Statistics GR8201-Spring 2023

Topics in Theoretical Statistics: Empirical Bayes

Nikolaos Ignatiadis Bodhisattva Sen

2023-04-30

Table of contents

Preface

These are lecture notes accompanying the course GR8201, "Topics in Theoretical Statistics: Empirical Bayes" that was taught at the Department of Statistics, Columbia University in Spring 2023.

We thank Arnab Auddy, Casey Bradshaw, Fangyi Chen, Maria-Cristiana Girjau, Russell Kunes, Zhenyuan Liu, Jonas Mikhaeil, and Yizi Zhang for providing feedback on these lecture notes.

1 Introductory formal perspectives on empirical Bayes

Our first goal is to demonstrate what's different in the empirical Bayes problem compared to more classical statistical decision theory. In other words, we will get a theoretical glimpse of the new possibilities presented by the empirical Bayes approach. In the next lecture we will see the relevance of these results to problems in applications.

This chapter will largely follow the three seminal papers of Herbert Robbins, who introduced the term "empirical Bayes" (Robbins 1951, 1956, 1964).

You may find an eloquent and beautiful introduction to empirical Bayes (and particularly its origins in Robbins (1951)) by Herbert Robbins himself in the following video:

https://youtu.be/id6YSycD5lc

1.1 Simple statistical decision problems

A simple statistical decision problem consists of the following ingredients:

Definition 1.1 (Simple statistical decision problem)**.**

A) Unknown parameter $\theta \in \Theta$.

B) Observed random variable $Z \sim p(\cdot \mid \theta)$, where $Z \in \mathcal{Z}$. $p(\cdot \mid \theta)$ is a density with respect to a measure λ on $\mathcal{Z}$. λ is typically the counting measure on $\mathcal{Z} = \mathbb{N}_{\geq 0}$ or the Lebesgue measure on $\mathcal{Z} = \mathbb{R}$. [In the empirical Bayes setting, $p(\cdot \mid \theta)$ is often called the likelihood or the noise distribution.]

C) Possible decisions $t \in \mathcal{T}$.

D) Loss $\ell(t, \theta)$ that is incurred by decision t when the true parameter is θ.

1.1.1 The risk function

The statistician's goal is to choose a data-driven decision $t = t(Z)$, that is, a decision that depends on the observed random variable Z, such that the risk ($\hat{=}$ expected loss)

$$R(t(\cdot), \theta) = \mathbb{E}_\theta \left[\ell(t(Z), \theta)\right] \tag{1.1}$$

is as small as possible. Note that by $\mathbb{E}_\theta$ we denote an expectation with respect to $Z \sim p(\cdot \mid \theta)$ with θ fixed. It is well known that seeking $t(\cdot)$ that minimizes Eq. 1.1 is not a well-defined task, i.e., there is no uniformly (over θ) best decision rule $t(\cdot)$ (Erich Leo Lehmann and Casella 1998, chap. 1.1). Hence one needs to find a way to collapse Eq. 1.1 into a single number (or alternatively to constrain the decision rules under consideration).

1.1.2 The Bayesian approach

In the Bayesian approach the statistician further posits that $\theta \sim G$ for a known distribution G, the **prior**. Then, it is natural to seek a rule $t(\cdot)$ such that the risk $R(t(\cdot), \theta)$ integrated over G is as small as possible:

$$R(t(\cdot), G) = \mathbb{E}_G \left[\ell(t(Z), \theta)\right] = \mathbb{E}_G \left[R(t(\cdot), \theta)\right] = \int R(t(\cdot), \theta) dG(\theta) \tag{1.2}$$

The optimal decision is called the Bayes decision $t_G(\cdot)$ and its risk is called the Bayes risk $R(G)$, that is:

$$t_G(\cdot) \in \operatorname*{argmin}_{t(\cdot)} \left\{R(t(\cdot), G)\right\}, \quad R(G) = R(t_G(\cdot), G) = \inf_{t(\cdot)} \left\{R(t(\cdot), G)\right\}.$$

In the Bayesian approach, typically one can base optimal decisions on the posterior distribution of θ given Z. The posterior distribution of θ given $Z = z$ has $dG(\theta)$-density $p(z \mid \theta)/f_G(z) dG(\theta)$, where $f_G(\cdot)$ is the marginal density of Z, i.e.,

$$f_G(z) = \int p(z \mid \theta) dG(\theta). \tag{1.3}$$

When G has a density g with respect to a measure ν, then we write the following for the $(d\nu)$-density of the posterior distribution of θ given $Z = z$,

$$p_G(\theta \mid z) = \frac{p(z \mid \theta)}{f_G(z)} g(\theta). \tag{1.4}$$

The optimal $t_G(z)$ can typically be computed by minimizing the posterior risk, that is the expectation of the loss with respect to the posterior Eq. 1.4.

Proposition 1.1. *Suppose that for all $z \in \mathcal{Z}$, there exists $t = t_z \in \mathcal{T}$ that minimizes $\mathbb{E}_G[\ell(t,\theta) \mid Z = z]$ over all $t \in \mathcal{T}$. Then the Bayes decision is given by:*

$$t_G(z) = \operatorname*{argmin}_{t \in \mathcal{T}} \{\mathbb{E}_G[\ell(t,\theta) \mid Z = z]\}.$$

Proof. Write

$$t^*(z) = \operatorname*{argmin}_{t \in \mathcal{T}} \{\mathbb{E}_G[\ell(t,\theta) \mid Z = z]\}.$$

We need to prove that the above minimizes the risk, so that $t_G(\cdot) = t^*(\cdot)$ indeed provides a Bayes optimal decision. To this end, take any other decision $\tilde{t}(\cdot)$. Then:

$$\begin{aligned}
R(t^*(\cdot), G) &= \mathbb{E}_G[\ell(t^*(Z), \theta)] \\
&= \mathbb{E}_G[\mathbb{E}_G[\ell(t^*(Z), \theta) \mid Z]] \\
&= \mathbb{E}_G\left[\inf_{t \in t} \mathbb{E}_G[\ell(t,\theta) \mid Z]\right] \\
&\leq \mathbb{E}_G\left[\mathbb{E}_G\left[\ell(\tilde{t}(Z), \theta) \mid Z\right]\right] \\
&= R(\tilde{t}(\cdot), G).
\end{aligned}$$

Since $\tilde{t}(\cdot)$ was arbitrary, we conclude.

□

Proposition 1.2. *Suppose $\theta \sim G$ and $Z \mid \theta \sim p(\cdot \mid \theta)$.*

1. *Suppose the decision space is equal to the parameter space, $\mathcal{T} = \Theta$, and we seek to estimate θ in squared error, that is, $\ell(t,\theta) = (t-\theta)^2$. Suppose further that $\mathbb{E}_G[\theta^2] < \infty$. Then the Bayes-optimal decision is given by:*

$$t_G(z) = \mathbb{E}_G[\theta \mid Z = z].$$

2. *Suppose that $\Theta = \{\theta_a, \theta_b\}$ consists of only two elements and that we need to choose between θ_a and θ_b, that is $\mathcal{T} = \{\theta_a, \theta_b\}$ with loss $\ell(t,\ \theta) = \mathbf{1}(t \neq \theta)$. Then a Bayes-optimal decision is given by:*

$$t_G(z) = \begin{cases} \theta_b, & \text{if } p_G(\theta_b \mid z) > p_G(\theta_a \mid z) \\ \theta_a, & \text{otherwise} \end{cases}.$$

3. *Let $\theta^* \in \Theta$ and suppose we seek to disambiguate between the following hypotheses: $H_{\leq} : \theta \leq \theta^*$ and $H_{>} : \theta > \theta^*$. Our decision space is $\mathcal{T} = \{H_{\leq}, H_{>}\}$. Suppose furthermore that we incur 0 loss when we choose correctly between $H_{\leq}$ and $H_{>}$, and a loss proportional to $|\theta - \theta^*|$ otherwise. In other words, our loss is:*

$$\ell(t,\theta) = \left(\mathbf{1}(t = H_{\leq}, \theta > \theta^*) + \mathbf{1}(t = H_{<}, \theta \geq \theta^*)\right)|\theta - \theta^*|.$$

If $\mathbb{E}_G[|\theta|] < \infty$, then a Bayes-optimal decision is given by:

$$t_G(z) = \begin{cases} H_>, & \text{if } \mathbb{E}_G[\theta \mid Z = z] > \theta^* \\ H_\leq, & \text{otherwise} \end{cases}.$$

Proof. For 1., we note that:

$$\mathbb{E}_G\left[(\theta - t)^2 \mid Z = z\right] = \mathrm{Var}_G[\theta \mid Z = z] + \left(\mathbb{E}_G[\theta \mid Z = z] - t\right)^2.$$

Hence the above is minimized at $t = t_G(z) = \mathbb{E}_G[\theta \mid Z = z]$ as claimed.

For 2.:

$$\mathbb{E}_G[\ell(t,\ \theta) \mid Z = z] = \mathbb{P}_G[t \neq \theta \mid Z = z] = p_G(\theta_a \mid z)\mathbf{1}(t = \theta_b) + p_G(\theta_b \mid z)\mathbf{1}(t = \theta_a).$$

Hence the posterior risk will either be $p_G(\theta_b \mid z)$ for $t = \theta_a$ and $p_G(\theta_a \mid z)$ for $t = \theta_b$. Thus the claimed rule $t_G(z)$ minimizes the posterior risk.

For 3., it is convenient to first note the following

$$\ell(H_>, \theta) - \ell(H_\leq, \theta) = \theta^* - \theta.$$

Hence

$$\mathbb{E}_G[\ell(H_>, \theta) \mid Z = z] - \mathbb{E}_G\left[\ell(H_\leq, \theta) \mid Z = z\right] = \theta^* - \mathbb{E}_G[\theta \mid Z = z]. \tag{1.5}$$

Hence e.g., $\mathbb{E}_G[\ell(H_>, \theta) \mid Z = z] < \mathbb{E}_G\left[\ell(H_\leq, \theta) \mid Z = z\right]$ when $\mathbb{E}_G[\theta \mid Z = z] > \theta^*$ which justifies the suggested rule.

□

1.1.3 The frequentist approach, particularly: minimax decisions

In the minimax approach one seeks a decision rule $t_{\text{minmax}}(\cdot)$ such that:

$$t_{\text{minmax}}(\cdot) \in \operatorname*{argmin}_{t(\cdot)} \sup_{\theta \in \Theta} \{R(t(\cdot), \theta)\}. \tag{1.6}$$

The worst-case risk of $t_{\text{minmax}}(\cdot)$ is called the minimax risk.

