by Evans (2015) (see section 4.7.1 therein) under IID sampling. In other words, as the amount of data increases, $\xi(x)$ converges to the proportion of H_{0i} that are true, each $RB(\psi_{0i} \mid x)$ converges to the largest possible value (always >1) when H_{0i} is true and converges to 0 when H_{0i} is false, and the evidence in favor or against converges to the strongest possible, depending on whether the hypothesis in question is true or false.

The following example demonstrates the characteristics of the algorithm.

Example 1. Location normal.

Suppose that there are k independent samples x_{ij} for $1 \le i \le k$, $1 \le j \le n$, where the i-th sample is from a $N(\mu_i, \sigma^2)$ distribution with μ_i unknown and σ^2 known. It is desired to assess the evidence as to whether or not $H_{0i} : \mu_i = \mu_0$ is true for $i = 1, \ldots, k$. It is easy to modify our development to allow the sample sizes to vary, and the case where σ^2 is unknown is considered in the section "Applications". This context is relevant to the analysis of microarray data. The statistic $T(x) = (\overline{x}_1, \ldots, \overline{x}_k)$ is an MSS for this model, and thus a natural model checking procedure is to compare the observed value of the statistic $\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \overline{x}_i)^2/\sigma^2$ to the chi-squared($k(n-1)$) distribution.

For the prior, the $\mu_1, \ldots, \mu_k$ are taken to be IID from a $N(\mu_0, \lambda_0^2\sigma^2)$ distribution. The value of λ_0^2 is determined via elicitation. For this it is supposed that it is known with virtual certainty that each $\mu_i \in (m_l, m_u)$ for specified values $m_l \le m_u$. Here, virtual certainty is interpreted to mean that the prior probability of this interval is at least γ, where γ is a large probability like 0.99. It is also supposed that $\mu_0 = (m_l + m_u)/2$. This implies that $\lambda_0 = (m_u - m_l)/(2\sigma\Phi^{-1}((1+0.99)/2))$. Following Evans and Jang (2011b), increasing the value of λ_0 implies a more weakly informative prior in this context and, as such, decreases the possibility of prior–data conflict, and this indicates how the prior is to be modified in case of prior–data conflict. Note that this elicitation argument also specifies μ_0 when this is not predetermined. The prior distribution of T is $N_k(\mu_0 1_k, \sigma^2(\lambda_0^2 + 1/n)I_k)$, where 1_k is the k-dimensional vector of 1s and I_k is the $k \times k$ identity matrix, and therefore the check on the prior becomes the probability $P(\chi_k^2 \ge \sum_{i=1}^{k}(\overline{x}_i - \mu_0)^2/\sigma^2(\lambda_0^2 + 1/n))$, where $\chi_k^2 \sim$ chi-squared(k).

The posteriors of the μ_i are independent $\mu_i \mid x \sim N(\mu_i(x), (n\lambda_0^2+1)^{-1}\lambda_0^2\sigma^2)$, where $\mu_i(x) = (n + 1/\lambda_0^2)^{-1}(n\overline{x}_i + \mu_0/\lambda_0^2)$. Given that the measurements are taken to finite accuracy it is not realistic to test $\mu_i = \mu_0$. A value $\delta > 0$ is specified so that $H_{0i} = (\mu_0 - \delta/2, \mu_0 + \delta/2]$ in a discretization of R^1 into a finite number of intervals, each of length δ, as well as two tail intervals. For some $D \in \mathbb{N}$ there are $2D+1$ intervals $I_d = (\mu_0 + (d-1/2)\delta, \mu_0 + (d+1/2)\delta]$ for $d \in \{-D, -D+1, \ldots, D\}$ that span (m_l, m_u), together with the tail intervals $(-\infty, \mu_0 - (D+1/2)\delta]$ and $(\mu_0 + (D+1/2)\delta, \infty)$. Then $RB_i(I_d \mid x) = \{\Phi((d+1/2)\delta/\lambda_0\sigma) - \Phi((d-1/2)\delta/\lambda_0\sigma)\}^{-1} \times \{\Phi((n\lambda_0^2+1)^{1/2}(\mu_0 + (d+1/2)\delta - \mu_i(x))/\lambda_0\sigma) - \Phi((n\lambda_0^2+1)^{1/2}(\mu_0 + (d-1/2)\delta - \mu_i(x))/\lambda_0\sigma)\}$, with a similar formula for the tail intervals. When δ is small this is approximated by the ratio of the posterior to prior densities of μ_i evaluated at $\mu_0 + d\delta$. Then $RB(I_0 \mid x) = RB_i(H_{0i} \mid x)$ gives the evidence for or against H_{0i} and the strength of this evidence is computed using the discretized posterior distribution. Notice that $RB_i(H_{0i} \mid x)$ converges to ∞ as $\lambda_0 \to \infty$ and this is characteristic of other measures of evidence such as Bayes factors. As discussed by Evans (2015), this is one of the reasons why calibrating eq. (1) via eq. (2) is necessary.

Now, consider the bias in the prior. To simplify matters, the continuous approximation is used as this makes little difference here (see **Tables 3** and **4**). The bias against $\mu_i = \mu_0$ equals

$$M(RB_i(\mu_0 \mid x) \leq 1 \mid \mu_0) = 2(1 - \Phi(a_n(1))) \tag{3}$$

where

$$a_n(q) = \begin{cases} \{(1 + 1/n\lambda_0^2)\log((n\lambda_0^2 + 1)/q^2)\}^{1/2}, & q^2 \leq n\lambda_0^2 + 1 \\ 0, & q^2 > n\lambda_0^2 + 1 \end{cases}$$

Note that eq. (3) converges to $2(1 - \Phi(1)) = 0.32$ as $\lambda_0 \to 0$ and to 0 as $\lambda_0 \to \infty$ and, for fixed λ_0, converges to 0 as $n \to \infty$. Thus, there is never strong bias against $\mu_i = \mu_0$; this is as expected because the prior is centered on μ_0. The bias in favor of $\mu_i = \mu_0$ is measured by

$$M(RB_i(\mu_0 \mid x) \geq 1 \mid \mu_0 \pm \delta/2) = \Phi(\sqrt{n}\delta/2\sigma + a_n(1)) - \Phi(\sqrt{n}\delta/2\sigma - a_n(1)) \tag{4}$$

As $\lambda_0 \to \infty$ eq. (4) converges to 1, thus there is bias in favor of $\mu_i = \mu_0$ and this reflects what was obtained for the limiting value of $RB_i(H_{0i} \mid x)$. Also, eq. (4) decreases with increasing δ and goes to 0 as $n \to \infty$; thus, bias of both types can be controlled by sample size. Perhaps the most important take away from this discussion, however, is that by using a supposedly noninformative prior with λ_0 large, bias in favor of the H_{0i} is being induced.

Consider, first, a simulated data set x when $k = 10$, $n = 5$, $\sigma = 1$, $\delta = 1$, $\mu_0 = 0$, $(m_l, m_u) = (-5, 5)$, so that $\lambda_0 = 10/2\Phi^{-1}(0.995) = 1.94$ and suppose $\mu_1 = \mu_2 = \ldots = \mu_7 = 0$, with the remaining $\mu_i = 2$. The relative belief ratio function $RB_\Xi(\cdot \mid x)$ is plotted in Fig. 1. In this case, the relative belief estimate $\xi(x) = 0.70$ is exactly correct. Table 1 gives the values of the $RB_i(0 \mid x)$ together with their strengths. It is clear that the multiple testing algorithm leads to 0 false positives and 0 false negatives. Therefore, the algorithm works perfectly on these data, but of course it can't be expected to perform as well when the three nonzero means move closer to 0. Also, it is worth noting that the strength of the evidence in favor of $\mu_i = 0$ is very strong for $i = 1, 2, 3, 5, 6, 7$, but only moderate when $i = 4$. The strength of the evidence against $\mu_i = 0$ is very strong for $i = 8, 9, 10$. The maximum possible value of $RB_i((\mu_0 - \delta/2, \mu_0 + \delta/2] \mid x)$ is $(2\Phi(\delta/2\lambda_0\sigma) - 1)^{-1} = 4.92$, thus some of the relative belief ratios are relatively large.

