Japanese Journal of Statistics and Data Science
https://doi.org/10.1007/s42081-023-00209-y

ORIGINAL PAPER

Stein Estimation and Statistical Shrinkage Methods

Machine learning and the James–Stein estimator

Bradley Efron[1,2]

Received: 24 March 2023 / Accepted: 13 May 2023

Abstract
It is now 62 years since the publication of James and Stein's seminal article on the estimation of a multivariate normal mean vector. The paper made a spectacular first impression on the statistical community through its demonstration of inadmissability of the maximum likelihood estimator. It continues to be influential, but not for the initial reasons. Empirical Bayes shrinkage estimation, now a major topic, found its early justification in the James–Stein formula. Less obvious downstream topics include Tweedie's formula and Benjamini and Hochberg's false discovery rate algorithm. This is a short and mainly non-technical review of the James–Stein rule and its effects on the machine learning era of statistical innovation.

Keywords Empirical bayes · Shrinkage · Tweedie's formula · Benjamini–Hochberg algorithm

1 Introduction

By and large, the statistics world is one of heuristics, approximations, and asymptotics. The James–Stein estimator arrived in that world in 1961 on a note of startling specificity: unseen parameters $\mu_1, \mu_2, \ldots, \mu_n$ produce independent observations

$$x_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1), \qquad i = 1, \ldots, n, \tag{1}$$

$n \geq 3$. The James–Stein rule in its simplest form proposed estimating the μ_i by

$$\hat{\mu}_i^{\text{JS}} = \left(1 - \frac{n-2}{S}\right) x_i \qquad \left(S = \sum_{i=1}^{n} x_i^2\right). \tag{2}$$

Dedicated to the memory of Carl Morris.

✉ Bradley Efron
efron@stanford.edu

[1] Department of Statistics, Stanford University, 390 Jane Stanford Way, Stanford, CA 94305, USA

[2] Department of Biomedical Data Science, Stanford School of Medicine, 1265 Welch Road, Stanford, CA 94305, USA

Published online: 30 June 2023

Formula (2) looked implausible: the estimate of μ_i depended on the *other* observations x_j, $j \neq i$ (through S), as well as x_i, despite the independence assumption. Nevertheless, James and Stein showed that Rule (2) *always* beat the obvious maximum likelihood estimates

$$\hat{\mu}_i^{\mathrm{ML}} = \bar{x}_i \qquad (i = 1, \ldots, n) \tag{3}$$

in terms of total expected squared error

$$E\left\{\sum_{i=1}^{n} \left(\hat{\mu}_i - \mu_i\right)^2\right\}. \tag{4}$$

That "always" was the shocking part: two centuries of statistical theory, ANOVA, regression, multivariate analysis, etc., depended on maximum likelihood estimation. Did everything have to be rethought?

One path forward involved Bayesian thinking. If we assumed that the μ_i themselves came from a normal distribution,

$$\mu_i \overset{\text{ind}}{\sim} \mathcal{N}(0, A) \qquad \text{for } i = 1, \ldots, n, \tag{5}$$

with variance $A \geq 0$, the Bayes estimates would be

$$\hat{\mu}_i^{\text{Bayes}} = Bx_i \qquad (B = A/(A+1))\,. \tag{6}$$

We don't know A or B but

$$\widehat{B} = 1 - (n-2)/S \tag{7}$$

is B's unbiased estimate: we can rewrite (2) as

$$\hat{\mu}_i^{\mathrm{JS}} = \widehat{B}x_i, \tag{8}$$

which at least looks more plausible.

In the language introduced by Robbins (1956), formula (8) is an *empirical Bayes* estimator, another shocking post-war statistical innovation. Carl Morris and I wrote a series of papers in the 1970 s exploring Bayesian roots of the James–Stein estimator (Efron and Morris, 1973). Something is lost in the empirical Bayes formulation, namely the frequentist "always" of expected square error minimization, but a lot is gained in flexibility and scope, as discussed in Sect. 2.

