J. Indian Inst. Sci.
A Multidisciplinary Reviews Journal
ISSN: 0970-4140 Coden-JIISAD

REVIEW ARTICLE

Empirical Bayes and Selective Inference

Daniel García Rasines[1] *and G. Alastair Young*[2*]

Abstract | We review the empirical Bayes approach to large-scale inference. In the context of the problem of inference for a high-dimensional normal mean, empirical Bayes methods are advocated as they exhibit risk-reducing shrinkage, while establishing appropriate control of frequentist properties of the inference. We elucidate these frequentist properties and evaluate the protection that empirical Bayes provides against selection bias.

Keywords: *Empirical Bayes, James–Stein estimator, Normal means problem, Randomisation, Selective inference, Shrinkage*

1 Introduction

Two significant challenges to contemporary statistical inference relate to: the scale of typical problems encountered nowadays; the need for adaptivity of inference to selection bias.

Modern scientific technology routinely produces thousands, potentially even more, parallel but related data sets, each with its own related testing or estimation problem: what is often characterised as 'large-scale inference'[5]. In such contemporary, data rich settings it is generally the case that data analysts examine some aspects of data before deciding on a formal statistical model or selecting the target parameters and inference to be performed. Inference is then adaptive, the same data being used *both* to define the statistical question of interest *and* to actually carry it out. Ignoring this adaptivity ('data snooping') typically results in loss of inferential guarantees and leads to flawed conclusions: it is, at least in part, responsible for the replicability crisis in science.

Our objective here is to review a statistical framework, the empirical Bayes approach, by which these two main challenges can be effectively met. The principal advocate of empirical Bayes methods is Professor Bradley Efron of Stanford University. Winner of the International Prize in Statistics in recognition of the 'bootstrap' approach to statistical inference, we assert his support for empirical Bayes methods of statistical inference is equally important to the contemporary statistical landscape. Neither fully frequentist or Bayesian, empirical Bayes methodology appears to provide a satisfactory compromise between these two main philosophies of inference. Sun[21] noted the main features: they exhibit Bayesian features such as risk-reducing shrinkage and selection adaptivity, while establishing appropriate control of frequentist properties of the inference. We consider here the empirical Bayes approach and illustrate these properties, while issuing the warning that it does not entirely solve the problems generated by data snooping.

2 Many Normal Means

The many normal means problem serves as template for analysis of contemporary large scale inference[5]. In the problem, we model data x as the outcome of a random variable $X = (X_1, \ldots, X_p)^{\mathrm{T}}, p \geq 3$, with a p-dimensional normal distribution with mean $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ and identity covariance matrix I_p, so that $X_1, \ldots, X_p$ are independent and

$$X_i \sim N(\theta_i, 1), \quad i = 1, \ldots, p. \tag{1}$$

Inference is required for the unknown θ assumed to have generated the data. In the frequentist perspective, θ is considered as fixed, but having some true, unknown value. In the Bayesian formulation, θ is itself considered as a random quantity, with some assumed prior distribution $\theta \sim g(\theta)$. Then, given the specified prior, Bayesian inference is extracted from the posterior

Bayesian inference: in Bayesian inference, (X, θ) are assumed both to be random, and inference about the value of θ which gave x is derived from the posterior distribution of θ, given $X = x$.

Frequentist inference: in frequentist inference, we treat data, x, say, as the observed value of a random variable X, with a distribution depending on a parameter θ, assumed to have some true, fixed (unknown) value. Inference on the value of θ is drawn by considering the sampling distribution of X, the hypothetical collection of datasets we might have seen.

[1] *Instituto de Ciencias Matemáticas, Consejo Superior de Investigaciones Científicas, Madrid, Spain.*
[2] *Department of Mathematics, Imperial College London, London, UK.*
**alastair.young@imperial.ac.uk*

Springer IISc PRESS

distribution $g(\theta|x)$, obtained through Bayes' Theorem[5, p.2]. We might, for instance, assume as prior a multivariate normal distribution in which $\theta_1, \ldots, \theta_p$ are independent, identically distributed $N(0, A)$, in which case the posterior distribution is multivariate normal with mean Bx and variance matrix BI_p, where $B = A/(A+1)$.

Bayes' Theorem: Bayes' Theorem is the rule for manipulation of conditional probabilities, $P(A|B) = P(B|A)P(A)/P(B)$. In Bayesian inference, its application tells us that the posterior distribution $_{g(\theta|X=x)}$ is proportional to the product of the prior density assumed on θ, $_{g(\theta)}$, and the likelihood function, the density of X evaluated at the observed data value x, $_{f(x|\theta)}$.

Multivariate normal distribution: a random vector has a multivariate normal distribution if any linear combination of its components has the univariate normal distribution, with the density of values around a central point being determined by the 'bell-shaped' Gaussian curve. The $N(\mu, \sigma^2)$ distribution has this density defined by $f(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \cdot \exp\{-\frac{1}{2\sigma^2}(x-\mu)^2\}$

3 Frequentist Analysis, Many Normal Means

We start our discussion with a frequentist analysis, described by Efron[3] as 'the most striking theorem of post-war mathematical statistics'. Given an estimator $\delta(X)$ of θ, define its risk function as

$$R(\theta, \delta(X)) = E_\theta \|\theta - \delta(X)\|^2,$$

where $\|\cdot\|$ is the Euclidean norm, $\|X\|^2 = X_1^2 + \cdots + X_p^2$, and E_θ means expectation with respect to repeated sampling of X from the model, for fixed parameter value θ. The James–Stein estimator is

$$\delta_{JS}(X) = \left(1 - \frac{p-2}{\|X\|^2}\right)X. \tag{2}$$

In the inference problem, the intuitively obvious estimator is $\delta(X) \equiv X$. This has constant risk: $R(\theta, X) \equiv p$, whatever the value of θ. It turns out though (see, for instance,[25, Chapter 3]) that the James–Stein estimator (2) has a risk which is strictly smaller: $R(\theta, \delta_{JS}(X)) < p$, whatever the value of θ. We speak of X as 'inadmissible' as an estimator of θ. This inadmissibility result was discussed by Stein[19] and James and Stein[16], though a simple proof was not provided until Stein[20]. Note the restriction $p \geq 3$ here: if $p = 1$ or $p = 2$ the estimator $\delta(X) = X$ is actually **admissible**. The James–Stein estimator incorporates shrinkage: the individual X_i are shrunk, towards 0 in this formulation, in providing estimators of $\theta_1, \ldots, \theta_p$, though the shrinkage factor $(1 - \frac{p-2}{\|X\|^2})$ involves all components of X, which are assumed independent.

Admissibility: An estimator $_{\delta(X)}$ is said to dominate an estimator $_{\delta^*(X)}$ if $_{R(\theta,\delta(X)) \leq R(\theta,\delta^*(X))}$ for all θ, and the inequality is strict for some θ. If an estimator can be dominated by some other estimator it is said to be *inadmissible*; otherwise it is *admissible*.

Practical applications in data analysis of the James–Stein and related estimators, are described, for example, by Efron and Morris[11,12], and, famously, in the general interest article Efron and Morris[13].