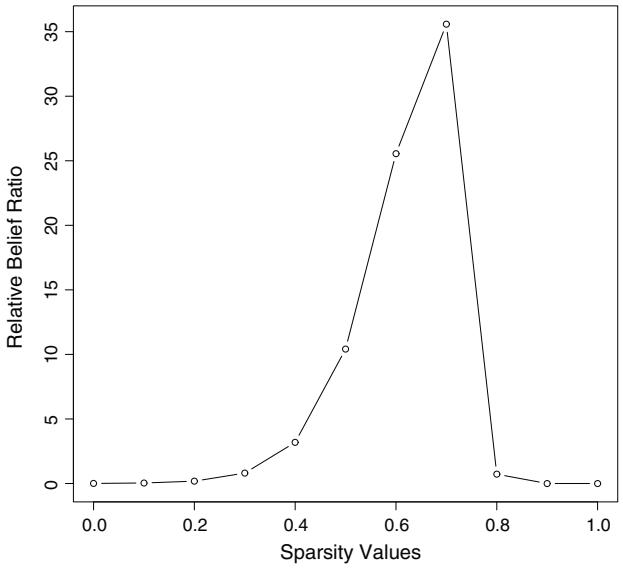

Fig. 1. A plot of the relative belief ratio of Ξ when $n = 5$, $k = 10$, and 7 means equal 0 with the remaining means equal to 2 in Example 1 with $\delta = 1$.

Table 1. Relative belief ratios and strengths for the μ_i in Example 1 with $k = 10$, $\delta = 1.0$.

i	1	2	3	4	5
μ_i	0	0	0	0	0
$RB_i(0 \mid x)$	3.27	3.65	2.98	1.67	3.57
Strength	1.00	1.00	1.00	0.37	1.00
i	6	7	8	9	10
μ_i	0	0	2	2	2
$RB_i(0 \mid x)$	3.00	3.43	2.09×10^{-4}	3.99×10^{-4}	8.80×10^{-3}
Strength	1.00	1.00	4.25×10^{-5}	8.11×10^{-5}	1.83×10^{-3}

Table 2. Relative belief ratios and strengths for the μ_i in Example 1 with $k = 10$, $\delta = 0.5$.

i	1	2	3	4	5
μ_i	0	0	0	0	0
$RB_i(0 \mid x)$	3.58	4.17	3.15	1.43	4.64
Strength	0.62	1.00	0.59	0.26	1.00
i	6	7	8	9	10
μ_i	0	0	2	2	2
$RB_i(0 \mid x)$	3.18	3.83	3.25×10^{-5}	6.76×10^{-5}	2.37×10^{-3}
Strength	0.59	1.00	3.30×10^{-6}	7.00×10^{-6}	2.47×10^{-4}

To investigate sensitivity to the choice of δ several smaller values were considered. Table 2 gives the relevant entries for the same sample as Table 1 when $\delta = 0.5$. The relative belief ratios do not change by much and still give evidence in the right direction. Some of the strengths do change, particularly for $i = 1$ and $i = 6$, which now indicate a bit weaker evidence in favor. In this case, $\xi(x) = 0.60$. Repeating these calculations with $\delta = 0.1$ gives similar results, with the relative belief ratios staying about the same but the strengths getting weaker, and now $\xi(x) = 0.50$. The insensitivity of the RB_i to δ is expected, as the data should increase belief in the interval $(\mu_0 - \delta/2, \mu_0 + \delta/2]$ when H_{0i} is true and decrease it when it is false. It is to be noted, however, that δ is not a tuning parameter of the algorithm but is determined by scientific knowledge in the application as the smallest difference from μ_0 of practical importance.

Now, consider basically the same context but with $k = 1000$, $\mu_1 = \ldots = \mu_{700} = 0$ and the remaining $\mu_i = 2$. In this case, $\xi(x) = 0.47$, which is a serious underestimate. As such, the multiple testing algorithm will not record enough acceptances and will fail. This problem arises due to the independence of the μ_i. For the prior distribution of $k\Xi(\theta)$ is binomial$(k, 2\Phi(\delta/2\lambda_0\sigma) - 1)$ and the prior distribution of $k\Upsilon(\theta)$ is binomial $(k, 2(1 - \Phi(\delta/2\lambda_0\sigma)))$. Thus, the a priori expected proportion of true hypotheses is $2\Phi(\delta/2\lambda_0\sigma) - 1$ and the expected proportion of false hypotheses is $2(1 - \Phi(\delta/2\lambda_0\sigma))$. When $\delta/2\lambda_0\sigma$ is small, as when the amount of sampling variability or the diffuseness of the prior are large, then the prior on Ξ suggests a belief in many false hypotheses. When k is small, the data can override this to produce accurate inferences about ξ or υ, but otherwise, large amounts of data are needed that may not be available. Contrary to what is sometimes claimed, testing multiple hypotheses is also a problem in a Bayesian framework.

Example 1 makes it clear that, in general, accurate inference about ξ and υ is not feasible in high-dimensional contexts without large amounts of data. Rather than focus on estimating the proportion of true or false hypotheses, however, we consider an approach designed to protect against false positives or false negatives. It is often the case that when evidence against a hypothesis is obtained it prompts some kind of action, and a user may wish to prevent too many that are spurious. Alternatively, the user may be concerned with too many false negatives, as this may conceal a discovery of real value.

The entries in **Tables 1** and **2** point to a feasible approach to these problems by focusing instead on the evidence concerning the individual μ_i, as these parameters do not depend on high-dimensional aspects of the full model parameter like ξ and υ do. To control the actions taken based on the evidence, constants q_R and q_A, where $0 < q_R \leq 1 \leq q_A$, are used as follows: classify H_{0i} as accepted when $RB_i(\psi_{0i} \mid x) > q_A$ and as rejected when $RB_i(\psi_{0i} \mid x) < q_R$. Note that those accepted always have evidence in favor, whereas those rejected always have evidence against. The strengths can also be quoted to assess the reliability of these inferences. Provided q_R is greater than the minimum possible value of $RB_i(\cdot \mid x)$, and this is typically 0, and the q_A chosen is less than the maximum possible value of $RB_i(\psi_{0i} \mid x)$, and this is 1 over the prior probability of H_{0i}, then this procedure is consistent as the amount of data increases. In fact, the related estimates of ξ and υ are also consistent. The price paid for this is that a hypothesis will not be classified whenever $q_R \leq RB_i(\psi_{0i} \mid x) \leq q_A$. Not classifying a hypothesis implies that there is not enough evidence for this purpose and more data are required. This approach is referred to as the relative belief multiple testing algorithm.

It remains to determine q_A and q_R. Consider, first, protecting against too many false positives. The a priori conditional prior probability, given that H_{0i} is true, of finding evidence against H_{0i} less than q_R satisfies $M(RB_i(\psi_{0i} \mid X) < q_R \mid \psi_{0i}) \leq q_R$ by Theorem 2. Naturally, we want the probability of a false positive to be small, and choosing q_R small accomplishes this. The a priori probability that a randomly selected hypothesis produces a false positive is

$$\frac{1}{k}\sum_{i=1}^{k} M(RB_i(\psi_{0i} \mid X) < q_R \mid \psi_{0i}) \tag{5}$$

which is bounded above by q_R and thus converges to 0 as $q_R \to 0$. Also, for fixed q_R, eq. (5) converges to 0 as the amount of data increases. More generally q_R can be allowed to depend on i, but when the ψ_i are similar in nature this does not seem necessary. Furthermore, it is not necessary to weight the hypotheses equally, therefore a randomly chosen hypothesis with unequal probabilities could be relevant in certain circumstances. In any case, controlling the value of eq. (5), whether by sample size or by the choice of q_R, is clearly controlling for false positives. Suppose there is proportion p_{FP} of false positives that is just tolerable in a problem. Then, q_R can be chosen so that eq. (5) is less than or equal to p_{FP}; note that $q_R = p_{FP}$ satisfies this.