Figure 1 illustrates an example of simultaneous estimation pursued in Sect. 2.1 of Efron (2010). A microarray study has compared expression levels between prostate cancer patients and control subjects for $n = 6033$ genes. For each gene, a statistic x_i has been calculated (essentially a "z-value"),

$$x_i \sim \mathcal{N}(\mu_i, 1), \qquad i = 1, \ldots, n, \tag{9}$$

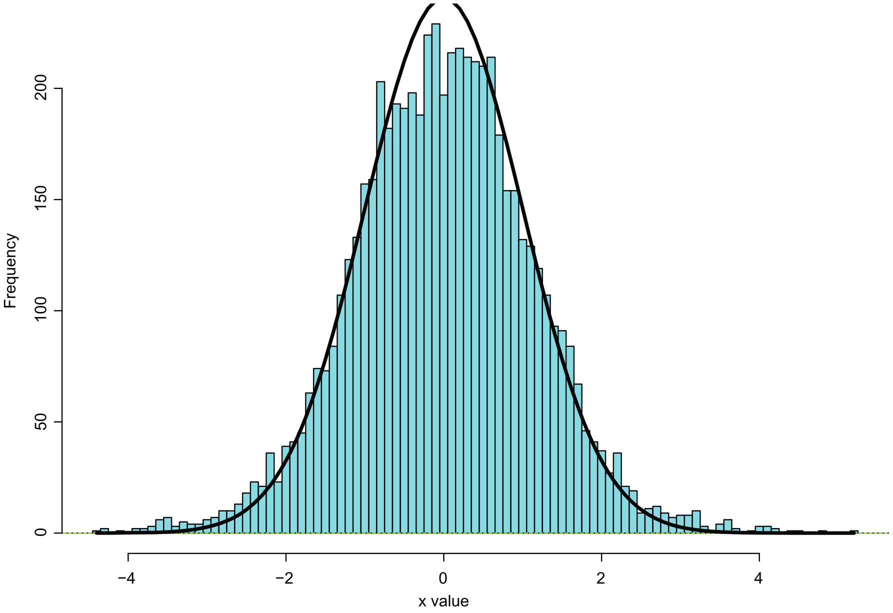

Fig. 1 Prostate data: 6033 x values; mean 0.003, sd $= 1.135$; curve is proportional to a $\mathcal{N}(0, 1)$ density

where μ_i measures the difference between cancer and control group levels.

The solid curve in Fig. 1 is a $\mathcal{N}(0, 1)$ density scaled to have the same area as the histogram of the 6033 x values. A bad result from the researchers' point of view would be a perfect fit of curve to histogram, which would imply all the genes have $\mu_i = 0$, the "null" value of no difference between cancer patients and controls.

That's not what happened: the histogram has mildly heavy tails in both directions. The researchers were hoping to find genes with large values of $\|\mu_i\|$—ones that might be a clue to prostate cancer etiology—as suggested by the heavy tails. How encouraged should they be?

Not very, according to the James–Stein rule. The 6033 x_i values have mean 0.003, which I'll take to be zero, and empirical variance

$$\hat{\sigma}^2 = 1.289. \tag{10}$$

The James–Stein estimate (2) is

$$\begin{aligned} \hat{\mu}_i^{\text{JS}} &= \left(1 - \frac{n-2}{n-1}\frac{1}{\hat{\sigma}^2}\right) x_i \\ &= 0.224 \cdot x_i, \end{aligned} \tag{11}$$

so even $x_i = 5$ yields an estimate barely exceeding 1. Section 2 suggests a more optimistic analysis.

2 Tweedie's formula

The impressive precision of the James–Stein theorem came at a cost in generality. Efforts to extend the theorem, say to Poisson rather than normal observations, or to measures of loss other than total squared error, gave encouraging asymptotic results but not the James–Stein kind of finite sample frequentist dominance.