3.1 Empirical Bayes Interpretation

The counterintuitive use in the James–Stein estimator of what may be termed indirect evidence, of the $X_j, j \neq i$, in the estimation of the mean of X_i has[9, page 282] always aroused controversy, but may be rationalized in empirical Bayes terms[10].

In the Bayes formulation suggested above, in which $\theta_1, \ldots, \theta_p$ are independent, identically distributed $N(0, A)$ and under the assumed measure of loss $\|\theta - \delta(X)\|^2$, the appropriate estimator of θ is the mean of the posterior distribution, $\delta_B(X) = BX$ in terms of the random variable underlying the data sample. This minimises the Bayes risk $r(g, \delta(X)) = \int R(\theta, \delta(X))g(\theta)d\theta$, the risk function averaged over the assumed prior $g(\theta)$ on θ. If the prior variance A is specified, this Bayes estimator can be immediately applied, and its Bayes risk is readily calculated (Young and Smith, Chapter 3) as

$$r(g, \delta_B(X)) = pB.$$

Under the model assumed, the X_i are marginally independent, identically distributed as $N(0, A+1)$ and a simple calculation shows that the shrinkage factor has expectation under this marginal distribution

$$E\left\{1 - \frac{p-2}{\|X\|^2}\right\} = B.$$

So, in the case when A is unspecified in the model formulation, we may replace the unknown B in the expression for the Bayes estimator by the estimator $\hat{B} = 1 - \frac{p-2}{\|X\|^2}$, giving precisely the James–Stein estimator. We then note[25, Section 3.5] that

$$r(g, \delta_{JS}(X)) = r(g, \delta_B(X)) + \frac{2}{A+1},$$

so that the increase in Bayes risk due to using the James–Stein estimator rather than the Bayes estimator $\delta_B(X)$ tends to zero as the prior variance $A \to \infty$. So, the JS estimator has desirable risk properties. Frequentist risk is uniformly smaller than that of the obvious estimator, and the Bayes risk will often be close to that of the Bayes estimator, a desirable situation, as the Bayes estimator has important theoretical properties, such as admissibility: there is no estimator $\delta(X)$ with risk $R(\theta, \delta(X))$ uniformly smaller than $R(\theta, \delta_B(X))$: see, for instance Young and Smith[25, Chapter 3].

Under our model assumptions, that we have $X_i|\theta_i$, $i = 1, \ldots, p$, independently distributed as $N(\theta_i, 1)$ and $\theta_1, \ldots, \theta_p$ are independent, identically distributed $N(0, A)$, and so with $B = A/(A+1)$, we have, given data outcome $x = (x_1, \ldots, x_p)$ that

$$\theta_i \in x_i \pm 1.96, \tag{3}$$

is a 95% confidence interval (in a frequentist sense) for θ_i. Since the posterior distribution for $\theta_i|x_i$ is $N(Bx_i, B)$, we have that

$$\theta_i \in Bx_i \pm 1.96\sqrt{B} \tag{4}$$

is a 95% posterior credible set for θ_i. Estimating B by $\hat{B}$, gives an empirical Bayes posterior interval $\hat{B}x_i \pm 1.96\sqrt{\hat{B}}$. Efron[5, Section 1.5] noted a result, which he attributes to Carl Morris, that taking into account the variability of $\hat{B}$ as an estimate of B, leads to a refined empirical Bayes posterior interval

$$\theta_i \in \hat{B}x_i \pm 1.96\left[\hat{B} + \frac{2}{p-2}\left\{x_i(1-\hat{B})\right\}^2\right]^{1/2}. \tag{5}$$

The above calculations are presented assuming the prior $g(\theta)$ under which $\theta_1, \ldots, \theta_p$ are independent, identically distributed $N(0, A)$. Such assumptions can be generalised. Suppose still the model (1), but consider now the prior assumption that $\theta_1, \ldots, \theta_p$ are independent, identically distributed with common density $g(\theta)$. An elegant characterisation of the posterior distribution is given by 'Tweedie's formula'[6,9, Section 20.3].

Suppose that X is distributed as $N(\theta, 1)$ and θ has prior $g(\theta)$. The marginal density of X is

$$f(x) = \int_{-\infty}^{\infty} g(\theta)\phi(x-\theta)\mathrm{d}\theta,$$

in terms of the density $\phi(\cdot)$ of $N(0, 1)$. Tweedie's formula provides an expression for the posterior expectation of θ having observed x:

$$E(\theta|x) = x + l'(x),$$

where $l'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\log f(x)$. The key point here is that the posterior expectation $E(\theta|x)$ is expressed directly in terms of the marginal density $f(x)$, the context for empirical Bayes. We do not know the prior $g(\theta)$, but in large-scale situations we can construct an estimate $\hat{f}(x)$ of $f(x)$ from the data $x = (x_1, \ldots, x_p)$, the realised value of X, using techniques such as Poisson regression.

In general, empirical Bayes analysis is characterised by the estimation of prior parameter values from marginal distributions of data. With the prior parameter values fixed at these estimates, we proceed as in a regular Bayes analysis, as if the values had been specified, without consideration of the data, at the beginning.

3.2 Properties of Empirical Bayes and Their Relevance

Empirical Bayes methods are advocated for contemporary large-scale problems of statistical inference on the basis that: they provide a synthesis between frequentist and Bayesian approaches; they ensure some degree of protection against selection bias.

Stressed throughout contemporary discussions of empirical Bayes is the notion that such methods yield, for the context of large scale simultaneous inference, procedures with interpretable frequentist properties. The desirability of this is supported by Cox[1, Appendix B] who comments that 'from a general perspective one view of Bayesian procedures is that, formulated carefully, they may provide a convenient algorithm for producing procedures that may have very good frequentist properties'. We will demonstrate this in empirical illustrations below, examining the frequentist coverage properties of empirical Bayes intervals (5). Efron[8] provides ingenious methods by which the frequentist properties of Bayesian procedures may be estimated directly from given data.

It is often asserted (see, for instance,[9, Chapter 3]) that Bayesian inference is immune to selection bias. Taking the assertion as justified offers[9, Section 20.3] some hope that empirical Bayes estimators, such as the James–Stein estimator and those constructed via Tweedie's formula, provide a realistic protection against selection bias, and will provide some cure for data snooping. Convincing evidence is given by Efron[6]. However, we discuss below that some care must be taken in trusting this assertion. In essence, the immunity only holds if selection is assumed to operate both on θ and X, rather than only on X (for fixed θ generated from its prior $g(\theta)$): see Sect. 4.2.

3.3 Testing Versus Estimation

Focus of the above is on estimation of θ in the model (1). The empirical Bayes analysis that we have sketched utilises what may be termed[9, Section 15.5] an effect size model: $\theta_i \sim h(\theta)$ and, given θ_i, $X_i \sim N(\theta_i, 1)$, with the assumed prior $h(\theta)$ not having an atom at $\theta = 0$. A major focus of large-scale inference is the application of empirical Bayes ideas to provide an effective untangling of the interpretation of simultaneous test results: see, for instance, Efron and Hastie[9, Section 15.3].

Confidence interval: in frequentist inference a confidence set is a random set $S(X)$, which under the assumed sampling distribution for X contains the true fixed value of the parameter θ determining this distribution a specified proportion of the time.