Similarly, if $\psi_{0i}' \neq \psi_{0i}$ then $M(RB_i(\psi_{0i} \mid X) > q_A \mid \psi_{0i}')$ is the prior probability of accepting H_{0i} when ψ_{0i}' is the true value. For a given effect size δ of practical importance it is natural to take $\psi_{0i}' = \psi_{0i} \pm \delta/2$. In typical applications this probability decreases the "farther" ψ_{0i}' is from ψ_{0i}, and choosing q_A to make this probability small will make it small for all meaningful alternatives. Under these circumstances the a priori probability that a randomly selected hypothesis produces a false negative is bounded above by

$$\frac{1}{k}\sum_{i=1}^{k} M(RB_i(\psi_{0i} \mid X) > q_A \mid \psi_{0i}') \tag{6}$$

As $q_A \to \infty$, or as the amount of data increases with q_A fixed, then eq. (6) converges to 0 and the number of false negatives can be controlled. If there is proportion p_{FN} of false negatives that is just tolerable in a problem, then q_A can be chosen so that eq. (6) is less than or equal to p_{FN}.

The following optimality result holds for relative belief multiple testing.

Corollary 1: (i) Among all procedures for which the prior probability of accepting H_{0i}, when it is true, is at least $M(RB_i(\psi_{0i} \mid X) > q_A \mid \psi_{0i})$ for $i = 1, \ldots, k$, the relative belief multiple testing algorithm minimizes the prior probability that a randomly chosen hypothesis is accepted. (ii) Among all procedures for which the prior probability of rejecting H_{0i}, when it is true, is less than or equal to $M(RB_i(\psi_{0i} \mid X) < q_R \mid \psi_{0i})$, then the relative belief multiple testing algorithm maximizes the prior probability that a randomly chosen hypothesis is rejected.

Proof: For (i) consider a procedure for multiple testing and let A_i be the set of data values where H_{0i} is accepted. Then, by hypothesis $M(RB_i(\psi_{0i} \mid X) > q_A \mid \psi_{0i}) \leq M(A_i \mid \psi_{0i})$ and by the analog of Theorem 1, $M(A_i) \geq M(RB_i(\psi_{0i} \mid X) > q_A)$. Applying this to a randomly chosen H_{0i} gives the result. The proof of (ii) is basically the same.

Applying the same discussion as after Theorem 1, it is seen that, under reasonable conditions, the relative belief multiple testing algorithm minimizes the prior probability of accepting a randomly chosen H_{0i} when it is false and maximizes the prior probability of rejecting a randomly chosen H_{0i} when it is false. This establishes an optimality result for the relative belief multiple testing algorithm.

Consider now the application of the relative belief multiple testing algorithm in the previous example.

Example 2. Location normal example, continued.

In this context, $M(RB_i(\mu_0 \mid x) < q_R \mid \mu_0) = 2(1 - \Phi(a_n(q_R))$ for all i and, therefore, this is the value of eq. (5). Therefore, q_R is chosen to make this number suitably small. Table 3 records values for eq. (5) for both the continuous and discretized cases. From this it is seen that for small n there can be some bias against H_{0i} when $q_R = 1$, and thus the prior probability of obtaining false positives is perhaps too large. Table 3 demonstrates that choosing a smaller value of q_R can adequately control the prior probability of false positives.

Table 3. Prior probability that a randomly chosen hypothesis produces a false positive when $\delta/\sigma = 1$, continuous and discretized () versions, in Example 2.

n	λ_0	q_R	(5)	n	λ_0	q_R	(5)
1	1	1	0.239 (0.228)	5	1	1	0.143 (0.097)
		1/2	0.041 (0.030)			1/2	0.051 (0.022)
		1/10	0.001 (0.000)			1/10	0.006 (0.001)
	2	1	0.156 (0.146)		2	1	0.074 (0.041)
		1/2	0.053 (0.045)			1/2	0.031 (0.013)
		1/10	0.005 (0.004)			1/10	0.005 (0.001)
	10	1	0.031 (0.026)		10	1	0.013 (0.004)
		1/2	0.014 (0.011)			1/2	0.006 (0.002)
		1/10	0.002 (0.002)			1/10	0.001 (0.001)

For false negatives, consider eq. (6), where

$$M(RB_i(\mu_0 \,|\, x) > q_A \,|\, \mu_0 \pm \delta/2) = \begin{cases} \Phi(\sqrt{n}\delta/2\sigma + a_n(q_A)) - \Phi(\sqrt{n}\delta/2\sigma - a_n(q_A)), & 1 \le q_A^2 \le n\lambda_0^2 + 1 \\ 0, & q_A^2 > n\lambda_0^2 + 1 \end{cases}$$

for all i. It is easy to show that this is monotone decreasing in δ, and therefore it is an upper bound on the expected proportion of false negatives among those hypotheses that are actually false. The cutoff q_A can be chosen to make this number as small as desired. When $\delta/\sigma \to \infty$, then eq. (6) converges to 0 and increases to $2\Phi(a_n(q_A)) - 1$ as $\delta/\sigma \to 0$. Table 4 records values for eq. (6) when $\delta/\sigma = 1$ so that the μ_i differ from μ_0 by one half of a standard deviation. There is clearly some improvement but the bias in favor of false negatives is still readily apparent. It would seem that taking $q_A = \sqrt{n\lambda_0^2 + 1}$ gives the best results, but this could be considered s quite conservative. It is also worth remarking that all the entries in Table 4 can be considered very conservative when large effect sizes are expected.

Now, consider the situation when $k = 1000$, $n = 5$, $\delta = 1$ and $\lambda_0 = 1.94$ is the elicited value. From Table 3 with $q_R = 1.0$ about 8% false positives are expected a priori, and from Table 4 with $q_A = 1.0$ a worst case upper bound on the a priori expected percentage of false negatives is about 75%. The top part of Table 5 indicates that with $q_R = q_A = 1.0$, then 4.9% (34 of 700) false positives and 0.1% (3 of 300) false negatives were obtained. With these choices of the cutoffs all hypotheses are classified. Certainly the upper bound 75% seems far too pessimistic in light of the results, but recall that Table 4

Table 4. Prior probability that a randomly chosen hypothesis produces a false negative when $\delta/\sigma = 1$, continuous and discretized () versions, in Example 2.

n	λ_0	q_A	(6)	n	λ_0	q_A	(6)
1	1	1.0	0.704 (0.715)	5	1	1.0	0.631 (0.702)
		1.2	0.527 (0.503)			2.0	0.302 (0.112)
		1.4	0.141 (0.000)			2.4	0.095 (0.000)
	2	1.0	0.793 (0.805)		2	1.0	0.747 (0.822)
		2.0	0.359 (0.304)			3.0	0.411 (0.380)
		2.2	0.141 (0.000)			4.5	0.084 (0.000)
	10	1.0	0.948 (0.955)		10	1.0	0.916 (0.961)
		5.0	0.708 (0.713)			10.0	0.552 (0.588)
		10.0	0.070 (0.000)			22.0	0.080 (0.000)

Table 5. Confusion matrices for Example 2 with $k = 1000$ when 700 of the μ_i equal 0 and 300 of the μ_i equal 2.