Better progress was possible on the empirical Bayes side of the street. *Tweedie's formula* (Efron, 2011) has been particularly useful. We wish to calculate Bayesian estimates

$$\mu_i^{\text{Bayes}} = E\{\mu_i \mid x_i\}, \qquad i = 1, \dots, n, \tag{12}$$

in the normal sampling model (1), starting from a given (possibly non-normal) prior $\pi(\mu)$, applying to all n cases. Let $f(x)$ be the marginal density

$$f(x) = \int_{\mathcal{R}} \pi(\mu)\phi(x-\mu)\, d\mu, \tag{13}$$

with ϕ the standard $\mathcal{N}(0, 1)$ density and $\mathcal{R}$ the range of μ. (It isn't necessary for $\pi(\cdot)$ to be a continuous distribution but it simplifies notation.)

Tweedie's formula provides an elegant statement for μ_i^{Bayes}, the posterior expectation of μ_i given x_i,

$$\begin{aligned} \mu_i^{\text{Bayes}} &= E\{\mu_i \mid x_i\} = x_i + l'(x_i) \\ \text{with} \qquad l'(x_i) &= \frac{d}{dx}\log\left(f(x_i)\right). \end{aligned} \tag{14}$$

In the empirical Bayes situation (1), where the prior $\pi(\cdot)$ is unknown, we can use the observed data $x_1, \dots, x_n$ to estimate the marginal density $f(x)$, say by $\hat{f}(x)$, giving empirical Bayes estimates

$$\hat{\mu}_i = x_i + \hat{l}'(x_i). \tag{15}$$

The Bayes estimate (14) can be thought of as the MLE x_i plus a Bayesian correction term $l'(x_i)$. When the prior $\pi(\mu)$ is the $\mathcal{N}(0, A)$ distribution (5), μ_i^{Bayes} equals Bx_i (6). Simple formulas for μ_i^{Bayes} give out for most other choices of $\pi(\mu)$ but now, in the machine learning era[1] of statistical research, numerical methods provide useful ways forward, as discussed next.

The *log polynomial class*[2] of marginal densities defines $f(x)$ by

$$\log\left(f_\beta(x)\right) = \beta_0 + \beta^\top c(x). \tag{16}$$

[1] Where algorithms can substitute for theorems.

[2] For general use, a natural spline basis is preferable to polynomials, to control the behavior of $\log \pi(\mu)$ at the extremes.

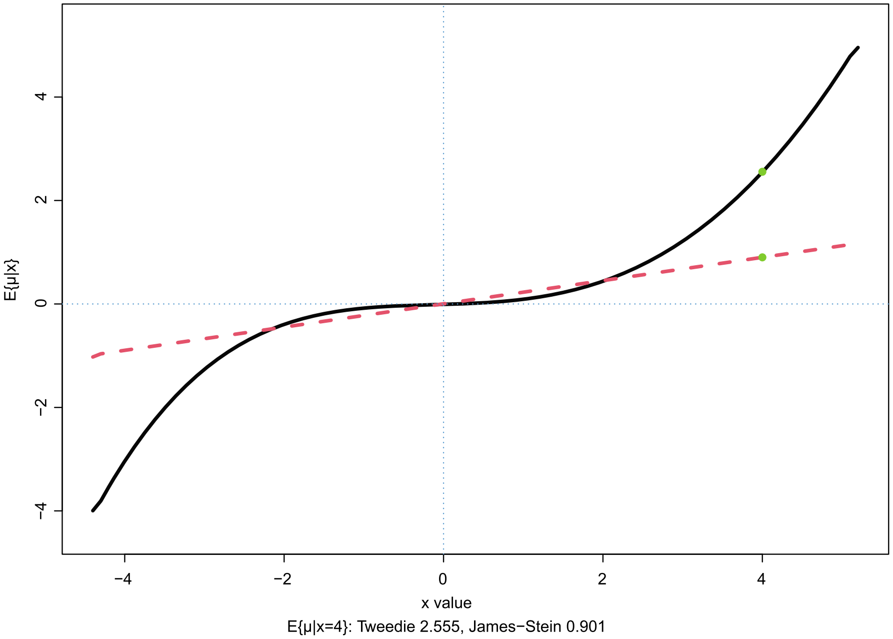

Fig. 2 Prostate data: Tweedie's estimate of $E\{\mu \mid x\}$, 5 degrees of freedom; dashed curve is James–Stein estimate