Posterior credible set: a posterior credible set $S(x)$ is a set which contains a specified proportion of the probability mass of the posterior distribution $g(\theta|x)$, for the given data x.

A simple Bayesian framework for simultaneous testing[9, Section 15.3] is provided by a two-groups model: each of the p 'cases' $(x_1, \ldots, x_p)$ is either null, with prior probability π_0 or non-null, with probability $\pi_1 = 1 - \pi_0$. The observation x then has density either $f_0(x)$ or $f_1(x)$. If $\pi_0 = \Pr(\text{null})$, the density underlying observation x is $f_0(x)$ if null, while if $\pi_1 = \Pr(\text{non-null})$, the density is $f_1(x)$ if non-null. We may reasonably assume $f_0(x)$ to be the density of the standard normal distribution $N(0, 1)$, while the non-null density $f_1(x)$ is to be estimated.

Let $F_0(x)$ and $F_1(x)$ be the cumulative distribution functions corresponding to $f_0(x)$ and $f_1(x)$ and $S_0(x) = 1 - F_0(x), S_1(x) = 1 - F_1(x), S(x) = \pi_0 S_0(x) + \pi_1 S_1(x)$ Suppose an observation x_i is seen to exceed some threshold x_0, and define the Bayes false discovery-rate $F\mathrm{d}r(x_0)$ to be the probability that the observation x_i is null, given that it exceeds x_0. Then $F\mathrm{d}r(x_0) = \pi_0 S_0(x_0)/S(x_0)$. We suppose that $S_0(x_0)$ is known, and π_0 may reasonably in typical applications be assumed to be close to 1. While $S(x_0)$ is unknown, in large-scale testing situations it can be estimated by $\hat{S}(x_0) = N(x_0)/p$, where $N(x_0)$ is the number of observations in the data sample $(x_1, \ldots, x_p)$ with $x_i \geq x_0$. Then we immediately have an empirical Bayes estimate of the Bayes false discovery rate:

$$\widehat{F\mathrm{d}r}(x_0) = \pi_0 S_0(x_0)/\hat{S}(x_0).$$

Efron and Hastie[9, Chapter 15] discuss the relationship of this empirical Bayes posterior probability of nullness with frequentist procedures of simultaneous hypothesis testing based around control of the false discovery rate: see also Efron ([5], Chapter 4). Efron and Hastie[9, page 282] note how, in contrast with James–Stein estimation, such methods of simultaneous hypothesis testing arouse little conceptual controversy.

Having observed x_i equal to some value x_0, we would be more interested in the probability of nullness given $x_i = x_0$, rather than given $x_i \geq x_0$. We can therefore define the local false-discovery rate as

$$f\mathrm{d}r(x_0) = P\{\text{Null}|x_i = x_0\}.$$

We have that

$$f\mathrm{d}r(x_0) = \pi_0 f_0(x_0)/f(x_0),$$

so a local false-discovery estimate

$$\widehat{f\mathrm{d}r}(x_0) = \pi_0 f_0(x_0)/\hat{f}(x_0),$$

can be constructed using a curve $\hat{f}(x)$ which smooths a histogram of the values $\{x_1, \ldots x_p\}$. The R package `locfdr` implements construction of the local false-discovery estimate, which in a data analysis can be used as a selection mechanism to identify parameters for formal inference. The null proportion π_0 can be estimated, or approximated to be 1. Similarly, the theoretical standard normal null density $f_0(x)$ can, in practice, be estimated: see Efron and Hastie[9, Section 15.5] for a summary and discussion. In a data analysis we might define an observation as being 'interesting' if, say, $\widehat{f\mathrm{d}r}(x_i) \leq 0.2$, and flag such for follow-up investigation, or as cases where we wish to do a formal inference. This, of course, has to be done in a way that accounts for the selection condition $\widehat{f\mathrm{d}r}(x_i) \leq 0.2$.

4 Selective Inference

Classical statistical methods are designed to give error guarantees in situations where the objectives of the inference are specified before collecting the data. In contemporary problems, though, such idealised settings are the exception rather than the norm. More realistically, an exploratory analysis of the data is performed before selecting the relevant inferential questions to examine, often, as in a regression setting, in the form of a selected model. Failing to acknowledge this adaptivity in the subsequent inferences can yield the reported error assessments invalid: for instance, frequentist Type 1 error guarantees of testing procedures are lost. This problem of selection bias has received considerable attention in recent years, particularly from a frequentist perspective. Efron[7] describes methods for error assessment in inference on parameters which account for model selection effects. Though[9, Chapter 20] there is no overarching general theory for inference after data snooping, prominent among approaches to remedy of the effects of selection bias is the conditional approach, which says that inference after selection should be based on hypothetical data samples which would lead to the same inference problem being tackled. This provides a broad framework for inference, which encapsulates the large-scale inference problem, expressed by (1), which is our focus here.

Suppose our data x represents the realisation of a random variable $X \in \mathcal{X}$, whose sampling distribution we model by some parametric family $\{F(x; \theta) : \theta \in \Theta\}$, with $F(x; \theta)$ denoting the distribution function of X under θ. In the example that is our focus here $F(x; \theta)$ denotes

the multivariate Gaussian distribution $N_p(\theta, I_p)$. The density function of X we denote by $f(x;\theta)$. We assume that there is a set of m potential parameters of interest, $\{\psi_1(\theta),\ldots,\psi_m(\theta)\}$, from which at most one is to be selected for inference after observing the data. This selection may, as we will discuss, be made according to a randomised procedure. In the many normal means problem, the set of potential parameters of interest would contain all subsets of the p means $\{\theta_1,\ldots,\theta_p\}$. We assume that we know the forms of functions $p_i : \mathcal{X} \to [0,1], i = 1,\ldots,m$, such that, having observed $X = x$, $\psi_i(\theta)$ is selected for inference with probability $p_i(x)$. In our illustrations later, we will specify selection to entail choice of a one-dimensional parameter for inference, specifically the mean θ_I corresponding to the largest element of X, $X_I = \max\{X_1,\ldots,X_p\}$. We therefore simplify notation by writing the selected parameter simply as ψ, and the corresponding selection probability as $p(y)$.

The conditional approach to frequentist inference[14] advocates that inference for the selected parameter ψ should be based on the conditional distribution of the data given selection. This distribution has density

$$f_S(x;\theta) = \frac{f(x;\theta)p(x)}{\varphi(\theta)}, \quad \varphi(\theta) = E_\theta\{p(X)\}, \tag{6}$$

so that the normalising constant $\varphi(\theta)$ is the probability that ψ gets selected when θ is the true parameter. In general, inference based on this selective density $f_S(x;\theta)$ may be awkward: $\varphi(\theta)$ may be intractable, and inference on ψ may be complicated by the presence of nuisance parameters. In the normal means example, inference on θ_I depends on the unknown nuisance parameters $\theta_j : j \neq I$. Simpler forms of inference, which achieve the same protection against selection bias, are desirable: an attractive idea is discussed below.