Exercise 1.1. Prove the following statements:

1. For any decision function $t(\cdot)$, it holds that

$$\sup_{\theta \in \Theta} \{R(t(\cdot), \theta)\} = \sup_G \{R(t(\cdot), G)\},$$

where the supremum in RHS above is taken with respect to all priors G supported on Θ.

2. Let $t_G(\cdot)$ be a Bayes-optimal decision function with respect to the prior G supported on Θ. Furthermore, suppose that $R(t_G(\cdot), \theta)$ is constant for all $\theta \in \Theta$. Then $t_G(\cdot)$ is minimax optimal as defined in Eq. 1.6.

The above exercise demonstrates that one may think of the minimax approach as a conservative approach to mimicking a Bayesian in the case wherein the analyst has no knowledge about the prior G.

1.2 Statistical decisions with parallel problems

In an empirical Bayes analysis, we simultaneously face multiple simple statistical decision problems as in Definition 1.1.

Definition 1.2 (Statistical decisions with parallel problems)**.**

A) Unknown parameters $\theta_1, \dots, \theta_n \in \Theta$. We write $\boldsymbol{\theta} = \boldsymbol{\theta}_{1:n} = (\theta_1, \dots, \theta_n)$ for the concatenation of all the θ_i.

B) Observed random variables $Z_i \sim p(\cdot \mid \theta_i)$, $i = 1, \dots, n$ (typically assumed conditionally independent). We write $\mathbf{Z} = \mathbf{Z}_{1:n} = (Z_1, \dots, Z_n)$ for the concatenation of all the Z_i.

C) Decisions $t_1, \dots, t_n \in \mathcal{T}$. As above, also write $\mathbf{t} = \mathbf{t}_{1:n} = (t_1, \dots, t_n)$.

D) Losses $\ell(t_i, \theta_i)$, $i = 1, \dots, n$ incurred by decision t_i when the true parameter is θ_i.

Parallel problems such as the above are also called *compound problems* when the goal is to minimize the average of the losses, that is:

$$\ell(\mathbf{t}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell(t_i, \theta_i). \tag{1.7}$$

Our decision t_i for the i-th problem can depend on data for all problems, that is $t_i = t_i(\mathbf{Z})$ instead of $t_i = t_i(Z_i)$. More broadly, our decisions for all the component problems take the form:

$$\mathbf{t}(\mathbf{Z}) = (t_1(\mathbf{Z}), \dots, t_n(\mathbf{Z})). \tag{1.8}$$

A, perhaps surprising,[1] take-home message of empirical Bayes theory is that it can be beneficial to make decisions for the i-th problem based on the data for all the problems!

[1] Robbins (1951) writes the following: "It is natural to suppose that the 'best' solution of the compound problem consists in applying to each of the Z_i the 'best' solution of the original simple problem."

1.2.1 The compound risk function

In the present setting, we typically evaluate procedures with respect to their compound risk, that is, with respect to the expectation of Eq. 1.7.

In frequentist terms we integrate Eq. 1.7. with respect to $\mathbf{Z}$ for fixed $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$:

$$R(\mathbf{t}(\cdot), \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}} \left[\ell(t_i(\mathbf{Z}), \theta_i) \right].$$

Analogously, in the Bayesian case, wherein we assume $\theta_i \sim G$ for $i = 1, \dots, n$:

$$R(\mathbf{t}(\cdot), G) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{G} \left[\ell(t_i(\mathbf{Z}), \theta_i) \right].$$

The expectation above is taken with respect to the n independent draws of the pairs (θ_i, Z_i), where $\theta_i \sim G$ and $Z_i \mid \theta_i \sim p(\cdot \mid \theta_i)$.

When $\theta_i \sim G$, Robbins considered one further way of evaluating procedures. He considered a hypothetical sequential task in which the analyst uses data from the first n experiments to construct a decision rule $\hat{t}(\cdot) = \hat{t}(\cdot; \mathbf{Z})$ to be applied to a hypothetical $n+1$-th decision problem in the future. The risk of $\hat{t}(\cdot)$ is then merely its Bayes risk with respect to the $n+1$-th simple decision problem:

$$R_{n+1}(\hat{t}(\cdot), G) = \mathbb{E}_{G} \left[\ell(\hat{t}(Z_{n+1}), \theta_{n+1}) \mid \hat{t} \right].$$

We note the quantity above is random since $\hat{t}$ is random (as a function of $Z_1, \dots, Z_n$).

1.2.2 The value of large scale data

Before proceeding to concrete demonstrations of the empirical Bayes approach, we take a moment to highlight three of the new possibilities unlocked by empirical Bayes methods (which were not available to us when we were faced with isolated simple statistical decision problems).

When we are faced with these parallel problems, several new features become available to us compared to the situation wherein we are dealing with an individual component problem.

1. For each simple component statistical problem we can typically never learn the parameter θ_i precisely.[2] However, we *can* precisely learn about properties of the bulk of parameters $\{\theta_1, \dots, \theta_n\}$.

2. Suppose we are willing to posit that $\theta_i \overset{\text{iid}}{\sim} G$, but *not* that G is known to us.[3] This setting is not amenable to analysis with data from a simple statistical decision

problem. However, with many parallel problems, it becomes possible to construct decision procedures that match the Bayes risk of a Bayes oracle that knows G.

3. Suppose we are not willing to assume that $\theta_i \sim G$ and instead assume that $\theta_1, \ldots, \theta_n$ are deterministic. Even in that case, it is often possible to come up with decision rules $\mathbf{t}(\mathbf{Z})$ (see Eq. 1.8) that depend on all the data that perform better in terms of the compound loss in Eq. 1.7 compared to the more "standard" approach wherein the i-th decision depends on only data for the i-th statistical problem (that is, $t_i = t_i(Z_i)$).

1.3 A simple vs. simple testing problem studied by Robbins (1951)

Example 1.1. Robbins (1951) considered the following decision problem.

A) Parameter: $\theta \in \Theta = \{-1, +1\}$.

B) Likelihood $Z \sim \mathcal{N}(\theta, 1)$.

C) $\mathcal{T} = \{-1, +1\}$.

D) Loss $\ell(t, \theta) = \mathbf{1}(t \neq \theta)$.

First suppose that we are willing to assume that $\theta \sim G$. Since $\theta \in \{-1, +1\}$, it follows that G is completely specified by $\pi = G(\{1\})$. We write for simplicity $G = G^\pi$ for the prior that assigns mass π to $+1$ and mass $1 - \pi$ to -1.

By applying Part 2. of Proposition 1.2, we get the following:

Proposition 1.3. *The Bayes decision for the prior G_π is equal to:*

$$t_{G^\pi}(z) = \begin{cases} +1, & \text{if } z > \frac{1}{2}\log\left((1-\pi)/\pi\right) \\ -1, & \text{otherwise} \end{cases}. \tag{1.9}$$

The Bayes risk then is equal to:

$$R(G^\pi) = \pi\Phi\left(\frac{1}{2}\log\left((1-\pi)/\pi\right) - 1\right) + (1-\pi)\Phi\left(-\frac{1}{2}\log\left((1-\pi)/\pi\right) - 1\right).$$

[3]In asymptotic statistics, we consider the regime wherein we collect more and more data for the simple statistical decision problem. In an empirical Bayes analysis instead the typical assumption is that we have *a lot* of parallel statistical problems, but *limited* data for each individual problem.

[3]As noted by Good (1992), for a Bayesian G often represents epistemic uncertainty. When an empirical Bayesian claims that "G exists", then the connotation is that G is a physical object: the frequency distribution of the parameters $\theta_1, \theta_2, \ldots$.

Here Φ is the standard normal distribution function.

The Bayes rule for $\pi = 1/2$, that is, $t_{G^{1/2}}(z) = \mathbf{1}(z > 0)$ is minimax optimal and has constant risk:

$$R(t_{G^{1/2}}, \theta) = \Phi(-1) \approx 0.1587 \text{ for } \theta \in \{-1, 1\}.$$

In contrast, for any $\pi \neq 1/2$:

$$\max_{\theta \in \{-1,+1\}} R(t_{G^\pi}, \theta) > \Phi(-1) \approx 0.1587.$$

Proof. The form of the Bayes-optimal decision follows from Proposition 1.2 (Part 2). The frequentist risk function of any such t_{G^π} is equal to:

$$R(t_{G^\pi}, \theta) = \begin{cases} \Phi\left(\frac{1}{2}\log\left((1-\pi)/\pi\right) - 1\right), & \text{if } \theta = 1 \\ \Phi\left(-\frac{1}{2}\log\left((1-\pi)/\pi\right) - 1\right), & \text{if } \theta = -1. \end{cases}$$

From here we can read off their Bayes risk. Furthermore we see that $R(t_{G^{1/2}}, \theta) = \Phi(-1)$ for both $\theta \in \{-1, +1\}$, so minimax optimality follows from Exercise 1.1.

□

In words: in this simple setting, the Bayesian approach can outperform the minimax approach by decreasing its loss for the $\theta \in \{-1, +1\}$ that has the largest prior probability. The price to pay is that the risk for the θ with the lower prior probability increases above that of the minimax rule that decides according to the sign of Z_i.

Robbins (1951) showed that this conundrum largely disappears when one faces many parallel problems.

1.3.1 Empirical Bayes results

Suppose we have n independent pairs (θ_i, Z_i) from the simple decision problem in Example 1.1 and further suppose that $\theta_i \sim G = G^\pi$.

Above, it was unclear whether we have any knowledge of π (perhaps we do in some settings; in many others we don't), and in the absence of it, it would be reasonable to proceed with the minimax decision, $t_i = \mathbf{1}(Z_i > 0)$. However, in the empirical Bayes setting, we can learn about π! For example:

$$\mathbb{E}_G[Z_i] = \pi \cdot 1 + (1-\pi) \cdot (-1) = 2\pi - 1, \text{ i.e.,} \mathbb{E}_G[(Z_i + 1)/2)] = \pi.$$

Hence for n sufficiently large, we can get a good estimate of π as follows:

$$\hat{\pi} = \frac{1}{2n}\sum_{i=1}^{n}(Z_i + 1).$$

Then we can consider the following empirical Bayes decision rule based on $\mathbf{Z} = (Z_1, \dots, Z_n)$:

$$\hat{t}_i^{\mathrm{EB}}(\mathbf{Z}) = \begin{cases} +1, & \text{if } Z_i > \frac{1}{2}\log\left((1-\hat{\pi})/\hat{\pi}\right) \\ -1, & \text{otherwise} \end{cases}. \tag{1.10}$$

The above decision is the Bayes decision Eq. 1.9 with π replaced by the estimate $\hat{\pi}$.