Here

$$c(x) = (x, x^2, \ldots, x^J)^\top \quad \text{and} \quad \beta = (\beta_1, \ldots, \beta_J)^\top, \tag{17}$$

with β_0 chosen to make $f_\beta(x)$ integrate to 1. The choice $J = 2$ gives normal marginals; larger values of J allow for marginal non-normality.

The choice $J = 5$ was applied to the prostate cancer data of Fig. 1: Tweedie's formula (14) gave $\hat{\mu}(x) = E\{\mu \mid x\}$, graphed as the solid curve in Fig. 2. It differs markedly from the James–Stein estimate $J = 2$, the dashed line. At $x = 4$ for example, the $J = 5$ estimate is[3]

$$E\{\mu \mid x = 4\} = 2.555 \tag{18}$$

compared to 0.901 for the James–Stein estimate.

The estimated curve $E\{\mu \mid x\}$ is *empirical Bayes* in the same sense as (8): the parameter vector β was selected by maximum likelihood, as discussed next. With $J = 5$, the prior was able to adapt to the "fishing expedition" nature of such microarray studies, where we expect most of the genes to be null or close to null, with μ_i nearly zero (corresponding here to the flat part of the curve for x between −2 and 2) and, hopefully, a small proportion of interestingly large μ_is.

[3] With an estimated bootstrap standard error of 0.192.

The sample size $n = 6033$ has much to do with Fig. 2. James and Stein (1961) was usually considered in terms of small samples, perhaps $n \leq 20$, for which there would be little hope of seeing the detail in Fig. 2. The term "machine learning era" seems less fanciful when considering the scale of problems statisticians are now asked to deal with, as well as the tools they use to solve them.

It looks like it might be hard work computing Fig. 2 but it's not. The histogram in Fig. 1 has 97 bins, with centerpoints

$$\boldsymbol{v} = (-4.4, -4.3, \ldots, 5.1, 5.2). \tag{19}$$

Let y_j be the count in bin j, that is, the number of the 6033 x_i values falling into it, with the vector of counts being

$$\boldsymbol{y} = (y_1, \ldots, y_{97}). \tag{20}$$

Then the single R command

$$\hat{\boldsymbol{l}} = \log\left(\text{glm}\left(\boldsymbol{y} \sim \text{poly}(\boldsymbol{v}, 5), \text{poisson}\right) \$ \,\text{fit}\right) \tag{21}$$

provides a close approximation to the MLE of $\log f(x)$ in (14); numerical differentiation of $\hat{\boldsymbol{l}}$ gives Tweedie's estimate. Section 3.4 of Efron (2023) shows why Poisson regression (21) is appropriate here.

The James–Stein theorem depends on the independence assumption in (1), unlikely to be true in the microarray study, but the estimates (2) have a certain marginal validity even under dependence. This is clearer from the empirical Bayes point of view. The Tweedie estimate $x_i + \hat{l}'(x_i)$ requires only that $\hat{l}'(x)$ be close to $l'(x)$, not that it be estimated from independent x_is.[4]

3 Shrinkage estimators

James and Stein's paper aroused excited interest in the statistics community when it arrived in 1961. Most of the excitement focused on the strict inadmissibility of the traditional maximum likelihood estimate demonstrated by the James–Stein rule. Other rules dominating the MLE were discovered, for instance the Bayes estimator of Strawderman (1971), that was itself admissible while rendering the MLE inadmissible.

Big new ideas can take a while to make their true impact felt. The James–Stein rule had an influential side effect on subsequent theory and practice in that it demonstrated, in an inarguable way, the virtues of *shrinkage estimation*: given an ensemble of problems, individual estimates are shrunk toward a central point; that is, a deliberate bias is introduced, pulling estimates away from their MLEs for the sake of better group performances.