We can interpret the conditional approach as a form of information splitting. For a given ψ, let R be the Bernoulli random variable which takes the value 1 if ψ gets selected for inference, and 0 otherwise, so that $R|X \sim \text{Bernoulli}\{p(X)\}$. Following Fithian et al.[14], the data generating process of X may be thought of as consisting of two stages. In the first, the value, r say, of R is sampled from its marginal distribution, and in the second stage X is sampled from the conditional distribution $X|r$. Since it is R which determines whether inference is provided for ψ or not, inference based on information revealed at the second stage in necessarily free of any selection bias, since it eliminates the information about the parameter provided by R. So, the information provided by the data is divided into two portions, one of which is used for selection (R) and the other is used for the actual inference ($X|R$).

There is a trade-off between the power of the selection mechanism (the ability to identify a parameter when it is truly interesting, such as a significant effect), and the power of the subsequent inferential method. If powerful inference is required and obtaining new data after selection is infeasible, we need to utilise the available information efficiently. The amount of information used for selection can be limited by applying the selection mechanism to a randomised version of the original data: see Tian and Taylor[22], Garcia Rasines and Young[15]. Formally, we can generate a random variable W, with known distribution and independent of the data, and apply the selection mechanism to $U = u(X, W)$, where u is some function of the data and the noise: a convenient case for the context of the model (1) is $U = X + W$. Note that, if $p_U(u)$ denotes the selection function in terms of U, the selection function of the data X would be computed as $p(x) = E\{p_U(x, W)\}$. For the model (1) which is our focus, we have $X \sim N_p(\theta, I_p)$. Suppose a parameter of interest ψ is selected for inference if and only if $U = Y + W \in E$, where $W \sim N_p(0, \gamma I_p)$ is a noise vector independent of X and $E \subseteq \mathbf{R}^p$ is some selection event, defining when the quantity ψ is chosen for inference. Then we note that $U = X + W$ and $V = X - \frac{1}{\gamma}W$ are independent, from properties of the normal distribution. Now, selection is defined only in terms of U, so a simple inference can be based on V, which is unaffected by the selection of ψ as our focus of inferential interest. Such inference is trivial, as V is distributed as $N_p(\theta, \{1+\gamma^{-1}\}I_p)$. Note that U is distributed as $N_p(\theta, \{1+\gamma\}I_p)$, so the noise parameter γ has the role of balancing how much information about θ we have in the selection and inferential stages. Garcia Rasines and Young[15] consider methods of inference based on V in regression models.

4.1 A Simple Univariate Model

To illustrate some of these ideas, consider the simple univariate ($p = 1$) normal model in which $X \sim N(\theta, 1)$, but suppose a selection, or truncation, condition $X > 0$ is imposed: any data provided for analysis satisfies $x > 0$. Under the condition on selection paradigm, inference on θ is based on the conditional distribution of $X|X > 0$,

with selective density $f_S(x;\theta) = \phi(x-\theta)/\Phi(\theta)$, in terms of the distribution function $\Phi(\cdot)$ of the standard normal distribution. Let $F(x;\theta)$ be the corresponding distribution function. For given observed data outcome x_o we can construct the appropriate selective confidence interval, of exact coverage $1-\alpha$ under repeated sampling of X subject to the selection event $X > 0$ as $\{\theta : \alpha/2 \le F(x_o;\theta) \le 1-\alpha/2\}$. Unfortunately, such inference is inappropriate. Consider the case $\theta \ll 0$: in such a situation the selection probability $P(X > 0)$ is vanishingly small, and the data outcome $X = x_o$ contains little information about the value of θ. Indeed, Kivaranovic and Leeb[17] show that such confidence intervals have infinite expected length under repeated sampling.

Suppose, instead, we apply the randomisation idea, and provide inference on θ if and only if $U = X + W > 0$, where W is random noise, independent of X, with distribution $N(0,\gamma)$. Then in the definition of the selective density (6), $p(x) = P(x + W > 0) = \Phi(x/\sqrt{\gamma})$ and $\varphi(\theta) = \Phi(\theta/\sqrt{1+\gamma})$. Now, with this randomisation, the confidence interval constructed from the selective density

$$f_S(x;\theta) = \phi(x-\theta)\Phi(x/\sqrt{\gamma})/\Phi(\theta/\sqrt{1+\gamma}) \tag{7}$$

is known to have finite expected length[18]. In fact, the length of the confidence interval is bounded above by the length of the confidence interval based on the $N(\theta, 1+\gamma^{-1})$ distribution of $V = X - \frac{1}{\gamma}W$. There is loss, in terms of the size of the confidence set, in providing inference here using V alone, rather than from the conditional distribution of $X|U > 0$. In general, however, the full conditional model $X|\{U \in E\}$ may be complicated or intractable. The cost incurred in using V alone for inference will depend on how informative the distribution of $U|\{U \in E\}$ is about the parameter of interest.

Figure 1 considers the length of confidence sets of coverage 90% constructed from the selective density (7), as a function of the true mean θ, in comparison with the length of the confidence set constructed from the normal distribution of V, which also has repeated sampling coverage 90%, and the length of the 'face value' interval constructed from the $N(\theta,1)$ distribution of X, ignoring selection. The latter does *not* have repeated sampling coverage close to the nominal 90%, unless $\theta \gg 0$. In this simple univariate normal model, the cost in terms of the length of the confidence set might be judged as very slight if the true value of θ is less than, say, about -1.

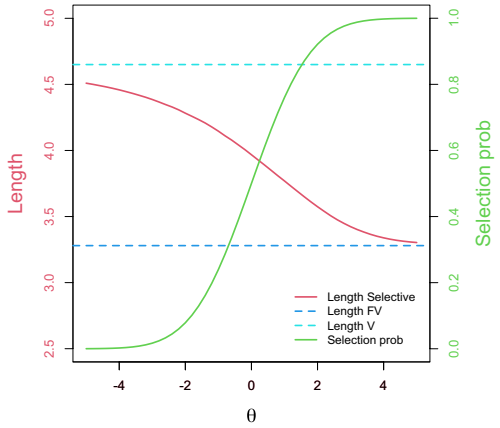


Figure 1: Univariate example. Lengths of confidence sets of coverage 90% constructed from selective density (7), compared to lengths of intervals constructed from the normal distribution of V and the lengths of the 'face value' interval (FV) constructed from the $N(\theta,1)$ distribution of X, ignoring selection, as a function of θ. Selection probability as a function of θ shown in green.

4.2 Selection Bias and Bayesian Inference

Why is selection bias a problem? Frequentist methods evaluate the accuracy of inferential procedures with respect to the sampling distribution of X at a fixed value of the parameter. Since selection modifies the sampling distribution, by favouring data values with higher selection probability, it is clear that inferential correctness requires that the reported accuracy be appropriately modified by accounting for the selection, through use of $f_S(x;\theta)$ as the basis for inference. The Bayesian viewpoint, as we have seen, is, instead, that once the data has been observed, the recognition that a different data realisation could have resulted in a different inferential problem being posed, or none at all, should have no effect on the inference[2]. This position has been challenged (see, for instance,[24]). Our central thesis here is that we must reassess the view that Bayesian and empirical Bayes methods necessarily provide the protection from selection effects that has been crucial to valid inference in large-scale problems.