Decision	$\mu = 0$	$\mu = 2$
Accept $\mu = 0$ using $q_A = 1.0$	666	3
Reject $\mu = 0$ using $q_R = 1.0$	34	297
Not classified	0	0
Accept $\mu = 0$ using $q_A = 3.0$	419	0
Reject $\mu = 0$ using $q_R = 0.5$	9	287
Not classified	272	13

is computed at the false values $\mu = \pm 0.5$. The relevant a priori expected percentage of false negatives when $\mu = \pm 2.0$ is about 3.5%. The bottom part of Table 5 gives the relevant values when $q_R = 0.5$ and $q_A = 3.0$. In this case, there are 2.1% (9 of 428) false positives and 0% false negatives, but 39.9% (272 of 700) of the true hypotheses and 4.3% (13 of 300) of the false hypotheses were not classified as the relevant relative belief ratio lay between q_R and q_A. Thus, being more conservative has reduced the error rates, but with the drawback that a large proportion of the true hypotheses don't get classified. The procedure has worked well in this example, but of course the error rates can be expected to rise when the false values move towards the null and improve when they move away from the null.

What is implemented in an application depends on the goals. If the primary purpose is to protect against false positives, then Table 3 indicates that this is accomplished fairly easily. Protecting against false negatives is more difficult; as the actual effect sizes are not known a decision has to be made. Note that choosing a cutoff is equivalent to saying that one will only accept H_{0i} if the belief in the truth of H_{0i} has increased by a factor at least as large as q_A. Computations such as those in Table 4 can be used to provide guidance, but there is no avoiding the need to be clear about what effect sizes are deemed to be important or the need to obtain more data when this is necessary. With the relative belief multiple testing algorithm error rates are effectively controlled, but there may be many true hypotheses not classified.

The idea of controlling the prior probability of a randomly chosen hypothesis yielding a false positive or a false negative via eq. (5) or eq. (6), respectively, can be extended. For example, consider the prior probability that a random sample of l from k hypotheses yields at least one false positive

$$\frac{1}{\binom{k}{l}} \sum_{\{i_1,\ldots,i_l\}\subset\{1,\ldots,k\}} M\left(\begin{array}{c}\text{at least one of } RB_{i_j}(\psi_{0i_j}\,|\,X) < q_R \\ \text{for } j = 1,\ldots,l\,|\,\psi_{0i_1},\ldots,\psi_{0i_l}\end{array}\right) \tag{7}$$

In the context of the examples in this paper, and many others, the term in eq. (7) corresponding to $\{i_1,\ldots,i_l\}$ equals $M(\text{at least one of } RB_{i_j}(\psi_{0i_j}\,|\,X) < q_R \text{ for } j = 1,\ldots,l\,|\,\psi_0)$. The following result leads to an interesting property for eq. (7).

Lemma 1: Let $(\Omega, \mathcal{F}, P)$ be a probability model and $\mathcal{B} = \{A_1,\ldots,A_k\} \subset \mathcal{F}$. The probability that at least one of $l \le k$ randomly selected events from $\mathcal{B}$ occurs is increasing in l.

Proof: Let $\Delta(i)$ be the event that exactly i of $A_1,\ldots,A_k \in \mathcal{F}$ occur, so that $\cup_{i=1}^k A_i = \cup_{i=1}^k \Delta(i)$; note that the $\Delta(i)$ are mutually disjoint. When $l < k$,

$$S_{l,k} = \sum_{\{i_1,\ldots,i_l\}\subset\{1,\ldots,k\}} I_{A_{i_1}\cup\ldots\cup A_{i_l}} = \binom{k}{l}\sum_{i=0}^{l-1} I_{\Delta(k-i)} + \sum_{i=l}^{k-1}\left[\binom{k}{l} - \binom{i}{l}\right] I_{\Delta(k-i)}$$
$$= \binom{k}{l}\sum_{i=0}^{k-1} I_{\Delta(k-i)} - \sum_{i=l}^{k-1}\binom{i}{l} I_{\Delta(k-i)}$$

and $S_{k,k} = I_{A_1\cup\ldots\cup A_k}$. Now, consider $\binom{k}{l}^{-1} S_{l,k} - \binom{k}{l-1}^{-1} S_{l,k}$, which equals

$$\frac{1}{\binom{k}{l}} \sum_{\{i_1,\ldots,i_l\}\subset\{1,\ldots,k\}} I_{A_{i_1}\cup\ldots\cup A_{i_l}} - \frac{1}{\binom{k}{l-1}} \sum_{\{i_1,\ldots,i_{l-1}\}\subset\{1,\ldots,k\}} I_{A_{i_1}\cup\ldots\cup A_{i_{l-1}}} \tag{8}$$

If $l = k$, then eq. (8) equals $I_{A_1 \cup \ldots \cup A_k} - \sum_{i=0}^{k-1} I_{\Delta(k-i)} + I_{\Delta(1)} = I_{A_1 \cup \ldots \cup A_k} - \sum_{i=0}^{k-2} I_{\Delta(k-i)}$, which is nonnegative. If $l < k$, then eq. (8) equals $\binom{k}{l-1}^{-1} I_{\Delta(k-l+1)} + \sum_{i=l}^{k-1} \left[\binom{i}{l-1} \binom{k}{l-1}^{-1} - \binom{i}{l} \binom{k}{l}^{-1} \right] I_{\Delta(k-i)}$, which is nonnegative because an easy calculation shows that each term in the second sum is nonnegative. The expectation of eq. (8) is then nonnegative and this establishes the result.

It follows, by taking $A_i = \{x : RB_i(\psi_{0i} \mid x) < q_R\}$, that eq. (7) is an upper bound on the prior probability that a random sample of l' hypotheses yields at least one false positive whenever $l' \leq l$. Thus, eq. (7) leads to a more rigorous control over the possibility of false positives. A similar result is obtained for false negatives.

Applications

We now consider the sparsity problem.

Example 3. Testing for sparsity.

Consider the context of Example 1. A natural approach to inducing sparsity is to estimate μ_i by μ_0 whenever $RB_i(\mu_0 \mid x) > q_A$. From the simulation it is seen that this works extremely well when $q_A = 1$ for both $k = 10$ and $k = 1000$. It also works when $k = 1000$ and $q_A = 3$, in the sense that the error rate is low, but it is conservative in the amount of sparsity it induces in that case. Again, the goals of the application will dictate what is appropriate.

Another Bayesian method for inducing sparsity is to use the Bayesian Lasso as per Park and Casella (2008) and based on Tibshirani (1996). The prior here is a product of independent Laplace distributions, namely $\prod_{i=1}^{k} [(\sqrt{2}\lambda_0\sigma)^{-k} \times \exp\{-(\sqrt{2}/\lambda_0\sigma) \sum_{i=1}^{k} |\mu_i - \mu_0|\}]$, where σ is assumed known and μ_0, λ_0 are hyperparameters. Note that each Laplace prior has mean μ_0 and variance $\lambda_0^2\sigma^2$. Using the elicitation algorithm provided in Example 1 but replacing the normal prior with a Laplace prior leads to the assignment $\mu_0 = (m_l + m_u)/2$, $\lambda_0 = (m_u - m_l)/2\sigma G^{-1}(0.995)$, where $G^{-1}(p) = 2^{-1/2} \log 2p$ when $p \leq 1/2$, $G^{-1}(p) = -2^{-1/2} \log 2(1-p)$, where $p \geq 1/2$ and G^{-1} denotes the quantile function of a Laplace distribution with mean 0 and variance 1. With the specifications used in the simulations of Example 1, this leads to $\mu_0 = 0$ and $\lambda_0 = 1.54$, which implies a smaller variance than the value $\lambda_0 = 1.94$ used with the normal prior, and therefore the Laplace prior is more concentrated about 0.

The posteriors for the μ_i are independent with the density for μ_i proportional to $\exp\{-n(\bar{x}_i - \mu_i)^2/2\sigma^2 - \sqrt{2}\,|\mu_i - \mu_0|\,/\lambda_0\sigma\}$ giving the MAP estimator

$$\mu_{i\text{MAP}}(x) = \begin{cases} \bar{x}_i + \sqrt{2}\sigma/\lambda_0 n, & \bar{x}_i < \mu_0 - \sqrt{2}\sigma/\lambda_0 n \\ \mu_0, & \mu_0 - \sqrt{2}\sigma/\lambda_0 n \leq \bar{x}_i \leq \mu_0 + \sqrt{2}\sigma/\lambda_0 n \\ \bar{x}_i - \sqrt{2}\sigma/\lambda_0 n, & \bar{x}_i > \mu_0 + \sqrt{2}\sigma/\lambda_0 n \end{cases}$$

The MAP estimate of μ_i is sometimes forced to equal μ_0, although this effect is negligible whenever $\sqrt{2}\sigma/\lambda_0 n$ is small.