Admissibility and inadmissibility aren't much in the air these days, while shrinkage estimation has gone on to play a major role in modern practice. A spectacular success

[4] The accuracy of the Tweedie estimate *does* suffer under dependence, so the previously quoted bootstrap standard error is likely to be optimistic.

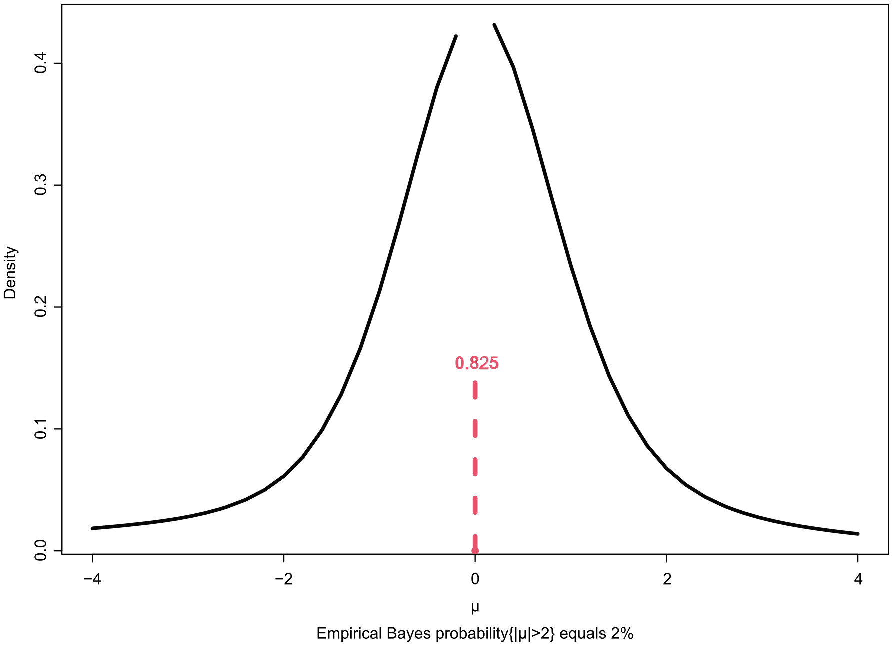

Fig. 3 Empirical Bayes conditional density of μ given μ not zero; $\Pr\{\mu = 0\}$ equals 0.825

story is the lasso (Tibshirani, 1996). Lasso shrinkage is extreme, pulling some (often most) of the coefficient estimates all the way back to zero.

Bayes and empirical Bayes rules tend to be strong shrinkers. Tweedie's estimate in Fig. 2 ($J = 5$) shrinks the estimate of $E\{\mu \mid x = 4\}$ from its MLE value 4 down to 2.555. For μ between -1 and 1, the shrinkage is almost all the way to zero.

The reader may have been surprised to see that neither Tweedie's formula (14) for $E\{\mu_i \mid x_i\}$ nor its empirical version (15) require estimation of the prior $\pi(\mu)$. This is a special property of the posterior expectation $E\{\mu_i \mid x_i\}$ and isn't available for say $\Pr\{\mu_i \geq 2 \mid x_i\}$, or most other Bayesian targets.

"Bayesian deconvolution" (Efron, 2016) uses low-dimensional parametric modeling of $\pi(\mu)$ for general empirical Bayes computations. It was applied to finding a prior density $\pi(\mu)$ that would give the distribution of x seen in Fig. 1, assuming the normal sampling model (1). The deconvolution model for $\pi(\mu)$ used a delta function at $\mu = 0$ (for the "null" genes) and a natural spline function with four degrees of freedom for the non-null cases.