According to Yekutieli[24], the correct Bayesian inference for a selected parameter depends on how the selection mechanism acts on the parameter space. Consider the joint sampling distribution of (θ, X), and a selection function $p(x)$. We say that θ is *random* if the joint sampling scheme for the parameter and data is such that the pairs (θ, X) are sampled from their joint distribution

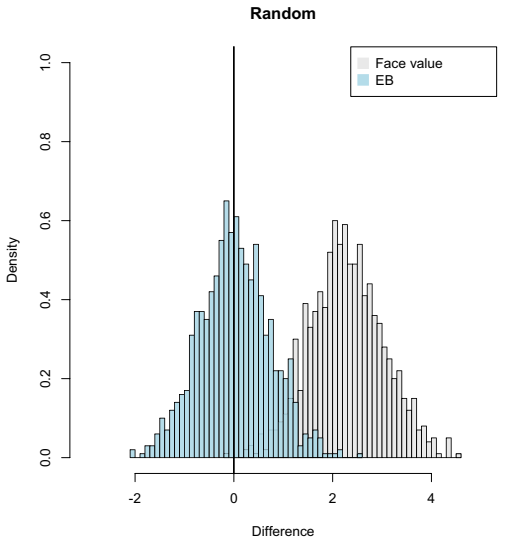


Figure 2: Distribution over 1000 replications of $X_I - \theta_I$ (Face value) and $\hat{\theta}_I - \theta_I$ (EB), where $X_I = \max\{X_1, \ldots, X_{1000}\}$ and $\hat{\theta}_I = \hat{B}X_I$ is the empirical Bayes estimator.

until $\psi \equiv \psi(\theta)$ gets selected and say that θ is *fixed* if θ is sampled from its marginal distribution, held fixed, and X sampled from its conditional distribution $X|\theta$ until ψ is selected for inference. Woody et al.[23] refer to these two scenarios as 'joint selection' and 'conditional selection', respectively.

As above, let R be the binary random variable that indicates if selection of the parameter ψ under consideration has happened. If θ is random, its density given selection and a prior density $\pi(\theta)$ is $\pi(\theta|R=1) \propto \pi(\theta)P_\theta(R=1) = \pi(\theta)\varphi(\theta)$. On the other hand, if θ is fixed, its conditional density is unchanged, $\pi(\theta|R=1) = \pi(\theta)$. The conditional density of the data x given θ and selection is $f_S(x;\theta) = f(x;\theta)p(x)/\varphi(\theta)$ in both cases. Then, the posterior distribution for a random parameter is

$$\pi(\theta|x) \propto \frac{\pi(\theta)\varphi(\theta)f(x;\theta)}{\varphi(\theta)} = \pi(\theta)f(x;\theta), \tag{8}$$

the usual Bayesian posterior, constructed without consideration of the selection. Hence Bayesian inference about $\psi = \psi(\theta)$, which is extracted from $\pi(\theta|x)$, is unaffected by selection in this case. In the case of a fixed parameter, the posterior is given by

$$\pi(\theta|x) \propto \frac{\pi(\theta)f(x;\theta)}{\varphi(\theta)}. \tag{9}$$

So, for a fixed parameter the posterior needs to be adjusted, and would formally be obtained by attaching the prior density $\pi(\theta)$ to the selective likelihood, $f_S(x;\theta)$. The viewpoint that Bayesian inference does not require an adjustment for selection, and protection against selection bias might be expected to be afforded by the empirical Bayesian approaches to inference sketched above, follows from the implicit assumption that the parameter is random. While posterior densities (8) and (9) are formally correct given the respective sampling mechanisms, it might be argued that it is not clear that a parameter can be labelled as random or fixed without explicit consideration of the sampling mechanism. In the context of the normal means problem, it may be reasonable to consider the parameter θ as random, but the sampling process might not be well-defined, and caution is appropriate in any assumption that Bayesian inference (or the empirical Bayes inference we have described) does provide protection against selection bias. In the context, say, of a genetic study where the quantity X_i is a measurement relating to gene i, with θ_i being some 'true effect' due to that gene, it might be reasonable to consider θ_i as an intrinsic quantity associated with that gene i.e. to consider, in terms of our discussion, θ_i as a fixed parameter.

5 Numerical Illustrations

5.1 Random and Fixed Parameter Models

We describe here a variant of the analysis carried out by Efron and Hastie[9, Section 20.3] to examine the idea that empirical Bayes estimates are a realistic approach to the problem of selection bias introduced by data snooping.

We consider the many normal means model (1), with $p = 1000$. We specify the distribution of $\theta = (\theta_1, \ldots, \theta_p)$ to be such that the components are independent $N(0, 1)$, so that the posterior distribution of $\theta_i|x_i$ is $N(Bx_i, B)$, with $B = 1/2$, and the Bayes estimator is $E(\theta_i|x_i) = Bx_i$. The empirical Bayes estimator, that is the James–Stein estimator, is $\hat{\theta}_i = \hat{B}x_i$, where $\hat{B} = 1 - (p-2)/\sum_{i=1}^{p} x_i^2$. We generate 50,000 replications from the specified joint distribution of (θ, X). For each we determine the index I corresponding to the largest observed data point, $I = \text{argmax}\{X_i\}$, and construct the face value interval (3), the Bayes interval (4) and empirical Bayes interval (5) for θ_I.

Figure 2 shows a histogram for the first 1000 replications of $X_I - \theta_I$, together with the corresponding histogram for $\hat{\theta}_I - \theta_I$. Selection bias is obvious: the fact that we have chosen to examine

the parameter value θ_I corresponding to the largest observation means that the uncorrected, face value differences are not centred on zero, but shifted to the right. By contrast, the empirical Bayes differences $\hat{\theta}_I - \theta_I$ are centred at zero. The coverages of the true θ_I over the 50,000 replications of the face value, Bayes and empirical Bayes intervals (3), (4) and (5) were 0.330, 0.950 and 0.949 respectively. The empirical Bayes interval delivers the desired frequentist property, of containing the parameter of interest θ_I in very close to 95% of the replications. Selection of the parameter of interest as θ_I from the data means that the face value interval has frequentist coverage very far from the nominal desired 95%. Note that the face value interval (3) has, for this context, constant width 3.92, while the Bayes interval (4) has constant width 2.77, and over the 50,000 replications the empirical Bayes interval had average width 2.80. The Bayes and empirical Bayes estimators of θ_I were virtually unbiased over the replications, while the face value estimator X_I displays substantial positive bias in this situation, demonstrating the need to correct the inference for selection.

As we have argued, to mitigate against the selection bias, we can utilise the idea of randomisation. For each specified noise level, we define the parameter of interest from a particular dataset $X = (X_1, \ldots, X_p)$, with the X_i independent, $X_i \sim N(\theta_i, 1)$, and independent noise $\{W_1, \ldots, W_p\}$, with the W_i independent, identically distributed $N(0, \gamma)$.

The above analysis reflects what we described in Sect. 4.2 as a random parameter context: on each of the replications (θ, X) was simulated from the specified joint distribution. Instead, we consider now repeating the simulation for a fixed parameter context. Here, a fixed value $\theta = (\theta_1, \ldots, \theta_p)$, again with $p = 1000$, was generated from the assumed prior, in which the elements of θ are independent $N(0, 1)$. Figure 3 shows a histogram of the 1000 values $\theta_1, \ldots, \theta_{1000}$, as a probability distribution, with the $N(0, 1)$ density from which they were generated superimposed.