The Lasso induces sparsity through estimation by taking λ_0 to be small. By contrast, the evidential approach, based on the normal prior and the relative belief ratio, induces sparsity through taking λ_0 large. The advantage to this latter approach is that by taking λ_0 large, prior–data conflict is avoided. When taking λ_0 small, the potential for prior–data conflict increases, as the true values can be deep

Table 6. Confusion matrices using Lasso with $k = 1000$ when 700 of the μ_i equal 0 and 300 of the μ_i equal 2 in Example 3.

Decision	$\mu = 0$	$\mu = 2$
Accept $\mu = 0$ using $q_A = 1.0$	227	0
Reject $\mu = 0$ using $q_A = 1.0$	473	300

into the tails of the prior. For example, for the simulations of Example 1, $\sqrt{2}\sigma/\lambda_0 n = 0.183$, which is smaller than the $\delta/2 = 0.5$ used in the relative belief approach with the normal prior. Therefore, it can be expected that the Lasso will do worse here, and this is reflected in Table 6 in which there are far too many false negatives. To improve this, the value of λ_0 needs to be reduced; however, note that this is determined by an elicitation and there is the risk of then encountering prior–data conflict. Another possibility is to implement the evidential approach with the elicited Laplace prior and the discretization as done with the normal prior, and then results similar to those obtained in Example 1 can be expected.

It is also interesting to compare the MAP estimation approach and the relative belief approach with respect to the conditional prior probabilities of μ_i being assigned the value μ_0 when the true value actually is μ_0. It is easily seen that, based on the Laplace prior, $M(\mu_{i\text{MAP}}(x) = \mu_0 \mid \mu_0) = 2\Phi(\sqrt{2}/\lambda_0\sqrt{n}) - 1$, and this converges to 0 as $n \to \infty$ or $\lambda_0 \to \infty$. For the relative belief approach $M(RB_i(\mu_0 \mid x) > q_A \mid \mu_0)$ is the relevant probability. With either the normal or Laplace prior $M(RB_i(\mu_0 \mid x) > q_A \mid \mu_0)$ converges to 1 both as $n \to \infty$ and as $\lambda_0 \to \infty$. Therefore, with enough data the correct assignment is always made using relative belief but not with MAP based on the Laplace prior.

The Laplace and normal priors work equally with the relative belief multiple testing algorithm but there are no advantages to using the Laplace prior. One could argue too that the singularity of the Laplace prior at its mode makes it an odd choice and there doesn't seem to be a good justification for this. Furthermore, the computations are harder with the Laplace prior, particularly with more complex models, and therefore using a normal prior is preferable overall.

An example with considerable practical significance is now considered.

Example 4. Full rank regression.

Suppose the basic model is given by $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + z = \beta_0 + x'\beta_{1:k} + z$, where the x_i are predictor variables, $z \sim N(0, \sigma^2)$ and β and σ^2 are unknown. The problem of interest is testing $H_{0i} : \beta_i = 0$ for $i = 1, \ldots, k$ to establish which variables have any effect on the response. It is assumed that the observed values of the predictor variable have been standardized so that for observations $(y, X) \in R^n \times R^{n\times(k+1)}$, where $X = (\mathbf{1}, \mathbf{x}_1, \ldots, \mathbf{x}_k)$ is of rank $k+1$, then $\mathbf{1}'\mathbf{x}_i = 0$ and $||x_i||^2 = 1$ for $i = 1, \ldots, k$. Note that (b, s), where $b = (X'X)^{-1}X'y$ and $s = ||y - Xb||$, is an MSS for this model, and model checking can be carried out by considering functions of the standardized residuals $r = (y - Xb)/s$ as this has a distribution independent of (β, σ^2). The skewness and kurtosis statistics are such functions and it is straightforward to simulate from their distributions to determine if their observed values are surprising.

The prior distribution of (β, σ^2) is taken to be

$$\beta \mid \sigma^2 \sim N_{k+1}(0, \sigma^2\Sigma_0), 1/\sigma^2 \sim \text{gamma}_{\text{rate}}(\alpha_1, \alpha_2) \tag{9}$$

for some hyperparameters Σ_0 and (α_1, α_2). Note that this may entail subtracting a known, fixed constant from each y value so that the prior for β_0 is centered at 0. Taking 0 as the central value for the priors on the remaining β_i seems appropriate when the primary concern is whether or not each x_i is having any effect. The marginal prior for β_i is then $\{(\alpha_2/\alpha_1)\sigma_{0ii}^2\}^{1/2} t_{2\alpha_1}$, where $t_{2\alpha_1}$ denotes the t distribution on $2\alpha_1$ degrees of freedom, for $i = 0, \ldots, k$. Hereafter, we will take $\Sigma_0 = \lambda_0^2 I_{k+1}$ although it is easy to generalize to more complicated choices.

The elicitation of the hyperparameters is carried out via an extension of a method developed by Cao et al. (2014) for the multivariate normal distribution. Suppose that it is known with virtual certainty, based on our knowledge of the measurements being taken, that $\beta_0 + x'\beta_{1:k}$ will lie in the interval $(-m_0, m_0)$ for some $m_0 > 0$ for all $x \in R$, where R is a compact set centered at 0. On account of the standardization, $R \subset [-1, 1]^k$. Again "virtual certainty" is interpreted as probability greater than or equal to γ, where γ is some large probability like 0.99. Therefore, the prior on β must satisfy $2\Phi(m_0/\sigma\lambda_0\{1 + x'x\}^{1/2}) - 1 \geq \gamma$ for all $x \in R$, and this implies that

$$\sigma \leq m_0/\lambda_0 \tau_0 z_{(1+\gamma)/2} \tag{10}$$

where $\tau_0^2 = 1 + \max_{x \in R} ||x||^2 \leq 1 + k$ with equality when $R = [-1, 1]^k$.

An interval that will contain a response value y with virtual certainty, given predictor values x, is $\beta_0 + x'\beta_{1:k} \pm \sigma z_{(1+\gamma)/2}$. Suppose that we have lower and upper bounds s_1 and s_2 on the half-length of this interval so that $s_1 \leq \sigma z_{(1+\gamma)/2} \leq s_2$ or, equivalently,

$$s_1/z_{(1+\gamma)/2} \leq \sigma \leq s_2/z_{(1+\gamma)/2} \tag{11}$$

holds with virtual certainty. Combining eq. (11) with eq. (10) implies $\lambda_0 = m_0/s_2\tau_0$.

To obtain the relevant values of α_1 and α_2 let $G(\alpha_1, \alpha_2, \cdot)$ denote the cdf of the $\text{gamma}_{\text{rate}}(\alpha_1, \alpha_2)$ distribution, and note that $G(\alpha_1, \alpha_2, z) = G(\alpha_1, 1, \alpha_2 z)$. Therefore, the interval for $1/\sigma^2$ implied by eq. (11) contains $1/\sigma^2$ with virtual certainty, when α_1, α_2 satisfy $G^{-1}(\alpha_1, \alpha_2, (1+\gamma)/2) = s_1^{-2} z_{(1+\gamma)/2}^2$, $G^{-1}(\alpha_1, \alpha_2, (1-\gamma)/2) = s_2^{-2} z_{(1-\gamma)/2}^2$, or equivalently

$$G(\alpha_1, 1, \alpha_2 s_1^{-2} z_{(1+\gamma)/2}^2) = (1+\gamma)/2 \tag{12}$$

$$G(\alpha_1, 1, \alpha_2 s_2^{-2} z_{(1-\gamma)/2}^2) = (1-\gamma)/2 \tag{13}$$

It is a simple matter to solve these equations for (α_1, α_2). For this choose an initial value for α_1 and, using eq. (12), find z such that $G(\alpha_1, 1, z) = (1+\gamma)/2$, which implies $\alpha_2 = z/s_1^{-2} z_{(1+\gamma)/2}^2$. If the left side of eq. (13) is less (or greater) than $(1-\gamma)/2$, then decrease (or increase) the value of α_1 and repeat step 1. Continue iterating this process until satisfactory convergence is attained.