The estimated prior[5] $\hat{\pi}(\mu)$ is shown in Fig. 3; it put probability 0.825 on $\mu = 0$, while the conditional distribution given $\mu \neq 0$ was a moderately heavy-tailed version of $\mathcal{N}(0, 1.33^2)$. Based on $\hat{\pi}(\mu)$ we can form estimates of *any* Bayesian target, for instance $\widehat{\Pr}\{\mu_i \geq 2 \mid x_i = 4\} = 0.80$. Figure 3 is a direct descendent of the James–Stein rule, now 60-plus years on.

[5] Estimated using the CRAN package `deconvolveR` (Narasimhan and Efron, 2020).

A less-direct descendent, but still on the family tree, arrived in 1995. The *false discovery rate* paper by Benjamini and Hochberg concerned simultaneous hypothesis testing. Looking at Fig. 1, which of the $n = 6033$ genes can confidently be labeled as non-null, that is as having $\mu_i \neq 0$?

Suppose for convenience that the x_is are ordered from smallest to largest. The right-sided significance level for testing $\mu_i = 0$ is

$$S_0(x_i) = 1 - \Phi(x_i), \tag{22}$$

where Φ is the standard normal cumulative distribution function. Of the 6033 genes, 401 had $S_i \leq 0.05$, the usual rejection level for individual testing, but even if actually *all* of the genes were null we would expect 302 such rejections, so individual testing can't be right. Benjamini and Hochberg proposed a novel simultaneous testing rule that safely controls the number of "false discoveries" — genes falsely labeled ”non-null" — while not being discouragingly strict. (My summary here won't give the BH rule its full due; see Chapter 4 of Efron (2010) for a more complete description.)

Let $\widehat{S}(x)$ be the observed proportion of x_is exceeding value x, and define

$$\widehat{\text{Fdr}}(x) = \pi_0 S_0(x)/\widehat{S}(x), \tag{23}$$

where π_0 is the proportion of null genes among all n.[6] For a fixed control level α, such as $\alpha = 0.1$, the BH rule says to reject the null hypothesis $\mu_i = 0$ for those genes having

$$\widehat{\text{Fdr}}(x_i) \leq \alpha. \tag{24}$$

The Benjamini–Hochberg theorem states that under independence assumptions like (1), the expected proportion of false discoveries by rule (24) is α.

Figure 4 shows $\widehat{\text{Fdr}}$ for the prostate cancer data and also for the left-sided Fdr estimate, where significance is defined by $S_0(x_i) = \Phi(x_i)$ rather than (22). I applied the BH rule with $\alpha = 0.1$ which labeled 60 genes as non-null, 32 on the left and 28 on the right. The BH theorem says that we can expect 6 of the 60 to actually be null.

The fdr story has evolved very much along the lines of its James–Stein predecessor. Intense initial interest focused on the exact frequentist control of false discovery rates. The Bayes and empirical Bayes implications came later: as at (5), we assume that each x_i is a realization of a random variable x given by

$$\mu \sim \pi(\mu) \quad \text{and} \quad x \mid \mu \sim p(x \mid \mu), \tag{25}$$

where $p(x \mid \mu)$ is a known probability kernel which I'll take here to be the normal sampling model (1). Then if $S(x)$ is 1 minus the cdf of the marginal density (13), Bayes rule gives

$$\Pr\{\mu = 0 \mid x\} = \pi_0 S_0(x)/S(x). \tag{26}$$

[6] π_0 can be estimated but in practice it is usually replaced by its upper bound 1 in applying rule (24). For cases like the prostate data where most of the genes are null, this doesn't much affect the outcome.