Two different analyses are then carried out. In the first simulation, for each of 20,000 replications we make inference for θ_I, $I = \text{argmax}\{X_i + W_i\}$. Therefore, as before, we are considering inference for a different target parameter on each replication. Note that we are not therefore considering coverages of confidence sets in any conventional frequentist sense, as the parameter for which inference is made is not held fixed over the replications. As before, we consider

Table 1: Coverages of face value, Bayes, empirical Bayes (EB) intervals, together with intervals based on randomisation.

γ	Face value (3)	Bayes (4)	EB (5)	Randomisation
0.0	0.308	0.976	0.976	–
0.25	0.326	0.976	0.974	0.947
0.5	0.388	0.975	0.973	0.951
1.0	0.539	0.967	0.966	0.952

The latter are based on V_I and require $\gamma > 0$. Fixed parameter sampling model, all figures based on 20,000 replications. In each replication interval constructed for θ_I, where $X_I + W_I = \max\{X_1 + W_1, \ldots, X_{1000} + W_{1000}\}$, W_i distributed as $N(0, \gamma)$

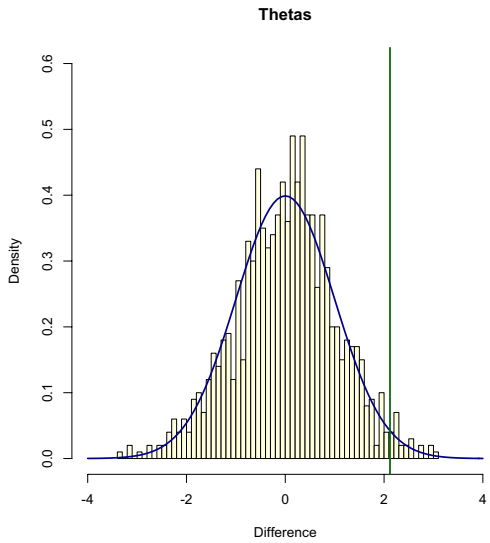


Figure 3: Histogram of the 1000 values $\{\theta_1, \ldots, \theta_{1000}\}$, as a probability distribution, with the $N(0, 1)$ density from which they were generated superimposed. Repeated sampling coverages assessed conditional on selected parameter for inference being value indicated by vertical line.

the repeated sampling coverages of the face value, Bayes and empirical Bayes intervals (3), (4) and (5), which we recall are all of nominal 95% coverage. Now also included in the analysis are coverages of confidence intervals for θ_I obtained from the normal distribution of $V_I = X_I - \frac{1}{\gamma} W_I$,

which we argued is unaffected by selection. Table 1 shows that, indeed, intervals based on this latter quantity yield the nominal desired repeated sampling properties in this fixed parameter sampling model, even under selection. The empirical Bayes intervals do not fully mitigate against data

Table 2: Coverages of face value, Bayes, empirical Bayes (EB) intervals, together with intervals based on randomisation.

γ	θ_I	Face value (3)	Bayes (4)	EB (5)	Randomisation
0.0	2.120	0.117	0.999	1.000	–
0.25	2.120	0.191	1.000	1.000	0.950
0.5	2.266	0.416	1.000	1.000	0.951
1.0	2.719	0.710	0.982	0.979	0.950

The latter are based on V_I and require $\gamma > 0$. Fixed parameter sampling model, all figures based on 20,000 replications, conditional on selected parameter being specified θ_I

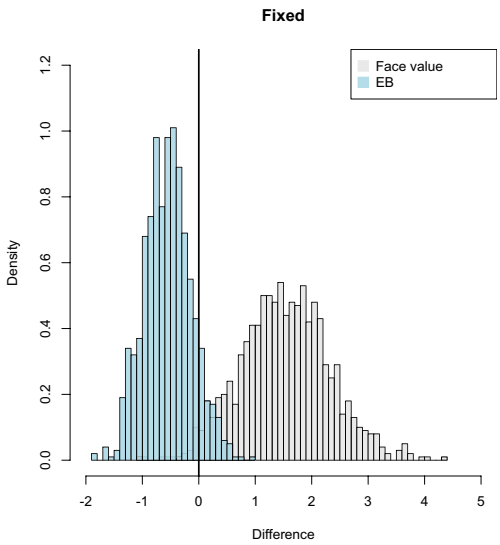


Figure 4: Distribution over 1000 replications of $X_I - \theta_I$ (Face value) and $\hat{\theta}_I - \theta_I$ (EB), where $X_I + W_I = \max\{X_1 + W_1, \ldots, X_{1000} + W_{1000}\}$, under fixed parameter assumption and randomised response, W_i distributed as $N(0, \gamma)$, the case $\gamma = 1$, $\hat{\theta}_I = \hat{B}X_I$ is the empirical Bayes estimator.

snooping in this sampling model, while inference based on V_I does. Table 1 shows coverages of the Bayes and empirical Bayes intervals to be some way off the desired 95%. A histogram of differences $\hat{\theta}_I - \theta_I$ for the 'no noise case', $\gamma = 0$, is still centred on zero, but is not symmetric, with positive skewness.

In a further simulation, a particular dataset $X = (X_1, \ldots, X_p)$, with the X_i independent, $X_i \sim N(\theta_i, 1)$ was generated, and $I = \text{argmax}\{X_I\} \equiv 269$ identified. It is worth noting that the selected parameter of interest is not the largest θ_i: in fact 22 values exceed θ_{269}, as shown by the vertical line in Fig. 3. Then we reconsider the repeated sampling coverages of the face value, Bayes and empirical Bayes intervals (3), (4) and (5), but now conditional on the fixed parameter value θ_{269} (actually equal to 2.120) being chosen as the parameter of interest on each replication. So, each of the 20,000 replications in this case had $X_{269} = \max\{X_j\}$. On each of the replications, the parameter of interest is the *same*, so in this analysis we are actually examining the coverages of the confidence sets in a strict frequentist sense. The empirical Bayes method does not protect against selection bias in this fixed parameter context. The whole analysis was repeated based on randomised data $(X_1 + W_1, \ldots, X_{1000} + W_{1000})$, for different noise levels γ. When $\gamma = 1.0$, for instance, the target parameter turned out to be defined as $\theta_I \equiv \theta_{115} = 2.719$. The coverages of the Bayes and empirical Bayes intervals, shown in Table 2 are now very far from the nominal desired 95%. Inference based on V_I does ensure strict frequentist control of confidence set coverage.

Figure 4 provides the analogue of Fig. 2 for this fixed parameter model, when the analysis is based on the randomised data $(X_1 + W_1, \ldots, X_{1000} + W_{1000})$, for the case $\gamma = 1$. The same bias due to selection of the interest parameter of a 'face value' inference as seen in Fig. 2 is evident. By contrast with Fig. 2, in this fixed parameter model, the differences $\hat{\theta}_I - \theta_I$ are no longer centered around zero. Figure 5 shows that the distribution over the replications of $V_I - \theta_I$ is centered at zero. The corresponding figures for the face value and empirical Bayes estimators X_I and $\hat{\theta}_I$ are very similar for other noise levels γ, including the case $\gamma = 0$, when no randomisation is employed.