It is possible to prove the following:

$$\mathbb{E}_{G^\pi}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\hat{t}_i^{\mathrm{EB}}(\mathbf{Z}), \theta_i)\right] \to R(G^\pi) \text{ as } n \to \infty.$$

The above means that by allowing ourselves to use decision rules that depend on *all* the data, we are able to match the Bayes risk without knowing the prior!

1.3.2 Results in the compound setting

Robbins (1951) presented the analysis above. However, in most of the paper he considered the case wherein $\theta_1, \dots, \theta_n$ are deterministic and not random according to G; he wrote "the assumption of an existing but unknown prior distribution G will be questionable in most applications of statistics."

A fascinating result is that even for deterministic θ_i, it turns out that it can be beneficial to use the decision rule Eq. 1.10. This was considered a breakthrough by Neyman (1962) and the conclusion defies the following argument: "At first sight it may seem that the use of decision functions of the general form $t_i(Z_1, \dots, Z_n)$ is pointless, since the values Z_j for $j \neq i$ can contribute no information concerning θ_i" (Robbins 1951).

So, what is going on? Observe the following. Let:

$$\pi_n = \pi_n(\theta_1, \dots, \theta_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(\theta_i = 1)$$

In the compound setting we have that:

$$\frac{1}{2n}\sum_{i=1}^{n}(Z_i + 1) \sim \mathcal{N}(\pi_n,\, 1/(4n)).$$

Thus

$$\hat{t}_i^{\mathrm{EB}}(\mathbf{Z}) \approx \begin{cases} +1, & \text{if } Z_i > \frac{1}{2}\log\left((1-\pi_n)/\pi_n\right) \\ -1, & \text{otherwise} \end{cases}.$$

Hence:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\ell(\hat{t}_i^{\mathrm{EB}}(\mathbf{Z}), \theta_i)\right] \approx \begin{cases} \Phi\left(\frac{1}{2}\log\left((1-\pi_n)/\pi_n\right) - 1\right), & \text{if } \theta_i = 1 \\ \Phi\left(-\frac{1}{2}\log\left((1-\pi_n)/\pi_n\right) - 1\right) & \text{if } \theta_i = -1. \end{cases}$$

Averaging over $i = 1, \dots, n$, we thus see:

$$\mathbb{E}_{\boldsymbol{\theta}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(\hat{t}_i^{\mathrm{EB}}(\mathbf{Z}), \theta_i) \right]$$
$$\approx \pi_n \Phi \left(\frac{1}{2} \log \left((1 - \pi_n)/\pi_n \right) - 1 \right) + (1 - \pi_n) \Phi \left(-\frac{1}{2} \log \left((1 - \pi_n)/\pi_n \right) - 1 \right).$$

Next note that the RHS is the Bayes risk $R(G^{\pi_n})$ under the prior G^{π_n}. Hence the above argument demonstrates that:

$$\mathbb{E}_{\boldsymbol{\theta}} \left[\frac{1}{n} \sum_{i=1}^{n} \ell(\hat{t}_i^{\mathrm{EB}}(\mathbf{Z}), \theta_i) \right] \approx R(G^{\pi_n}) \leq \Phi(-1).$$

In other words, the unconventional decision rule Eq. 1.10 that uses all data to make decisions for individual simple problems outperforms the more conventional (and minimax) rule $t_i(Z_i) = \mathbf{1}(Z_i > 0)$ for almost all configurations of θ_i. For this reason, Robbins called the decision rule Eq. 1.10 an asymptotically subminimax solution. We visualize the conclusions in the figure below:

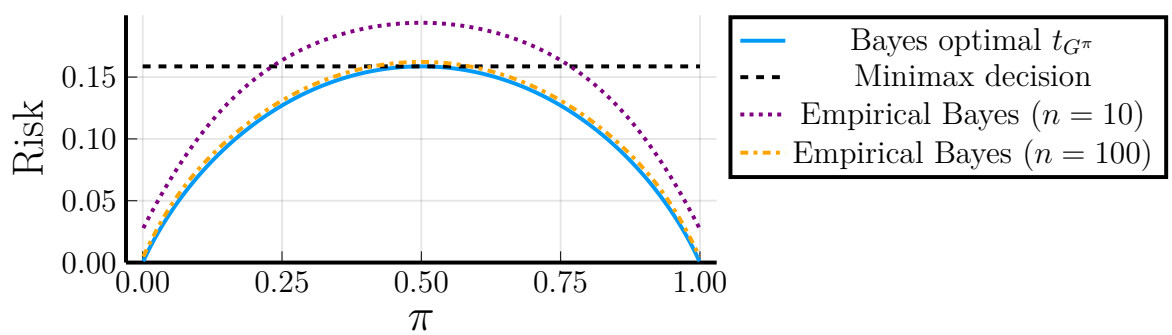


Figure 1.1: Risk functions in the stylized example of testing for $\theta_i = 1$ vs. $\theta_i = -1$ based on $Z_i \sim \mathcal{N}(\theta_i, 1)$, $i = 1, \dots, n$.

Robbins (1951) infamously wrote that "Z_1 could be an observation on a butterfly in Ecuador, Z_2 on an oyster in Maryland, Z_3 the temperature of a star" to emphasize that the parameters θ_i above may be deterministic and completely unrelated to each other. Nowadays, most statisticians would agree that the empirical Bayes approach only makes conceptual sense when applied to problems that are in fact related. Indeed, positing the compound loss in Eq. 1.7 makes most conceptual sense when the problems are related (in some form). Results in the compound setting are nevertheless considered important to this day. For example, they showcase a form of modeling robustness of an empirical Bayes analysis.

1.4 The Poisson posterior mean problem

Example 1.1 may seem like a toy problem. Hence we now consider a more substantial decision problem, that is, the Poisson decision problem (which is perhaps one of the most famous use-cases of the empirical Bayes approach).

The simple problems we consider are as follows:

Example 1.2 (Poisson problem)**.**

A) Unknown parameter $\theta \in [0, \infty)$.

B) Observed random variable $Z \sim \text{Poisson}(\theta)$.

In his work, Robbins (1964) studied the above problem along with the decision spaces and losses in Parts 2. and 3. of Proposition 1.2.

The following is perhaps one of the most famous empirical Bayes formulas:

Proposition 1.4. *In the above situation and when $\theta \sim G$, then:*

$$\mathbb{E}_G\left[\theta \mid Z = z\right] = \frac{(z+1) f_G(z+1)}{f_G(z)}. \tag{1.11}$$

Proof. We have that

$$\begin{aligned}
\mathbb{E}_G\left[\theta \mid Z = z\right] &= \int \theta \frac{p(z \mid \theta)}{f_G(z)} dG(\theta) \\
&= \frac{1}{f_G(z)} \int \theta \frac{\exp(-\theta)\theta^z}{z!} dG(\theta) \\
&= \frac{z+1}{f_G(z)} \int \frac{\exp(-\theta)\theta^{z+1}}{(z+1)!} dG(\theta) \\
&= \frac{(z+1) f_G(z+1)}{f_G(z)}.
\end{aligned}$$

□

1.4.1 Poisson F-modeling

Here's the fascinating observation about Proposition 1.4: $f_G(z) = \mathbb{P}_G[Z = z]$ is easily estimated since it depends on the *marginal* distribution of Z, which we directly observe. For example, a straight-forward estimator is the following:

$$\hat{f}_n(z) = \frac{\#\{i : Z_i = z\}}{n}.$$

Hence we can easily estimate the posterior mean in the Poisson model. Hence we can estimate the posterior mean as follows:

$$\hat{\mathbb{E}}[\theta \mid Z = z] = \frac{(z+1)\,\#\{i : Z_i = z+1\}}{\#\{i : Z_i = z\}}. \tag{1.12}$$

The above can be used e.g., to construct decision rules that are asymptotically optimal in mean squared error. Here instead we will demonstrate asymptotic Bayes optimality for the hypothesis testing loss in Definition 1.1 (Part 3). Building upon Eq. 1.12, we have the following empirical Bayes decision rule:

$$\hat{t}_n(z) = \begin{cases} H_>, & \text{if } \frac{(z+1)\,\#\{i:Z_i=z+1\}}{\#\{i:Z_i=z\}} > \theta^* \\ H_\leq, & \text{otherwise} \end{cases}. \tag{1.13}$$

Theorem 1.1. *Suppose that $\mathbb{E}_G[\theta] < \infty$. Then:*

$$R_{n+1}(\hat{t}_n(\cdot),\ G) \to R(G) \text{ as } n \to \infty \text{ almost surely.}$$

Proof. Consider the event,

$$A = \left\{\hat{f}_n(z) \to f_G(z) \text{ as } n \to \infty \text{ for all } z \in \mathbb{N}_{\geq 0}\right\}.$$

By Glivenko Cantelli it holds that $\mathbb{P}_G[A] = 1$.