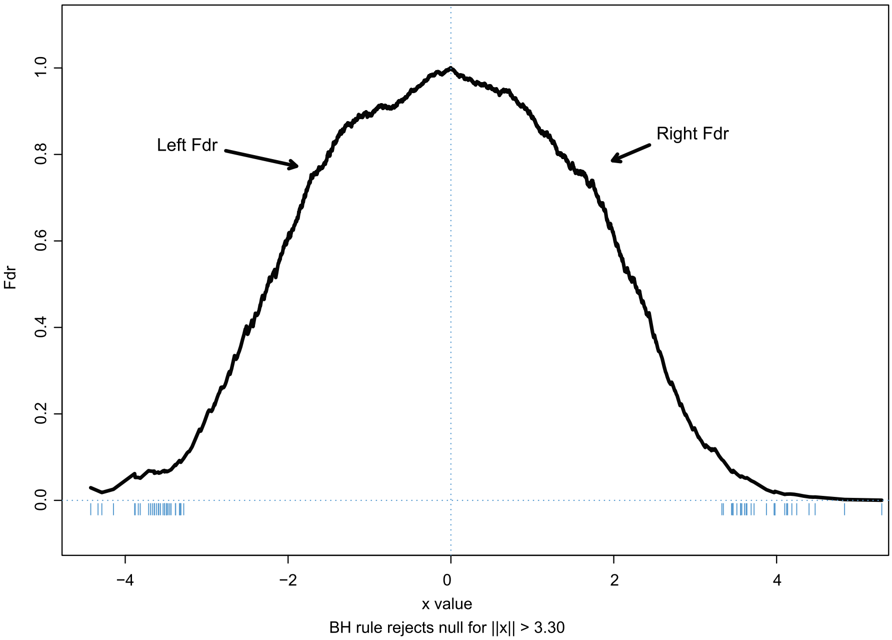

Fig. 4 Prostate data: left Fdr and right Fdr; dashes show 60 genes with Fdr < 0.1

Comparing (26) with (23) says that the BH rule amounts to labeling case i as non-null if its obvious empirical Bayes estimate of nullness is less than α. This is less precise than the frequentist control theorem but, as with the James–Stein estimator, is more robust in not demanding independence among the x_is. The family resemblance between JS and BH is through shrinkage: in the BH case the shrinkage of significance levels. For instance, $x_i = 3$ has individual significance level 0.001 against nullness, whereas $\widehat{\text{Fdr}} = 0.164$ for the prostate data, i.e, still with about a $1/6$ chance of gene i being null.

So what does machine learning have to do with the James–Stein estimator? Nothing to its birth but, as the articles in this volume show, a great deal to its downstream effects on statistical theory and practice. Charles Stein, who was a good applied statistician when he put his mind to it, might have enjoyed these developments, but maybe not; his heart was always with the mathematics.

Funding No funds, grants, or other support was received.

Declarations

Conflict of interest The author has no relevant financial or non-financial interests to disclose.

References

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B, 57*(1), 289–300.

Efron, B. (2010). *Large-scale inference: Empirical Bayes methods for estimation, testing, and prediction* (Vol. 1). Cambridge: Cambridge University Press.

Efron, B. (2011). Tweedie's formula and selection bias. *Journal of the American Statistical Association, 106*(496), 1602–1614. https://doi.org/10.1198/jasa.2011.tm11181

Efron, B. (2016). Empirical Bayes deconvolution estimates. *Biometrika, 103*(1), 1–20. https://doi.org/10.1093/biomet/asv068

Efron, B. (2023). *Exponential Families in Theory and Practice*. Cambridge: Cambridge University Press.

Efron, B., & Morris, C. (1973). Stein's estimation rule and its competitors—An empirical Bayes approach. *Journal of the American Statistical Association, 68*, 117–130.

James, W., & Stein, C. (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Sympos. Math. Statist. and Prob.* (Vol. I, pp. 361–379). Berkeley: University of California Press.

Narasimhan, B., & Efron, B. (2020). deconvolveR: A G-Modeling Program for Deconvolution and Empirical Bayes Estimation. *Journal of Statistical Software, 94*(11), 1–20. https://doi.org/10.18637/jss.v094.i11

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proc. 3rd Berkeley Sympos. Math. Statist. and Prob.* (Vol. I, pp. 157–163). Berkeley: University of California Press.

Strawderman, W. E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Annals of Mathematical Statistics, 42*(1), 385–388. https://doi.org/10.1214/aoms/1177693528

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