5.2 A Two-Groups Model

We consider now the two-groups model considered in Sect. 3.3. In this situation, as discussed, it is reasonable to suppose that the proportion π_0 of the elements of θ that are null, $\theta_i = 0$, is large. We take, as before, $p = 1000$, and set $\theta_1 = \cdots = \theta_{900} = 0$, with the remaining components of θ as a set of independent

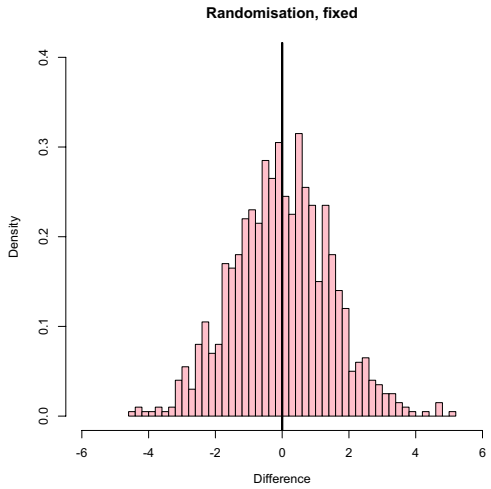


Figure 5: Distribution over 1000 replications of $V_I - \theta_I$, with $V_I = X_I - W_I/\gamma$, where $X_I + W_I = \max\{X_1 + W_1, \ldots, X_{1000} + W_{1000}\}$, under fixed parameter assumption and randomised response, W_i distributed as $N(0, \gamma)$, the case $\gamma = 1$.

realisations of $N(0, 1)$, held fixed over a series of 20,000 replications of the model (1). We report here results for the situation where inference is made on θ_I, this chosen parameter of interest being selected on the basis of randomised data: $X_I + W_I = \max\{X_1 + W_1, \ldots, X_{1000} + W_{1000}\}$, with the noise variables $W_1, \ldots, W_{1000}$ independent $N(0, \gamma)$. So, again, the target parameter varies over the replications. In this context, since we might primarily be interested in identifying true, non-zero, effects, rather than in just examining the overall coverage properties of the empirical Bayes interval and the interval based on V_I, we examine: $P(\theta_I \in \text{Interval} \mid \theta_I = 0)$ and $P(\theta_I \notin \text{Interval} \mid \theta_I \neq 0)$. If an interval contains zero, we might conclude that there is no evidence to suggest that the corresponding effect is non-null, while if the interval does not include zero, we might infer evidence of a non-null effect. Results are given in Table 3. The empirical Bayes intervals for θ_I contain zero too high a proportion of times, while intervals based on V_I correctly contain θ_I, when the true value of this selected parameter is $\theta_I = 0$, on the specified proportion 95% of replications. The inference based on the randomisation quantity V_I is more powerful, in the sense that the intervals for non-zero selected θ_I do not include zero in a higher proportion of replications. Of course, as the noise level γ increases, the proportion of replications for which the selected parameter of interest is actually null increases.

5.3 Data Analysis

Efron and Hastie[9, Section 13.3 and elsewhere] discuss analysis of data from a prostate cancer study. The data consists of a set of $p = 6033$ observations $(X_1, \ldots, X_p)$, each measuring the effect of one gene. Efron and Hastie[9, Section 15.1] describe how these observations are extracted from raw gene expression data comparing a set of prostate cancer patients and a set of control patients. The objective is to identify non-null genes, for which the patients and the controls respond differently: a reasonable model for both null and non-null genes is the normal means model (1). Suppose we use $\widehat{fdr}(x_i) < 0.2$ as a selection rule, based on the data on all $p = 6033$ genes, giggled by injection of small levels of random $N(0, \gamma)$ noise, with $\gamma = 0.25$. This identifies 15 'interesting' cases. The plot produced by `locfdr` with default settings is shown as Fig. 6. Note that the estimated null distribution, by both maximum likelihood and the central matching estimate method[4], are normal distributions with standard deviation close to $\sqrt{(1 + 0.25)}$, which we expect, as `locfdr` is applied to $\{X_1 + W_1, \ldots, X_p + W_p\}$, with the X_i assumed to have variance 1 and the independent noise variables W_i specified to have variance 0.25.

Of the $p = 6033$ cases, 478 of the face value intervals (3), 130 of the Bayes intervals (4) and 9 of the empirical Bayes intervals (5), for genes 332, 364, 579, 610, 914, 1068, 1720, 3940, 4546, all among the set of 15 interesting cases selected by `locfdr`, do not contain zero. Over the full set of

Table 3: Two-groups model.

		$P(\theta_I \in \text{Interval} \mid \theta_I = 0)$		$P(\theta_I \notin \text{Interval} \mid \theta_I \neq 0)$	
γ	$P(\theta_I = 0)$	EB (5)	Randomisation	EB (5)	Randomisation
0.1	0.38	0.995	0.949	0.004	0.057
0.25	0.41	0.995	0.948	0.004	0.073
0.5	0.48	0.994	0.950	0.005	0.134
1.0	0.66	0.992	0.951	0.006	0.230

Coverage of true null values, correct identification of non-null values of θ_I, where $X_I + W_I = \max\{X_1 + W_1, \ldots, X_{1000} + W_{1000}\}$, under fixed parameter assumption and randomised response, W_i distributed as $N(0, \gamma)$. Empirical Bayes (EB) intervals, randomised intervals based on V_I

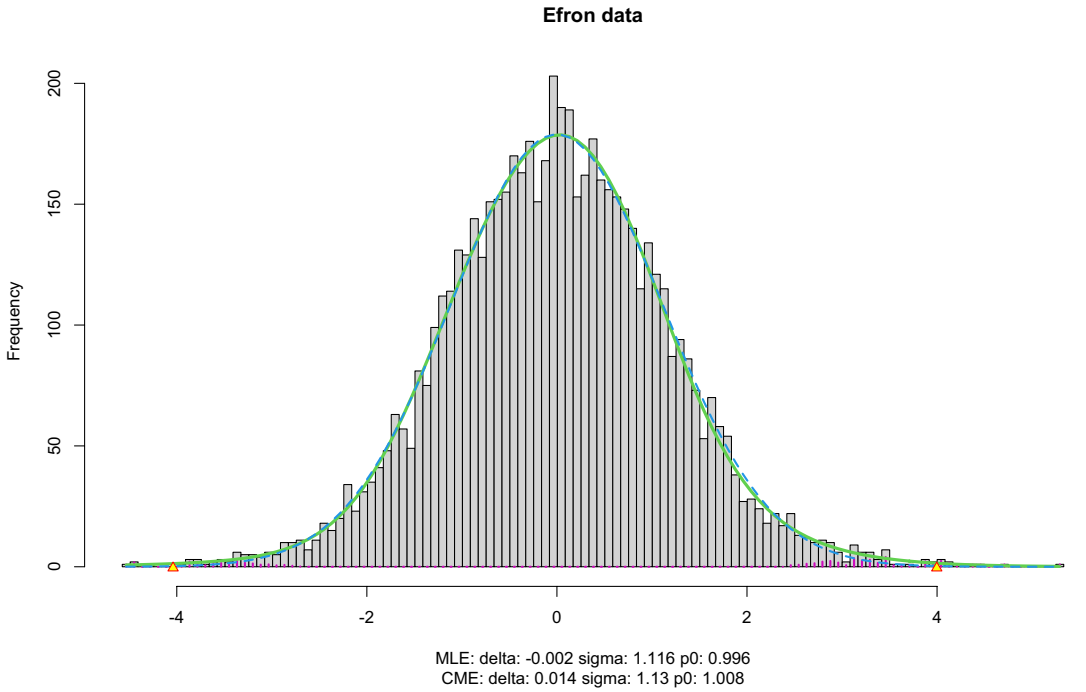


Figure 6: Output of running R package `locfdr` on the prostate gene data.