Now take $z \in \mathbb{N}_{\geq 0}$ such that $\mathbb{E}_G[\theta \mid Z = z] \neq \theta^*$. Then on the event A, one gets that:

$$\hat{t}_n(z) \to t_G(z) \text{ as } n \to \infty,$$

as well as:

$$\ell(\hat{t}_n(\cdot), \theta) \to \ell(t_G(\cdot), \theta) \text{ as } n \to \infty.$$

We also note that $\int \sup_t \ell(t,\theta) p(z \mid \theta) dG(\theta) < \infty$ since $p(z \mid \theta) \leq 1$, $\ell(t,\theta) \leq \theta + \theta^*$, and $\mathbb{E}_G[\theta] < \infty$ by assumption. Hence by dominated convergence we get that on the event A as $n \to \infty$:

$$\int \ell(\hat{t}_n(z), \theta) p(z \mid \theta) dG(\theta) \to \int \ell(t_G(z), \theta) p(z \mid \theta) dG(\theta). \tag{1.14}$$

In fact, if z is such that $\mathbb{E}_G[\theta \mid Z = z] = \theta^*$ then the above also holds since for such z it holds that:[4]

$$\int \ell(\hat{t}_n(z), \theta) p(z \mid \theta) dG(\theta) = \int \ell(t_G(z), \theta) p(z \mid \theta) dG(\theta). \tag{1.15}$$

Hence Eq. 1.14 holds for all z. To conclude we note the following:

$$\begin{aligned}\int \sup_t \int \ell(t, \theta) p(z \mid \theta) dG(\theta) d\lambda(z) &\leq \int (\theta + \theta^*) \int p(z \mid \theta) d\lambda(z) dG(\theta) \\ &= \int (\theta + \theta^*) dG(\theta) < \infty.\end{aligned}$$

Hence by one more application of dominated convergence, we deduce that on the event A as $n \to \infty$:

$$\int \int \ell(\hat{t}_n(z), \theta) p(z \mid \theta) dG(\theta) d\lambda(z) \to \int \int \ell(t_G(z), \theta) p(z \mid \theta) dG(\theta) d\lambda(z).$$

The above is equivalent to the sought-after conclusion:

$$R(\hat{t}_n(\cdot), G) \to R(t_G(\cdot), G) = R(G) \text{ almost surely.}$$

□

1.4.2 Poisson G-modeling

The construction of the empirical Bayes decision $\hat{t}_n(\cdot)$ in Eq. 1.13 is very simple and follows directly from a straightforward estimate of the marginal density $f_G(z)$.

One disadvantage is that the estimated decisions do not satisfy the natural monotonicity that the "oracle" Bayes decisions satisfy.

Exercise 1.2. In the Poisson setting it holds that:

$$\mathbb{E}_G[\theta \mid Z = z] \geq \mathbb{E}_G[\theta \mid Z = z'] \text{ for } z \geq z'.$$

[4] To see this, first observe that by Eq. 1.5,

$$\mathbb{E}_G[\ell(H_>, \theta) \mid Z = z] = \mathbb{E}_G[\ell(H_\leq, \theta) \mid Z = z],$$

when $\mathbb{E}_G[\theta \mid Z = z] = \theta^*$. Multiplying by $f_G(z)$, this is equivalent to

$$\int \ell(H_>, \theta) p(z \mid \theta) dG(\theta) = \int \ell(H_\leq, \theta) p(z \mid \theta) dG(\theta).$$

Hence, no matter what value $t_G(z) \in \{H_>, H_\leq\}$ and $\hat{t}_n(z) \in \{H_>, H_\leq\}$ take, Eq. 1.15 will hold. As a side remark, we note that when $\mathbb{E}_G[\theta \mid Z = z] = \theta^*$ holds, then the Bayes optimal decision $t_G(z)$ is not unique; we may take either $t_G(z) = H_>$ or $t_G(z) = H_\leq$.

In contrast $\hat{t}_n$ above is not necessarily monotonic, so it could be that for $Z = z$, we would claim that $\theta > \theta^*$, but then for $Z = z'$ with $z' > z$ we would claim instead $\theta \leq \theta^*$! This seems undesirable.

Second, the approach works for the specific setting we considered (Poisson observation, testing loss functions from Proposition 1.2, Part 3) and does not necesserily generalize to other settings.

An alternative is the so-called G-modeling approach. Here we first estimate $\widehat{G}$ based on $Z_1, \dots, Z_n$ and then consider the plug-in decision:

$$\hat{t}_n = t_{\widehat{G}},$$

that is we consider the Bayes decision evaluated at the estimated prior $\widehat{G}$ rather than the (unknown) true prior G.

F-modeling and G-modeling

the core idea of the empirical Bayes approach is to estimate the prior distribution either directly or indirectly using the available data, wherein the final inference is based on the posterior distribution when using this estimated prior. Bradley Efron (2014) classified empirical Bayes approaches as pursing one of two strategies: (i) *F-modeling*, which is modeling on the data scale; and (ii) G-modeling, which is modeling on the parameter scale. Under F-modeling, the resulting empirical Bayes rule usually depends on the prior indirectly via the marginal probability density function; under G-modeling, the prior distribution is estimated and then plugged into the posterior calculation. See N. Laird (1978), and Jiang and Zhang (2009) for important works in the G-modeling approach, and see Robbins (1956) Brown and Greenshtein (2009), and Bradley Efron (2011) for some F-modeling approaches. The approach we described in Section 1.4.1 is an example of an F-modeling approach, while in this section we present an example of a G-modeling approach.

Estimates $\widehat{G}$ are often furnished through nonparametric maximum likelihood (which we will treat in later lectures). Robbins (1964) instead suggested the following "minimum distance" type estimator:[5]

$$\widehat{G} \in \operatorname{argmin} \left\{ d_{\mathrm{KS}}(F_{\widetilde{G}}, \widehat{F}_n) \,:\, \widetilde{G} \text{ distribution supported on } \Theta \right\}. \tag{1.16}$$

Above, $F_{\widetilde{G}}$ is the marginal distribution of Z_i (that is, the distribution with density $f_{\widetilde{G}}(\cdot)$) and $\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(Z_i \leq t)$ is the empirical distribution of the Z_i. Finally, for two distribution functions F_1, F_2 on $\mathbb{R}$, $d_{\mathrm{KS}}(F_1, F_2)$ is the Kolmogorov-Smirnov distance:

$$d_{\mathrm{KS}}(F_1, F_2) = \sup_{z \in \mathbb{R}} |F_1(z) - F_2(z)|.$$

[5] Deely and Kruse (1968) go on to explain how this estimator may be computationally evaluated by solving a linear program after a careful discretization.

The conclusion of Theorem 1.1 also holds for the G-modeling-based estimate.

Proof. We apply first the triangle inequality and then the definition of $\widehat{G}$ as the minimizer in Eq. 1.16.

$$d_{\mathrm{KS}}(F_{\widehat{G}}, F_G) \leq d_{\mathrm{KS}}(F_{\widehat{G}}, \hat{F}_n) + d_{\mathrm{KS}}(F_G, \hat{F}_n) \leq 2d_{\mathrm{KS}}(F_G, \hat{F}_n).$$

Now, $d_{\mathrm{KS}}(F_G, \hat{F}_n) \to 0$ almost surely by the Glivenko-Cantelli theorem, and so also $d_{\mathrm{KS}}(F_{\widehat{G}}, F_G) \to 0$ almost surely.

In the Poisson case, the above implies that for any $z \in \mathbb{N}_{\geq 0}$, we have that $f_{\widehat{G}}(z) \to f_G(z)$. For example, for $z > 0$:

$$f_{\widehat{G}}(z) = F_{\widehat{G}}(z) - F_{\widehat{G}}(z-1) \to F_G(z) - F_G(z-1) = f_G(z),$$

and also $f_{\widehat{G}}(0) = F_{\widehat{G}}(0) \to F_G(0) = f_G(0)$.

The rest of the argument is identical to the proof of Theorem 1.1.

□

We conclude this chapter by showing that in fact the estimator $\widehat{G}$ is asymptotically consistent for the "true" G.

Theorem 1.2. *$\widehat{G} = \widehat{G}_n$ (defined in Eq. 1.16) converges weakly to G as $n \to \infty$ almost surely.*

Proof. Let A be the event on which $d_{\mathrm{KS}}(F_G, \widehat{F}_n) \to 0$. As we explained above, $\mathbb{P}_G[A] = 1$. Hence it suffices to argue that on the event A, $\widehat{G}_n$ converges weakly to G.

There are two steps to our proof. First, we will prove that on the event A, the sequence of probability distributions $(\widehat{G}_n)_{n\in\mathbb{N}}$ is tight. Next we will prove that the weak limit of $\widehat{G}_{n_k}$ along any subsequence n_k must be G. To see this, suppose $\widehat{G}_{n_k}$ converges weakly to $\widetilde{G}$ as $k \to \infty$. For any $z \in \mathbb{N}_{\geq 0}$, $\theta \mapsto p(z \mid \theta)$ is continuous and bounded, and so as $k \to \infty$:

$$f_{\widehat{G}_{n_k}}(z) = \int p(z \mid \theta) d\widehat{G}_{n_k}(\theta) \to \int p(z \mid \theta) d\widetilde{G}(\theta) = f_{\widetilde{G}}(z).$$

On the other hand, on the event A, we also know that:

$$f_{\widehat{G}_{n_k}}(z) \to f_G(z) \text{ as } k \to \infty.$$

The above imply that $f_G(z) = f_{\widetilde{G}}(z)$ for all $z \in \mathbb{N}_{\geq 0}$. Below we will prove the identifiability of Poisson mixtures, which means that $G = \widetilde{G}$. Our proof concludes by standard arguments for establishing weak convergence of distributions (see e.g., Billingsley (1995, chap. 25, corollary on pg. 337)).

Tightness: Suppose $(\widehat{G}_n)_{n\in\mathbb{N}}$ is not tight (on the event A). Then there exist (random) $\varepsilon > 0$ and sequences $n_k \in \mathbb{N}, M_k > 0$ with $n_k, M_k \to \infty$ as $k \to \infty$ so that:

$$\widehat{G}_{n_k}(M_k) \leq 1 - \varepsilon \text{ for all } k \in \mathbb{N}.$$

Next pick $z' \in \mathbb{N}$ such that $F_G(z') \geq 1 - \varepsilon/2$. Observe the following:

$$\begin{aligned} F_{\widehat{G}_{n_k}}(z') &= \int_0^\infty \sum_{z=0}^{z'} p(z \mid \theta) d\widehat{G}_{n_k}(\theta) \\ &\leq \widehat{G}_{n_k}(M_k) + \sup_{\theta > M_k} \left\{ \sum_{z=0}^{z'} p(z \mid \theta) \right\}. \end{aligned}$$

Now, for any z, $p(z \mid \theta) \to 0$ as $\theta \to \infty$. Hence taking $k \to \infty$ we find:

$$\limsup_{k\to\infty} F_{\widehat{G}_{n_k}}(z') \leq 1 - \varepsilon.$$

However, on the event A, we have that:

$$\lim_{k\to\infty} F_{\widehat{G}_{n_k}}(z') = F_G(z') \geq 1 - \varepsilon/2.$$

This is a contradiction. Hence $(\widehat{G}_n)_{n\in\mathbb{N}}$ is tight on A.