Evans and Moshonov (2006) showed that when checking for prior–data conflict in such a context it is better to check the components of the prior sequentially as this helps to pinpoint where any failure in the prior occurs. First, the prior on σ^2 is checked using the tail probability based on the prior predictive for s and, if this component passes, then the prior on β is checked based on the conditional prior-predictive of b given s. If conflict is found, the methods discussed by Evans and Jang (2011b) are available to modify the prior.

Assuming that X is of rank $k+1$, the posterior of (β, σ^2) is given by

$$\beta \,|\, y, \sigma^2 \sim N_{k+1}(\beta(X, y), \sigma^2 \Sigma(X)), \quad 1/\sigma^2 \,|\, y \sim \text{gamma}_{\text{rate}}((n + 2\alpha_1)/2, \alpha_2(X, y)/2) \tag{14}$$

where $\beta(X,y)=\Sigma(X)X'Xb, \Sigma(X)=(X'X+\Sigma_0^{-1})^{-1}$ and $\alpha_2(X,y)=||y-Xb||^2+(Xb)'(I_n-X\Sigma(X)X')Xb+2\alpha_2$. Then the marginal posterior for β_i is given by $\beta_i(X,y)+\{\alpha_2(X,y)\sigma_{ii}(X)/(n+2\alpha_1)\}^{1/2}t_{n+2\alpha_1}$ and the relative belief ratio for β_i at 0 equals

$$RB_i(0\mid X,y)=\frac{\Gamma\left(\frac{n+2\alpha_1+1}{2}\right)\Gamma(\alpha_1)}{\Gamma\left(\frac{2\alpha_1+1}{2}\right)\Gamma\left(\frac{n+2\alpha_1}{2}\right)}\left(1+\frac{\beta_i^2(X,y)}{\alpha_2(X,y)\sigma_{ii}(X)}\right)^{-\frac{n+2\alpha_1+1}{2}}\times\left(\frac{\alpha_2(X,y)\sigma_{ii}(X)}{\alpha_2^2\lambda_0^2}\right)^{-\frac{1}{2}} \tag{15}$$

Rather than using eq. (15), however, the distributional results are used to compute the discretized relative belief ratios as in Example 1. For this $\delta>0$ is required to determine an appropriate discretization and it will be assumed here that this is the same for all the β_i, although the procedure can be easily modified if this is not the case in practice. Note that such a δ is effectively determined by the amount that $x_i\beta_i$ will vary from 0 for $x\in R$. As $x_i\in[-1,1]$, then $|x_i\beta_i|\le\delta$ provided that $|\beta_i|\le\delta$. When this variation is suitably small as to be immaterial, then such a δ is appropriate for saying β_i is effectively 0. Determination of the hyperparameters and δ is dependent on the application.

Again inference can be made concerning $\xi=\Xi(\beta,\sigma^2)$, the proportion of the β_i effectively equal to 0. As in Example 1, however, we can expect bias when the amount of variability in the data is large relative to δ or the prior is too diffuse. To implement the relative belief multiple testing algorithm, the quantities eqs. (5) and (6) need to be computed to determine q_R and q_A, respectively. The conditional prior distribution of $(b,||y-Xb||^2)$, given (β,σ^2), is $b\sim N_{k+1}(\beta,\sigma^2(X'X)^{-1})$, statistically independent of $||y-Xb||^2\sim\text{gamma}((n-k-1)/2,\sigma^{-2}/2)$. Thus, computing eqs. (5) and (6) can be carried out by generating (β,σ^2) from the relevant conditional prior, generating $(b,||y-Xb||^2)$ given (β,σ^2), and using eq. (15).

To illustrate these computations the diabetes data set discussed by Efron et al. (2004) and Park and Casella (2008) is now analyzed. With $\gamma=0.99$, the values $m_0=100$, $s_1=75$, $s_2=200$ were used to determine the prior together with $\tau_0=1.05$ determined from the X matrix. This led to the values $\lambda_0=0.48$, $\alpha_1=7.29$, $\alpha_2=13641.35$ being chosen for the hyperparameters. Using the methods developed by Evans and Moshonov (2006), a first check was made on the prior on σ^2 against the data, and a tail probability equal to 0.19 was obtained indicating there is no prior–data conflict with this prior. Given no prior–data conflict at the first stage, the prior on β was then checked and the relevant tail probability of 0.00 was obtained indicating a strong degree of conflict. Following the argument of Evans and Jang (2011b), the value of λ_0 was increased to choose a prior that was weakly informative with respect to our initial choice. This led to choosing the value $\lambda_0=5.00$, and the relevant tail probability equals 0.32, so there is no conflict.

Using this prior, the relative belief estimates, ratios, and strengths are recorded in **Table 7**. This shows that there is strong evidence against $\beta_i=0$ for the variables sex, bmi, map, and ltg and no evidence against $\beta_i=0$ for any other variables. There is strong evidence in favor of $\beta_i=0$ for age and ldl, moderate evidence in favor of $\beta_i=0$ for the constant, tc, tch, and glu and perhaps only weak evidence in favor of $\beta_i=0$ for hdl.

As previously discussed, it is necessary to consider the issue of bias, namely to compute the prior probability of getting a false positive for different choices of q_R and the prior probability of getting a false negative for different choices of q_A. The value of eq. (5) is 0.0003 when $q_R=1$, and therefore there is virtually no bias in favor of false positives and one can feel confident that the predictors identified as having an effect do so. The story is somewhat different, however, when considering the possibility of false negatives via eq. (6). For example, with $q_A=1$ then eq. (6) equals 0.9996 and when $q_A=100$ then eq. (6) equals 0.7998. Thus, there is substantial bias in favor of the null hypotheses and undoubtedly this is due to the diffuseness of the prior. The implication is that we cannot be entirely confident

Table 7. Relative belief estimates, relative belief ratios, and strengths for assessing no effect for the diabetes data in Example 4.

Variable	Estimates	$RB_i(0 \mid X, y)$	Strength
Constant	2	2454.86	0.44
age	−4	153.62	0.95
sex	−224	0.13	0.00
bmi	511	0.00	0.00
map	314	0.00	0.00
tc	162	33.23	0.36
ldl	−20	57.65	0.90
hdl	167	27.53	0.15
tch	114	49.97	0.37
ltg	496	0.00	0.00
glu	77	66.81	0.23

concerning those β_i assigned to be equal to 0. Recall that the first prior proposed led to prior–data conflict, and thus a much more diffuse prior obtained by increasing λ_0 was substituted. The bias in favor of false negatives with this prior could be reduced by making the prior less diffuse by lowering λ_0, but we know that if it is lowered too much prior–data conflict arises. Thus, there is a trade-off between lowering the bias in favor and avoiding prior–data conflict. In any case, determining a value of λ_0 in such a fashion seems inappropriate becauase then the prior becomes too dependent on the data and we do not advocate this. The real cure for the bias in an application is to collect more data, and the amount necessary can be determined by the bias calculations.

Next we consider the application to regression with $k+1>n$.

Example 5. Non-full rank regression.