$p = 6033$ genes, 350 of the intervals based on $X_i - \frac{1}{\gamma} W_i$ do not contain zero. Of the 15 cases selected by `locfdr`, intervals based on $X_i - \frac{1}{\gamma} W_i$ do not contain zero only for *one* case, gene with label 914, which might temper any willingness to read too much into the fact that 9 of the empirical Bayes intervals suggest non-null effects among the 15 cases selected from the full set of 6033 genes for detailed inspection.

6 Discussion

Many contemporary problems of large-scale inference may, perhaps after transformation and data scaling, be expressed in terms of the many normal means model (1), with interest typically being in some subset of $\theta = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ chosen after examination of the data, such as the element of θ corresponding to the largest observed data point. The empirical Bayes approach to inference in this model provides a framework with attractive properties. If estimation is required for the whole parameter vector θ, empirical Bayes estimators incorporate shrinkage, through the indirect evidence provided by all of the elements of $X = (X_1, \ldots, X_p)$ in estimation of all of the individual components of θ: the result is desirable frequentist and Bayes risk properties. The empirical Bayes inference can be seen to be adaptive to the data-driven specification of the parameter chosen for inference, maintaining appropriate control of frequentist properties of the inference, at least under a random parameter or joint selection assumption. Under a random parameter assumption, for instance, an empirical Bayes 95% confidence set will contain the target parameter of interest on close to 95% of instances. This is not necessarily true under a fixed parameter or conditional selection model. Such frequentist properties do not, of course, relate formally to those demanded by the condition on selection paradigm of selective inference, which requires a 95% confidence set to contain a specified target parameter for 95% of instances for which *that* fixed target parameter is chosen by the selection mechanism. Some care is required on attributing to empirical Bayes methods such strong frequentist control. If this is demanded, methods based on selection of the target parameters from randomised versions of sample data offer a simple alternative.

A further key context where there is data-driven choice of the parameters selected for formal inference concerns high-dimensional regression, where inference is carried out after model selection using the same data. Formal examination of the ability of empirical Bayes methods to account for selection bias in that

context is unexplored, but would add to the conclusions reached here for many normal means problem.

Publisher's Note

Declarations

Conflict of interest

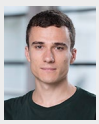
On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access

Received: 14 October 2021 Accepted: 3 January 2022
Published online: 7 March 2022

References

1. Cox DR (2006) Principles of statistical inference. Cambridge University Press, Cambridge
2. Dawid AP (1994) Selection paradoxes of Bayesian inference. In: Anderson TW, Fang KT, Olkin I (eds) Multivariate analysis and its applications, vol 24. Institute of Mathematical Statistics lecture notes, monograph series, pp 211–220
3. Efron B (1992) Introduction to James and Stein (1961) 'Estimation with quadratic loss'. In: Kotz S, Johnson NL (eds) Breakthroughs in statistics, vol 1. Springer, New York, pp 437–442
4. Efron B (2007) Size, power and false discovery rates. Ann Stat 35:1351–1377
5. Efron B (2010) Large-scale inference: empirical Bayes methods for estimation, testing and prediction. Cambridge University Press, Cambridge
6. Efron B (2011) Tweedie's formula and selection bias. J Am Stat Assoc 106:1602–1614
7. Efron B (2014) Estimation and accuracy after model selection. J Am Stat Assoc 109:991–1007
8. Efron B (2015) Frequentist accuracy of Bayesian estimates. J R Stat Soc Ser B 77:617–646
9. Efron B, Hastie T (2016) Computer age statistical inference: algorithms, evidence and data science. Cambridge University Press, Cambridge
10. Efron B, Morris C (1973) Stein's estimation rule and its competitors–an empirical Bayes approach. J Am Stat Assoc 68:117–130
11. Efron B, Morris C (1973) Combining possibly related estimation problems (with discussion). J R Stat Soc Ser B 35:379–421
12. Efron B, Morris C (1975) Data analysis using Stein's estimator and its generalizations. J Am Stat Assoc 70:311–319
13. Efron B, Morris C (1977) Stein's paradox in statistics. Sci Am 236:119–127
14. Fithian W, Sun DL, Taylor JE (2017) Optimal inference after model selection. arXiv:1410.2597v4
15. Garcia Rasines D, Young GA (2021) Splitting strategies for post-selection inference. arXiv:2102.02159
16. James W, Stein C (1961) Estimation with quadratic loss. In: Proceedings of 4th Berkeley symposium on mathematical statistics and probability, vol 1. University of California Press, Berkeley, pp 361–379
17. Kivaranovic D, Leeb H (2021) On the length of post-model-selection confidence intervals conditional on polyhedral constraints. J Am Stat Assoc 116:845–857
18. Kivaranovic D, Leeb H (2021b) A (tight) upper bound on the length of confidence intervals with conditional coverage. arXiv:2007.12448v2
19. Stein C (1956) Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proceedings of 3rd Berkeley symposium on mathematical statistics and probability, vol 1. University of California Press, Berkeley, pp 197–206
20. Stein C (1981) Estimation of the mean of a multivariate normal distribution. Ann Stat 9:1135–1151
21. Sun L (2020) Topics on Empirical Bayes normal means. PhD thesis, University of Chicago
22. Tian X, Taylor JE (2018) Selective inference with a randomized response. Ann Stat 46:679–710
23. Woody S, Padilla OHM, Scott JG (2021) Optimal post-selection inference for sparse signals: a nonparametric empirical Bayes approach. Biometrika. https://doi.org/10.1093/biomet/asab014
24. Yekutieli D (2012) Adjusted Bayesian inference for selected parameters. J R Stat Soc Ser B 74:515–541
25. Young GA, Smith RL (2005) Essentials of statistical inference. Cambridge University Press, Cambridge



Daniel García Rasines holds a postdoctoral research position at ICMAT-CSIC. He completed BSc in Mathematics at the University of Santiago de Compostela and obtained MSc and PhD in Statistics from Imperial College London. After his doctoral studies, he held a short predoctoral research position at ICMAT-CSIC before joining as a postdoctoral researcher.

G. Alastair Young is Professor of Statistics in the Department of Mathematics at Imperial College London. Having studied at the Universities of Edinburgh and Cambridge, he received his PhD from Cambridge in 1987. He then held faculty positions in the Statistical Laboratory at the University of Cambridge, before joining Imperial College London in 2005. Alastair is a Fellow of the Institute of Mathematical Statistics, and former Joint Editor of the Journal of the Royal Statistical Society, Series B.