Identifiability: Take two distributions G_1, G_2 on $[0, \infty)$ such that:

$$f_{G_1}(z) = f_{G_2}(z) \text{ for all } z \in \mathbb{N}_{\geq 0}.$$

This means that:

$$\int \exp(-\theta)\theta^z dG_1(\theta) = \int \exp(-\theta)\theta^z dG_2(\theta) \text{ for all } z \in \mathbb{N}_{\geq 0}.$$

Write $f_{G_1}(0) = f_{G_2}(0) = c$. If $c = 1$ then it must be that $G_1 = G_2 = \delta_0$, a Dirac point mass at 0. Suppose otherwise, i.e., that $c < 1$. Then define the following measures H_1, H_2:

$$dH_1(\theta) = \frac{1}{c}\exp(-\theta)dG_1(\theta), \;\; dH_2(\theta) = \frac{1}{c}\exp(-\theta)dG_2(\theta).$$

By the above condition,

$$\int \theta^z dH_1(\theta) = \int \theta^z dH_2(\theta) \text{ for all } z \in \mathbb{N}_{\geq 0}.$$

This means that we have two distributions on $[0, \infty)$ for which all moments are identical. Also we may note that both distributions have a moment generating function close to the origin. In particular, for any $|t| < 1$, we have that:

$$\mathbb{E}_{H_1}[\exp(t\theta)] = \frac{1}{c}\int \exp(t\theta)\exp(-\theta)dG_1(\theta) \leq \frac{1}{c} < \infty,$$

and analogously for H_2. These two facts demonstrate that $H_1 = H_2$ (see e.g., Billingsley (1995, chap. 30)). But then it must also hold that $G_1 = G_2$.

□

1.5 Bibliographic remarks

The terms F-modeling and G-modeling are due to Bradley Efron (2014). The two modeling strategies for the Poisson case indeed go back to the early works of Herbert Robbins. Nevertheless, not all questions are settled, e.g., Shen and Wu (2022) provided some new results on mean squared error optimal empirical Bayes estimation in the Poisson problem and contrasted the F- and $G-$modeling approaches.

2 Illustrative applications of empirical Bayes

2.1 Actuarial statistics

2.1.1 Historical setting

Empirical Bayes ideas have been fruitful in the rate-making of automobile insurances. The foundation of these ideas was developed by actuaries in Europe in the 1960s, e.g., by Thyrion (1960), and Bichsel (1964). At least initially it appears that they were not aware of Robbins' work on empirical Bayes.

Suppose an insurance company has a portfolio with drivers $i = 1, \dots, n$. The portfolio has been constructed in a way that accounts for some covariates that are predictive of insurance claims risk. Suppose the drivers enroll into their policy in a yearly basis. After a single year, some drivers have made multiple claims, while other drivers did not make a single claim. Should the insurance increase the premium of the former, and decrease the premium of the latter? By what amount? Thyrion (1960) explains:

> "To create homogeneous rate classes, all the factors influencing the risk must theoretically be identified and their effects quantified. If this is done, the fluctuation of individual results around the average is only the accidental effect of chance [...]: there is nothing unfair about policyholders who have not had losses paying for others [...]. In general—and this is the case in Belgium—the rate classes take into account one characteristic of the vehicle's power (displacement or fiscal power), the use it is put to (tourism and business, transport for one's own account, transport of others, etc.), sometimes also one or another factor specific to the driver (profession). [...] Now the studies done [...] indicate that a high percentage of accidents are due to recklessness, [...] drunkenness, etc... in short to the driver's behavior itself. A significant factor of the risk is therefore hardly taken into account in the pricing. It is therefore not unreasonable to ask if—in the absence of something better—it would not be appropriate to try to take it into account a posteriori."

2.1.2 Formalizing the setup

Let us formalize the above following Thyrion (1960). We index time by t, with $t = 1$ marking the end of the first year of all contracts. Let $Z_i(t) \in \mathbb{N}_{\geq 0}$ denote that number of claims by

the i-th individual until time t. Furthermore, we let the claim amount of the j-th claim of driver i be equal to the random variable Ξ_{ij}. The total cost the driver incurs to the insurance company until time t is thus equal to

$$C_i(t) = \sum_{j=1}^{Z_i(t)} \Xi_{ij}.$$

To make further progress, we make the following assumptions:

1. $Z_i(\cdot)$ is a homogeneous Poisson process [1] with intensity θ_i that is specific to the i-th driver and represents the driver's accident proneness. Thus we posit that conditionally on θ_i the following holds: for any finite collection of times $(t_k)_k$ with $t_k \geq 0$ and $t_k < t_{k+1}$, $(Z_i(t_{k+1}) - Z_i(t_k))_k$ are jointly independent. Furthermore, $Z_i(t_{k+1}) - Z_i(t_k) \sim \text{Poisson}(\theta_i \cdot (t_{k+1} - t_k))$. For a two year period, the Poisson process assumption entails that:

$$Z_i(1),\ Z_i(2) - Z_i(1) \mid \theta_i \overset{\text{iid}}{\sim} \text{Poisson}(\theta_i).$$

2. The distribution of Ξ_{ij} does not depend on i or j and is independent of the claims process $Z_i(\cdot)$. Hence we write Ξ for a generic random variable with the same distribution as Ξ_{ij}.[2] Assuming also that $\mathbb{E}[\Xi] < \infty$ we get by Wald's Lemma:

$$\mathbb{E}_{\theta_i}[C_i(t)] = \mathbb{E}_{\theta_i}[Z_i(t)]\,\mathbb{E}[\Xi] = t \cdot \theta_i \cdot \mathbb{E}[\Xi].$$

Hence, under our simplifying assumptions, the expected cost $\mathbb{E}_{\theta_i}[C_i(t)]$ of the i-th driver to the insurance company is proportional to the expected number of claims $\mathbb{E}_{\theta_i}[Z_i(t)]$. The above also means that the expected cost of the i-th driver in the second year will be proportional to $\mathbb{E}_{\theta_i}[Z_i(2) - Z_i(1)] = \theta_i$. This remains true conditionally on the number of claims $Z_i(1)$ in the first year, that is, $\mathbb{E}_{\theta_i}[Z_i(2) - Z_i(1) \mid Z_i(1)] = \theta_i$.

Hence if we had started with a truly homogeneous portfolio, that is, $\theta_i = \bar{\theta}$ for all $i = 1, \ldots, n$, then it would be fair to keep everyone's premium the same no matter how many claims they made in the first year!

However, the basic supposition of Thyrion (1960) and others (e.g., Bühlmann (1964), Bichsel (1964)) is that the θ_i vary from driver to driver (because the insurance company cannot control for all possible risk factors), and these θ_i may be taken as random draws:

$$\theta_i \sim G.$$

[1] Thyrion (1960) argues that this assumption is a reasonable starting point: "This hypothesis is reasonable if the risk does not vary systematically over time. Of course, the automobile risk has seasonal, daily and even hourly peaks, and the instantaneous rate of accidents would be better represented by a periodic function of time $\theta_i(t)$. But, as only whole numbers of periods (years) of insurance are considered in practice, there is no major objection to taking θ_i constant."

[2] This assumption represents the belief that to first order it suffices to capture the frequency of claims and not their severity: it is a sensible starting point for more involved modeling.

Bichsel (1964) called G the "structural function" representing the portfolio's heterogeneity.

Since θ_i of driver i is not known, it is impossible to assess their expected number of claims in the next year. However, it will turn out that it *is* possible to assess the expected number of claims in the 2nd year among all drivers who made the same number of claims in the 1st year. We have that:

$$\mathbb{E}_G\left[Z_i(2) - Z_i(1) \mid Z_i(1)\right] = \mathbb{E}_G\left[\mathbb{E}_{\theta_i}\left[Z_i(2) - Z_i(1) \mid Z_i(1)\right] \mid Z_i(1)\right] = \mathbb{E}_G\left[\theta_i \mid Z_i(1)\right].$$

Thyrion (1960) continues saying that "as the a posteriori structure function [...] is given by the Bayes formula, the expression for $\mathbb{E}$ can be written more simply", and derives that:

$$\mathbb{E}_G\left[\theta_i \mid Z_i(1) = z\right] = (z+1)\frac{\mathbb{P}_G\left[Z_i(1) = z+1\right]}{\mathbb{P}_G\left[Z_i(1) = z\right]}.$$

In other words, Thyrion (1960) had discovered Robbins' famous formula Eq. 1.11 for the posterior mean in the Poisson problem!

There are several attractive features to assigning premiums proportionally to $\mathbb{E}_G\left[\theta_i \mid Z_i(1)\right]$ with proportionality constant $\mathbb{E}\left[\Xi\right]$.

1. The premium system is financially balanced in the following sense:

$$\mathbb{E}_G\left[\text{Premium}_i(Z_i(1))\right] = \mathbb{E}_G\left[\mathbb{E}\left[\Xi\right]\mathbb{E}_G\left[\theta_i \mid Z_i(1)\right]\right] = \mathbb{E}\left[\Xi\right]\mathbb{E}_G\left[\theta_i\right] = \mathbb{E}_G\left[C_i(2) - C_i(1)\right].$$

2. The premium is assigned in the optimal way as a function of $Z_i(1)$ according to squared error loss.

3. The premium is non-decreasing in the number of accidents (cf. Exercise 1.2).

However, the above presupposes knowledge of the "structure function" G. Both Thyrion (1960) and Bichsel (1964) realized that they could use $Z_1(1), \ldots, Z_n(1)$ in order to estimate G, and to then determine the premium for the next year. Hence they had managed to reduce the problem to precisely the Poisson empirical Bayes problem we discussed in Section 1.4.

As an aside, we note that Vernon Johns, a pioneer in empirical Bayes theory, considered the above to be one of the best applications of empirical Bayes. In a discussion (Johns 1974) of Bühlmann (1976), Vernon Johns wrote:

> "I would like to take this opportunity to interject a parenthetical remark to the effect that insurance rate making provides one of the best examples I know where the pure Bayesian approach based on subjective prior probabilities is **not** appropriate. The point here is that the actuary's subjective prior may well be substantially different from those of the insurance commissioner or the client even if they have access to similar collateral information, since their interests do not coincide. The Bayesian philosophy does not really provide for the negotiation of such differences."