In a number of applications $k+1>n$ and thus X is of rank $l<n$. In this situation, suppose $\{\mathbf{x}_1,\ldots,\mathbf{x}_l\}$ forms a basis for $\mathcal{L}(\mathbf{x}_1,\ldots,\mathbf{x}_k)$, perhaps after relabeling the predictors, and write $X=(\mathbf{1}\ X_1\ X_2)$, where $X_1=(\mathbf{x}_1\ldots\mathbf{x}_l)$. For given $r=(X_1\ X_2)\beta_{1:k}$ there will be many solutions $\beta_{1:k}$. A particular solution is given by $\beta_{1:k^*}=(X_1(X_1'X_1)^{-1}\ 0)'r$. The set of all solutions is then given by $\beta_{1:k^*}+\ker(X_1\ X_2)$, where $\ker(X_1\ X_2)=\{(-B'\ I_{k-l})'\eta:\eta\in R^{k-l}\}$, $B=(X_1'X_1)^{-1}X_1'X_2$, and the columns of $C=(-B'\ I_{k-l})'$ give a basis for $\ker(X_1\ X_2)$. As sparsity is expected for $\beta_{1:k}$, it is natural to consider the solution that minimizes $||\beta_{1:k}||^2$ for $\beta_{1:k}\in\beta_{1:k^*}+\mathcal{L}(C)$. Using $\beta_{1:k^*}$, and applying the Sherman–Morrison–Woodbury formula to $C(C'C)^{-1}C'$, this is given by the Moore–Penrose solution

$$\beta_{1:k}^{MP}=(I_k-C(C'C)^{-1}C')\beta_{1:k^*}=(I_l\ B)'\omega_{1:l} \tag{16}$$

where $\omega_{1:l}=(I_l+BB')^{-1}(\beta_{1:l}+B\beta_{l+1:k})$.

From eq. (9) with $\Sigma_0=\lambda_0^2 I_{k+1}$, the conditional prior distribution of $(\beta_0,\ \omega_{1:l})$ given σ^2 is $\beta_0\mid\sigma^2\sim N(0,\sigma^2\lambda_0^2)$, independent of $\omega_{1:l}\mid\sigma^2\sim N_l(0,\ \sigma^2\lambda_0^2(I_l+BB')^{-1})$, which, using eq. (16), implies $\beta_{1:k}^{MP}\mid\sigma^2\sim N_k(0,\ \sigma^2\Sigma_0(B))$, conditionally independent of β_0, where

$$\Sigma_0(B)=\lambda_0^2\begin{pmatrix}(I_l+BB')^{-1} & (I_l+BB')^{-1}B\\ B'(I_l+BB')^{-1} & B'(I_l+BB')^{-1}B\end{pmatrix}$$

With $1/\sigma^2 \sim \text{gamma}_{\text{rate}}(\alpha_1, \alpha_2)$, this implies that the unconditional prior of the i-th coordinate of $\beta_{1:k}^{MP}$ is $(\lambda_0^2 \alpha_2 \sigma_{ii}^2(B)/\alpha_1)^{1/2} t_{2\alpha_1}$.

Putting $X_* = (\mathbf{1}\ X_1 + X_2 B')$ gives the full rank model $y \,|\, \beta_0, \omega_{1:l}, \sigma^2 \sim N_n(X_*(\beta_0, \omega_{1:l}')', \sigma^2 I_n)$. As in Example 4, then $(\beta_0, \omega_{1:l}) | y, \sigma^2 \sim N_l(\omega(X_*, y), \sigma^2 \Sigma(X_*)), 1/\sigma^2 | y \sim \text{gamma}_{\text{rate}}((n+2\alpha_1)/2, \alpha_2(X_*, y)/2)$ where $\omega(X_*, y) = \Sigma(X_*) X_*' X_* b_*, b_* = (X_*' X_*)^{-1} X_*' y$ and

$$\Sigma^{-1}(X_*) = \begin{pmatrix} n & 0 \\ 0 & (X_1 + X_2 B')'(X_1 + X_2 B') \end{pmatrix} + \lambda_0^{-2} \begin{pmatrix} 1 & 0 \\ 0 & (I_l + BB') \end{pmatrix},$$
$$\alpha_2(X_*, y) = ||y - X_* b_*||^2 + (X_* b_*)'(I_n - X_* \Sigma(X_*) X_*') X_* b_* + 2\alpha_2$$

Now, noting that $(X_1 + X_2 B')'(X_1 + X_2 B') = (I_l + BB') X_1' X_1 (I_l + BB')$, this implies $b_*' = (\bar{y}, (I_l + BB')^{-1} b_1)$, where $b_1 = (X_1' X_1)^{-1} X_1' y$ is the least-squares estimate of $\beta_{1:l}$, and

$$\Sigma(X_*) = \begin{pmatrix} n + \lambda_0^{-2} & 0 \\ 0 & (I_l + BB') X_1' X_1 (I_l + BB') + \lambda_0^{-2} (I_l + BB') \end{pmatrix}^{-1},$$
$$\omega(X_*, y) = \Sigma(X_*) X_*' X_* b_* = \begin{pmatrix} n\bar{y}/(n + \lambda_0^{-2}) \\ (I_l + BB' + \lambda_0^{-2}(X_1' X_1)^{-1})^{-1} b_1 \end{pmatrix}$$

Using eq. (16), then $\beta_0 \,|\, y, \sigma^2 \sim N(n(n + \lambda_0^{-2})^{-1} \bar{y}, \sigma^2 (n + \lambda_0^{-2})^{-1})$, independent of $\beta_{1:k}^{MP} \,|\, y, \sigma^2 \sim N_k(\beta^{MP}(X, y), \sigma^2 \Sigma^{MP}(X))$, where

$$\beta^{MP}(X, y) = \begin{pmatrix} D b_1 \\ B' D b_1 \end{pmatrix},\ \Sigma^{MP}(X) = \begin{pmatrix} E & EB \\ B'E & B'EB \end{pmatrix}$$

with $D = (I_l + BB' + \lambda_0^{-2}(X_1' X_1)^{-1})^{-1}$ and $E = ((I_l + BB')(X_1' X_1)(I_l + BB') + \lambda_0^{-2}(I_l + BB'))^{-1}$. The marginal posterior for β_i^{MP} is then given by $\beta_i^{MP}(X, y) + \left\{ \alpha_2(X_*, y) \sigma_{ii}^{MP}(X)/(n + 2\alpha_1) \right\}^{1/2} t_{n+2\alpha_1}$. Relative belief inferences for the coordinates of $\beta_{1:k}^{MP}$ can now be implemented just as in Example 4.

We consider a numerical example in which there is considerable sparsity. For this let $X_1 \in R^{n \times l}$ be formed by taking the second through l-th columns of the $(l+1)$-dimensional Helmert matrix, repeating each row m times and then normalizing. Thus, $n = m(l+1)$ and the columns of X_1 are orthonormal and orthogonal to $\mathbf{1}$. It is supposed that the first $l_1 \leq l$ of the variables giving rise to the columns of X_1 have $\beta_i \neq 0$, whereas the last $l - l_1$ have $\beta_i = 0$ and that the variables corresponding to the first $l_2 \leq k - l$ columns of $X_2 = X_1 B \in R^{n \times (k-l)}$ have $\beta_i \neq 0$, whereas the last $k - l - l_2$ have $\beta_i = 0$. The matrix B is obtained by generating $B = \text{diag}(B_1, B_2)$, where $B_1 = (z_1/||z_1|| \ldots z_{l_2}/||z_{l_2}||)$ with $z_1, \ldots, z_{l_2} \overset{i.i.d.}{\sim} N_{l_1}(0, I)$ independent of $B_2 = (z_{l_2+1}/||z_{l_2+1}|| \ldots z_{k-l-l_2}/||z_{l_{k-l-l_2}}||)$ with $z_{l_2+1}, \ldots, z_{k-l-l_2}$ IID $N_{l-l}(0, I)$. Note that this ensures that the columns of X_2 are all standardized. Furthermore, because it is assumed that the last $l - l_1$ variables of X_1 and the last $k - l - l_2$ variables of X_2 don't have an effect, B is necessarily of the diagonal form given. For, if it was allowed that the last $k - l - l_2$ columns of X_2 were linearly dependent on the the first l_1 columns of X_1, then this would induce a dependence on the corresponding variables, and this is not the intention in the simulation. Similarly, if the first l_2 columns of X_2 were dependent on the last $l - l_1$ columns of X_1, then this would imply that the variables associated with these columns of X_1 have an effect, and this is not the intention.