2.1.3 An example: La Royale Belge

We make the above discussion concrete by reanalyzing the portfolio that Thyrion (1960) analyzed. He had access to a year of data (claims) for a portfolio of the "La Royale Belge" insurance company that covered vehicles in the category "Tourism and Business". Table 2.1 shows the full dataset.[3]

```
using Empirikos
using TypedTables
using MarkdownTables
using Optim
using MosekTools
```

```
thyrion_tbl = Thyrion.load_table()
thyrion_eb = PoissonSample.(thyrion_tbl.z)
thyrion_summary = Empirikos.MultinomialSummary(thyrion_eb, thyrion_tbl.count)
markdown_table(thyrion_tbl)
```

Table 2.1: The number of drivers (second column) who made a given number of claims throughout the year in which they were insured (first column). For example, 7840 drivers made zero claims, and 1317 drivers made a single claim.

z	count
0	7840
1	1317
2	239
3	42
4	14
5	4
6	4
7	1

We compare a few different choices for estimating the prior G:

1. We model G as a Dirac mass. In this case there is no unobserved heterogeneity. Maximum likelihood over G is identical to maximum likelihood estimation of the common Poisson rate.

2. We model G as a Gamma distribution; the conjugate prior for the Poisson likelihood. This is a parametric model with two unknown parameters.

[3]We include Julia code to reproduce the following analysis.

3. Finally we consider the nonparametric estimate Eq. 1.16.

We start by fitting the above models:

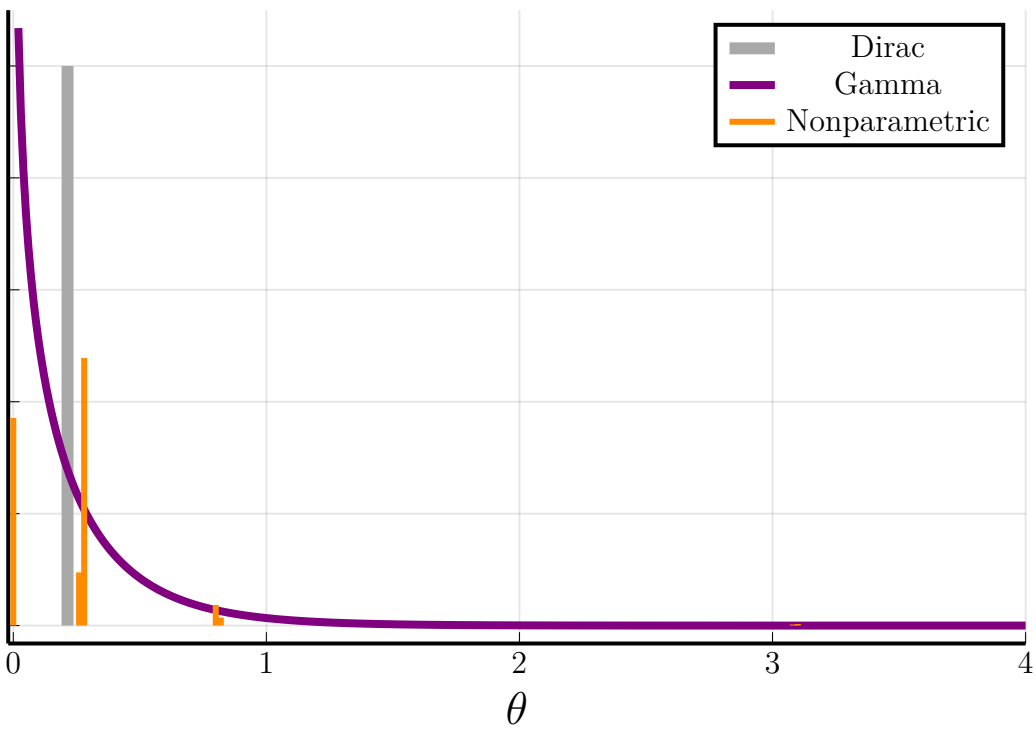
```
mean_Zs = mean(response.(thyrion_summary), weights(thyrion_summary))

fitted_dirac = Dirac(mean_Zs)

fitted_gamma = fit(
    Empirikos.ParametricMLE(model = Gamma(), solver = NewtonTrustRegion()),
    thyrion_summary,
)

ks_nonparametric = fit(
    Empirikos.KolmogorovSmirnovMinimumDistance(DiscretePriorClass(0:0.02:8), Mosek.Optimizer),
    thyrion_summary,
)
fitted_ks = Empirikos.clean(ks_nonparametric.prior)
```

We plot the estimated priors:



Above it is difficult to see all masses of the nonparametric $\widehat{G}$. The table below shows all the masses and corresponding probabilities of $\widehat{G}$:

```
markdown_table(
    Table(
        loc = round.(support(fitted_ks), digits=3),
                     prob = round.(probs(fitted_ks), digits=3)
        )
)
```

Table 2.2: Nonparametric prior $\hat{G}$ estimated by the minimum distance method. $\hat{G}$ is a finite discrete distribution with point mass locations.

loc	prob
0.0	0.371
0.26	0.095
0.28	0.478
0.8	0.037
0.82	0.014
3.08	0.002
3.1	0.003

Let us look at how well our models fit with the observed counts. We see that the model of no heterogeneity provides a very subpar fit. Assuming that G is a Gamma distribution improves the fit, and the nonparametric model improves the fit further.

```
n = nobs(thyrion_summary)

markdown_table(
    Table(
    Z = thyrion_tbl.z,
    Empirical = thyrion_tbl.count,
    Dirac = round.(pdf.(fitted_dirac, thyrion_eb) .* n, digits=1),
    Gamma = round.(pdf.(fitted_gamma, thyrion_eb) .* n, digits=1),
    Nonparametric = round.(pdf.(fitted_ks, thyrion_eb) .* n, digits=1)
    )
)
```

Z	Empirical	Dirac	Gamma	Nonparametric
0	7840	7635.6	7847.0	7839.3
1	1317	1636.7	1288.4	1318.5
2	239	175.4	256.5	237.5
3	42	12.5	54.1	43.5
4	14	0.7	11.7	12.5
5	4	0.0	2.6	5.5
6	4	0.0	0.6	2.5
7	1	0.0	0.1	1.1

Finally let us look at the estimated posterior means.

```
robbins_postmeans = ((0:7).+1) ./ thyrion_tbl.count .* [thyrion_tbl.count[2:end]; 0]
```

```
postmeans = PosteriorMean.(PoissonSample.(0:7))
markdown_table(
    Table(
        Estimand = postmeans,
        Dirac = round.(postmeans(fitted_dirac), digits=2),
        Gamma = round.(postmeans(fitted_gamma), digits=2),
        NonparametricG = round.(postmeans(fitted_ks), digits=2),
        Robbins = round.(robbins_postmeans, digits=2)
    )
)
```

Table 2.4: Posterior means

Estimand	Dirac	Gamma	NonparametricG	Robbins
$\mathbb{E}[\mu \mid \mathcal{P}oi(0;\ \mu)]$	0.21	0.16	0.17	0.17
$\mathbb{E}[\mu \mid \mathcal{P}oi(1;\ \mu)]$	0.21	0.4	0.36	0.36
$\mathbb{E}[\mu \mid \mathcal{P}oi(2;\ \mu)]$	0.21	0.63	0.55	0.53
$\mathbb{E}[\mu \mid \mathcal{P}oi(3;\ \mu)]$	0.21	0.87	1.16	1.33
$\mathbb{E}[\mu \mid \mathcal{P}oi(4;\ \mu)]$	0.21	1.1	2.17	1.43
$\mathbb{E}[\mu \mid \mathcal{P}oi(5;\ \mu)]$	0.21	1.33	2.81	6.0
$\mathbb{E}[\mu \mid \mathcal{P}oi(6;\ \mu)]$	0.21	1.57	3.02	1.75
$\mathbb{E}[\mu \mid \mathcal{P}oi(7;\ \mu)]$	0.21	1.8	3.07	0.0

Observe that Robbins' F-modeling estimator can behave quite eratically for large values of Z.

2.2 The missing species problem

Our next applications also pertains to an ingenious empirical Bayes solution to a seemingly impossible problem.

During World War II, Alexander Corbet, a renowned naturalist, spent two years in Malaysia (then called Malaya) trapping butterflies. Throughout his time, he captured 620 species of butterfly. For 118 of these species, he had captured a single specimen, while 74 species had been captured twice, 44 species had been captured three times, and so on.

```
butterfly_tbl = Butterfly.load_table()
markdown_table(butterfly_tbl)
```

x	y
1	118
2	74
3	44
4	24
5	29
6	22
7	20
8	19
9	20
10	15
11	12
12	14
13	6
14	12
15	6
16	9
17	9
18	6
19	10
20	10
21	11
22	5
23	3
24	3

In addition to the above 501 species, Corbet captured another 119 species, with 25 specimen or more for each.

His seemingly impossible question was the following: how many new species that he had never encountered before could he expect to catch if he returned to Malaysia for one more year?

He posed the question to R.A. Fisher and the latter came up with an extraordinary empirical Bayes solution (Fisher, Corbet, and Williams 1943). Fisher posited that in total there are S species of butterfly in Malaysia. Let Z_i be the number of butterflies of species i that Corbet captured. Fisher then assumed the following:

$$Z_i \sim \text{Poisson}(\theta_i),$$

for some $\theta_i > 0$. θ_i represents how abundant a species is, how easy it is to be captured and so forth. Furthermore, Fisher assumed that:

$$\theta_i \sim G,$$

where he took G to be a Gamma distribution with unknown parameters.[4] Let us figure out how many new butterfly species Corbet could expect under the above hierarchical model. Fix the species i. The probability that Corbet did not capture any butterfly of species i in his two years at Malaysia, but would capture such a butterfly after one more year, is equal to $\mathbb{E}_G\left[\exp(-\theta)\left(1-\exp(-\theta/2)\right)\right]$.[5] Hence the expected number of new species in one more year is equal to:

$$\sum_{i=1}^{S} \mathbb{E}_G\left[\exp(-\theta_i)\left(1-\exp(-\theta_i/2)\right)\right] = S \cdot \mathbb{E}_G\left[\exp(-\theta)\left(1-\exp(-\theta/2)\right)\right]. \tag{2.1}$$

In an typical empirical Bayes fashion, one could hope to estimate G based on the data from the first two years, and then to estimate the quantity above. However, there is one more challenge: in fact, not even S, the total number of species is known! Corbet only knew the number of species that he observed one or more times, that is:

$$\hat{S}_{\geq 1} = \sum_{i=1}^{S} \mathbf{1}(Z_i \geq 1).$$

The expectation of the quantity above is equal to:

$$\mathbb{E}_G\left[\hat{S}_{\geq 1}\right] = S \cdot \mathbb{E}_G\left[\left(1-\exp(-\theta)\right)\right].$$

Hence we can rewrite the expected number of new species in one more year Eq. 2.1 as:

$$\mathbb{E}_G\left[\hat{S}_{\geq 1}\right] \cdot \frac{\mathbb{E}_G\left[\exp(-\theta)\left(1-\exp(-\theta/2)\right)\right]}{\mathbb{E}_G\left[\left(1-\exp(-\theta)\right)\right]}.$$

Hence if we have access to an estimate $\widehat{G}$ of G, then we could estimate the expected number of new species Corbet would trap by the following:

$$\hat{S}_{\geq 1} \cdot \frac{\mathbb{E}_{\widehat{G}}\left[\exp(-\theta)\left(1-\exp(-\theta/2)\right)\right]}{\mathbb{E}_{\widehat{G}}\left[\left(1-\exp(-\theta)\right)\right]}.$$

[4] Good (1992) notes that Fisher did not use the term "prior" for G "in case anyone thought he'd become a covert Bayesian."

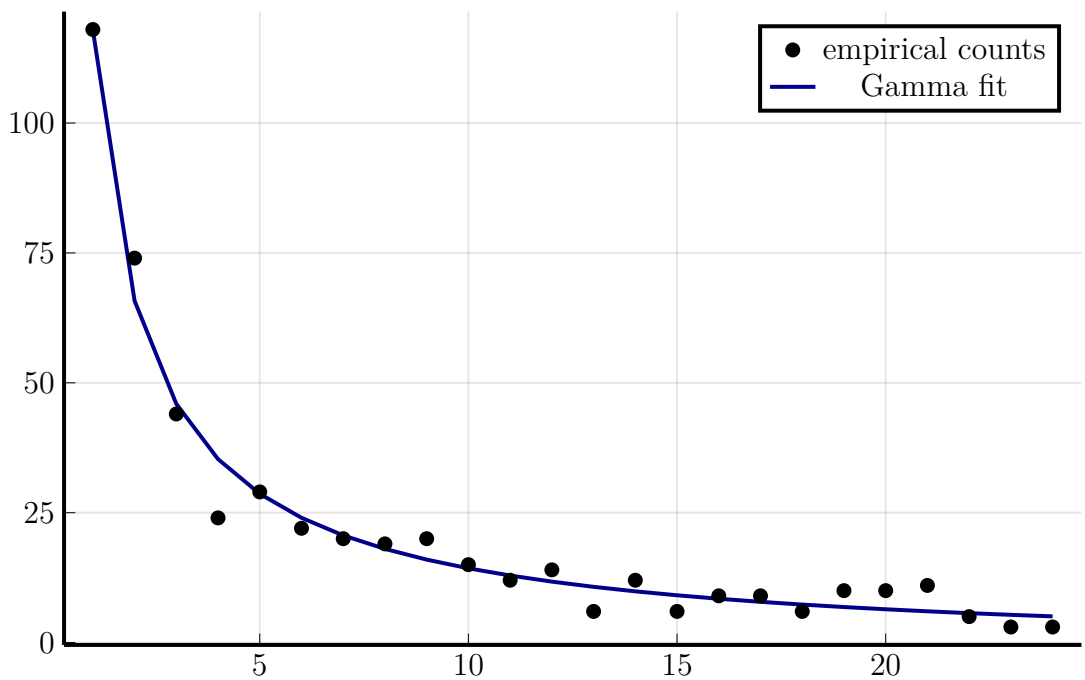
[5] Here we made a Poisson process assumption, similar to the one we made in the actuarial application of Section 2.1.

2.2.1 A parametric approach

Below we carry out this computation by making the parametric assumption that G is a Gamma distribution with unknown parameters and then estimating G by maximum likelihood. [6]

```
butterfly_Zs = [
    Empirikos.MarginallyTruncatedPoissonSample.(butterfly_tbl.x)
    Empirikos.MarginallyTruncatedPoissonSample(Interval(25, nothing))
]
butterfly_summary = Empirikos.MultinomialSummary(butterfly_Zs, [butterfly_tbl.y; 119])

gamma_butterfly = fit(
    Empirikos.ParametricMLE(model = Gamma(), solver = NewtonTrustRegion()),
    butterfly_summary,
)
```



We can now estimate the number of new species that Corbet would encounter after one more year:

[6] One needs to pay attention here and account for the fact that we do *not* observe species with $Z_i = 0$ and also that we do not know the exact value of Z_i for species with $Z_i \geq 25$, but only that Z_i is at least as large as 25. Hence we are facing a problem with left-truncation and right-censoring.

```
expected_new_species =
    nobs(butterfly_summary) *
    (mgf(gamma_butterfly, -1) - mgf(gamma_butterfly, -1 - 1 / 2)) /
    (1 - mgf(gamma_butterfly, -1))
expected_new_species
```

```
46.67776925110288
```

Hence our estimate is that Corbet would catch about 46.68 new species if butterfly.

2.2.2 An F-modeling approach due to Good, Toulmin, and Turing

Good and Toulmin (1956) proposed a nonparametric approach to answering the above question; Good attributed the core ideas of what follows to Alan Turing. The starting point is to revisit Eq. 2.1. Expanding $1 - \exp(-\theta/2)$ in its Taylor series, we get:

$$S \cdot \mathbb{E}_G \left[\exp(-\theta)\left(1 - \exp(-\theta/2)\right)\right] = S \cdot \mathbb{E}_G \left[\sum_{j=1}^{\infty} \exp(-\theta) \frac{(-1)^{j-1}\theta^j}{2^j j!}\right].$$

Also note that:

$$\mathbb{E}_G \left[\# \left\{i : Z_i = j\right\}\right] = \sum_{i=1}^{S} \mathbb{P}_G \left[Z_i = j\right] = S \cdot \mathbb{E}_G \left[\exp(-\theta) \frac{\theta^j}{j!}\right].$$

Hence by Tonelli-Fubini and by matching terms:

$$S \cdot \mathbb{E}_G \left[\exp(-\theta)\left(1 - \exp(-\theta/2)\right)\right] = \sum_{j=1}^{\infty} \mathbb{E}_G \left[\# \left\{i : Z_i = j\right\}\right] \frac{(-1)^{j-1}}{2^j}.$$

The punchline is the following: just as we could estimate Robbins' formula in Eq. 1.12 by plugging in the observed counts, we can do the same above! We get the following nonparametric estimator

$$\sum_{j=1}^{\infty} \# \left\{i : Z_i = j\right\} \frac{(-1)^{j-1}}{2^j},$$

which of course is just a finite sum. Let us compute this!

```
good_toulmin_turing = sum(-(-0.5).^(1:24) .* butterfly_tbl.y )
good_toulmin_turing
```

```
45.17149204015732
```

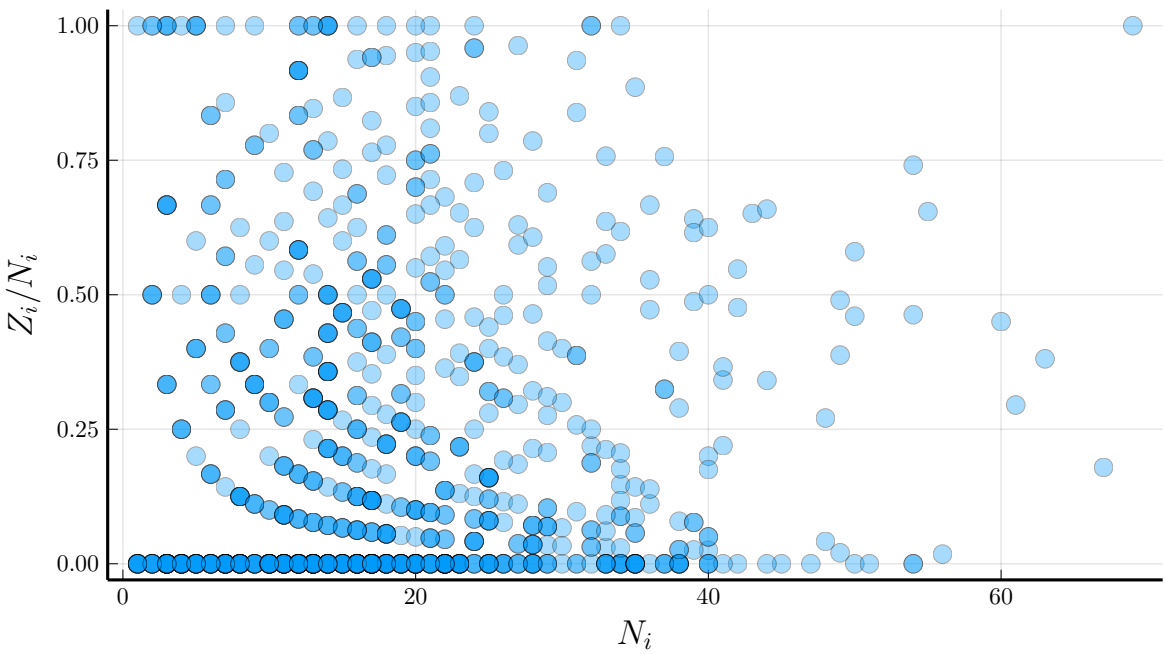
In this case, the nonparametric estimate 45.17 is very close to the parametric estimate 46.68 that we derived above.

2.3 Intestinal surgery dataset

Gholami et al. (2015) studied the problem of staging gastric adenocarcinoma in patients who had undergone intestinal surgery. The following table (Bradley Efron and Hastie 2016, chap. 6.3) contains data on 844 patients. For each patient, the data consists of N_i, the total number of surgically removed lymph nodes, as well as Z_i, the number of lymph nodes that were positive.