The sampling model is then prescribed by setting $l = 10$, $l_1 = 5$, $l_2 = 2$, with $\beta_i = 4$ for $i = 1, \ldots, 5, 11, 12$ with the remaining $\beta_i = 0$, $\sigma^2 = 1$, $m = 2$, therefore $n = 22$ and we consider various values of $k \geq l$. Note that a different data set was generated for each value of k. The prior is specified as in Example 4, where the values $\lambda_0^2 = 4$, $\alpha_1 = 11$, $\alpha_2 = 12$ were chosen so that there will be no prior–data conflict arising with the generated data. Also, we considered several values for the discretization parameter δ.

Table 8. Confusion matrices for the numerical example in Example 5.

$k = 10$	Classified positive	Classified negative	Total
True positive	5	0	5
True negative	1	4	5
Total	6	4	10
$k = 20$	Classified positive	Classified negative	Total
True positive	7	0	7
True negative	0	13	13
Total	7	13	20
$k = 50$	Classified positive	Classified negative	Total
True positive	7	0	7
True negative	0	43	43
Total	7	43	50
$k = 100$	Classified positive	Classified negative	Total
True positive	7	0	7
True negative	0	93	93
Total	7	93	100

A hypothesis was classified as true if the relative belief ratio was >1 and classified as false if it was <1. Table 8 gives the confusion matrices with $\delta = 0.1$. The value $\delta = 0.5$ was also considered, but there was no change in the results.

One fact stands out immediately, namely that in all of these examples only one misclassification was made and this was in the full rank ($k = 10$) case where one hypothesis that was true was classified as a positive. The effect sizes that exist are reasonably large, and thus it can't be expected that the same performance will arise with much smaller effect sizes, but it is clear that the approach is robust to the number of hypotheses considered. It should also be noted, however, that the amount of data is relatively small and the success of the procedure will only improve as this increases. This result can, in part, be attributed to the fact that a logically sound measure of evidence is being used.

Conclusions

The relative belief approach to inference has been applied to problems of practical significance. The central feature is that the inferences are based upon a proper measure of evidence. This approach avoids many of the problems that arise with p-values. For example, there is a natural cutoff to determine when there is either evidence for or against. Given a measure of evidence, a concern with Bayesian methodology can be addressed, namely determining whether or not the ingredients bias the results. Bias calculations play a key role in the multiple testing algorithm and its application to sparsity through the a priori control of false positives and negatives.

There are a number of ingredients that need to be selected to implement the relative belief multiple testing algorithm. Perhaps the most important of these is the model and the most controversial is the prior. For the prior, elicitation algorithms have been provided for each example based on the user being able to specify bounds on parameters that hold with virtual certainty. Given that a measurement process was used in the data collection, this implies restrictions for the values of parameters. For

example, suppose interest is in the mean of a response variable corresponding to some kind of length. Each length is measured to a certain accuracy and there is an upper bound on what length can be obtained using a particular measurement technology. Thus, such bounds on the mean response are definitely available and how tight they are depends on what additional information is available on the context. It is also worth noting that there is no reason why some other elicitation algorithm cannot be used if this is felt to be appropriate. There is also the choice of (q_R, q_A), but these are chosen based on the bias calculations to control for false positives and false negatives and the user will have to select these after considering what proportions of errors are tolerable.

The value of δ in hypothesis assessment problems is seemingly another choice but practical aspects of the measurement process involved in data collection dictate what values make sense. For example, there is no point in considering differences from a mean >0.5 cm if the measurements producing the data are only taken to this accuracy. This provides a lower bound on δ and the application may allow for a larger value. It is comforting, however, that results are reasonably robust to this choice. Determining δ for an arbitrary parameter of interest ψ is not necessarily straightforward, but some guidance, when ψ is a probability and δ is either absolute or relative error, can be found in the work of Al-Labadi et al. (2017).

No mention has been made in the paper concerning the false discovery rate (FDR) approach to multiple testing. Current approaches base this on p-values, but presumably there is no reason why a valid measure of evidence such as the relative belief ratio couldn't be used instead. It should also be noted that the FDR approach is somewhat different as it does not imply control over both false positive and false negatives, which has been our intent here. The relationship between the approach of this paper and controlling something like the FDR is a topic for further investigation.

Acknowledgements

Michael Evans was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The authors thank two reviewers for a number of helpful comments.

Author contributions

ME and JT conceived and designed the study. ME and JT performed the experiments/collected the data. ME and JT analyzed and interpreted the data. ME and JT contributed resources. ME and JT drafted or revised the manuscript.

Competing interests

ME is currently serving as a Subject Editor for FACETS, but was not involved in review or editorial decisions regarding this manuscript.

Data accessibility statement

All relevant data are within the paper.

References

Al-Labadi L, Baskurt Z, and Evans M. 2017. Goodness of fit for the logistic regression model using relative belief. Journal of Statistical Distributions and Applications, 4: 17. DOI: 10.1186/s40488-017-0070-7

Cao Y, Evans M, and Guttman I. 2014. Bayesian factor analysis via concentration. *In* Current trends in Bayesian methodology with applications. *Edited by* SK Upadhyay, U Singh, DK Dey, and A Loganathan. CRC Press, Boca Raton, Florida, USA. pp. 181–201.

Carvalho CM, Polson NG, and Scott JG. 2009. Handling sparsity via the horseshoe. Journal of Machine Learning Research, 5: 73–80.

Efron B, Hastie T, Johnstone I, and Tibshirani R. 2004. Least angle regression. The Annals of Statistics, 32: 407–499.

Evans M. 2015. Measuring statistical evidence using relative belief. Monographs on Statistics and Applied Probability 144. CRC Press, Boca Raton, Florida, USA.

Evans M, and Jang GH. 2011a. A limit result for the prior predictive applied to checking for prior-data conflict. Statistics & Probability Letters, 81: 1034–1038. DOI: 10.1016/j.spl.2011.02.025

Evans M, and Jang GH. 2011b. Weak informativity and the information in one prior relative to another. Statistical Science, 26(3): 423–439. DOI: 10.1214/11-STS357

Evans M, and Moshonov H. 2006. Checking for prior-data conflict. Bayesian Analysis, 1(4): 893–914. DOI: 10.1214/06-BA129

George EI, and McCulloch RE. 1993. Variable selection via Gibbs sampling. Journal of the American Statistical Association, 88: 881–889. DOI: 10.1080/01621459.1993.10476353

Howson C, and Urbach P. 2006. Scientific reasoning: the Bayesian approach. 3rd edition. Open Court, Chicago, Illinois, USA.

Park R, and Casella G. 2008. The Bayesian Lasso. Journal of the American Statistical Association, 103: 681–686. DOI: 10.1198/016214508000000337

Rockova V, and George EI. 2014. EMVS: the EM approach to Bayesian variable selection. Journal of the American Statistical Association, 109(506): 828–846. DOI: 10.1080/01621459.2013.869223

Royall R. 1997. Statistical evidence: a likelihood paradigm. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press, Boca Raton, Florida, USA.

Rudin W. 1974. Real and complex analysis. 2nd edition. McGraw-Hill, New York, New York, USA.

Salmon W. 1973. Confirmation. Scientific American, 228(5): 75–83. DOI: 10.1038/scientificamerican0573-75

Strug LJ, and Hodge SE. 2006a. An alternative foundation for the planning and evaluation of linkage analysis. I. Decoupling 'error probabilities' from 'measures of evidence'. Human Heredity, 61: 166–188. PMID: 16865000 DOI: 10.1159/000094709

Strug LJ, and Hodge SE. 2006b. An alternative foundation for the planning and evaluation of linkage analysis. II. Implications for multiple test adjustments. Human Heredity, 61: 200–209. PMID: 16877867 DOI: 10.1159/000094775

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B, 58(1): 267–288.