Below we provide a scatterplot of the full dataset:

```
surgery_samples = Surgery.ebayes_samples()
surgery_props = response.(surgery_samples) ./ ntrials.(surgery_samples)
```



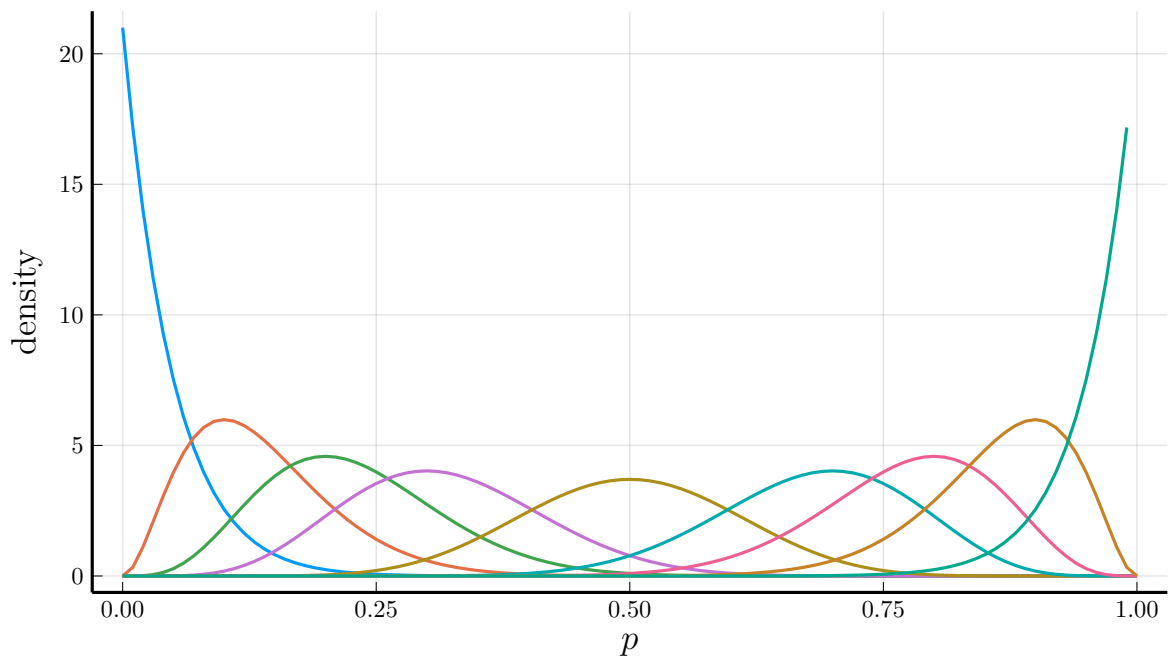
We model the data as follows:[7]

$$p_i \sim G,\ Z_i \mid p_i, N_i \sim \text{Binomial}(p_i, N_i)$$

We will fit a nonparametric class of smooth priors to our dataset by nonparametric maximum likelihood (which we will describe in a later class). The class of smooth priors we consider is a mixture of smooth Beta densities, of which we show plot some components below:

[7] In contrast to the empirical Bayes models we have been looking at so far, here the likelihood is different for different units since it is a function of N_i. The empirical Bayes framework can handle such heterogeneous settings as well.

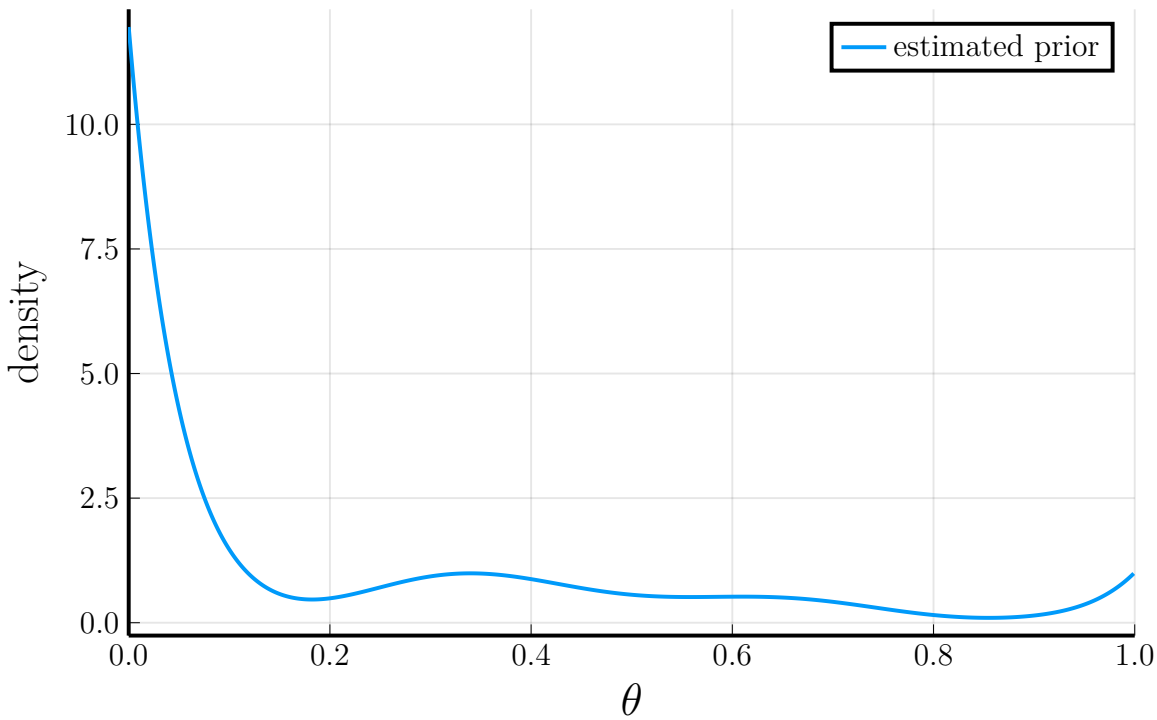
```
smooth_beta_class = Empirikos.auto_convexclass(
    Empirikos.BetaMixtureClass(),
    0.05, 0:0.01:1
)
```



Hence we seek to capitalize on what Efron calls the "bet on smoothness" principle. [8] The density of the estimated prior is shown next:

```
fitted_betamix = fit(NPMLE(smooth_beta_class, Mosek.Optimizer), surgery_samples)
```

[8]The class of priors considered in Bradley Efron and Hastie (2016) to "bet on smoothness" is instead an exponential family with a fourth degree polynomial as the sufficient statistic. This is a different class of priors than the one we considered above; however the conclusions are similar.



We may answer questions as the following: what is the chance that $p_i \leq 0.1$? As we can see, the (estimated) answer is over 50%.

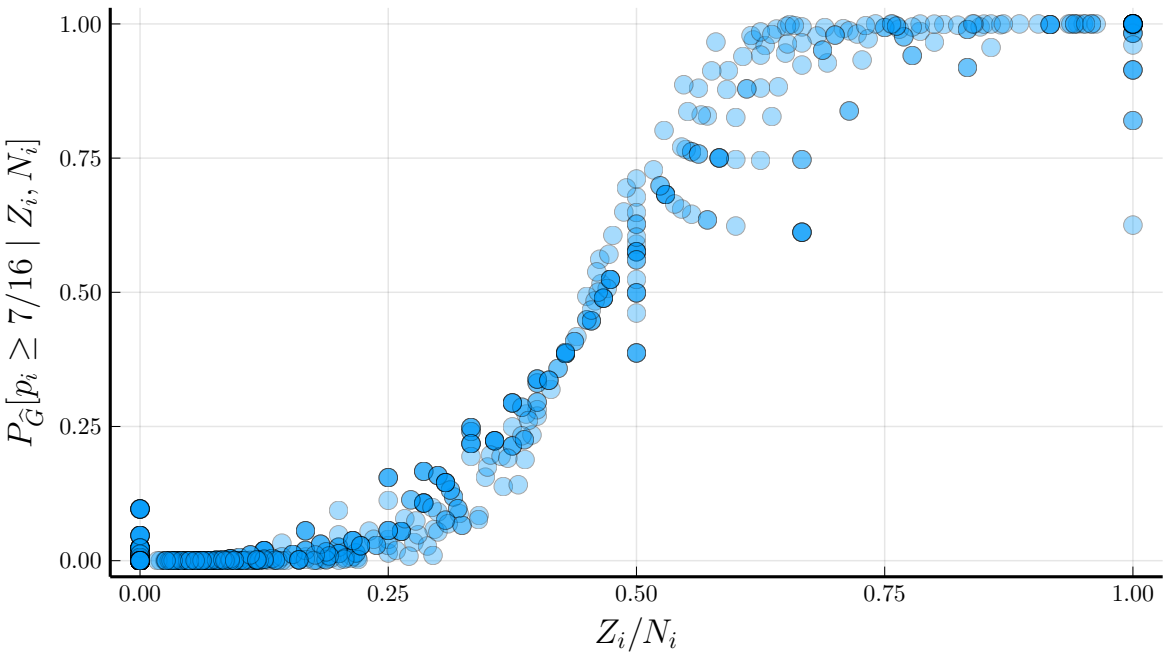
```
round(cdf(fitted_betamix.prior, 0.1), digits = 2)
```

```
0.51
```

Furthermore, Gholami et al. (2015) were interested in figuring out which patients have $p_i \geq 7/16$. Our empirical Bayes approach allows us to compute the posterior probability that $p_i \geq 7/16$ for each patient.

```
post_prob_estimands =
    Empirikos.PosteriorProbability.(surgery_samples, Interval(7 / 16, 1.0))

estimated_post_probs = post_prob_estimands(fitted_betamix.prior)
```



2.4 Bibliographic remarks

We refer the reader to the following excellent resources that provide an overview of further applications of empirical Bayes: Bradley Efron (2010) and Bradley Efron and Hastie (2016, chap. 6) provide illuminating descriptions of several of the applications we discussed above. Koenker and Gu (2017) and Narasimhan and Efron (2020) describe empirical Bayes applications alongside reproducible software code. Bühlmann (2005) is a textbook that describes the application of empirical Bayes ideas (and more) to actuarial problems.

3 The James-Stein estimator and Empirical Bayes

3.1 Introduction

Consider the following simple model where we observe

$$X_1, \dots, X_m \overset{iid}{\sim} N(\theta, \sigma^2), \qquad \text{where} \quad m \geq 1, \tag{3.1}$$

and we are interested in estimating the *unknown* $\theta \in \mathbb{R}$, and σ^2 is assumed *known*. We know that in this model, $\bar{X}_m := \frac{1}{m}\sum_{i=1}^m X_i$ is the "best" estimator for θ—we know that $\bar{X}_m$ has a number of appealing properties, some of which we list below:

1. it is complete *sufficient* for θ;
2. it is the *uniformly minimum variance unbiased estimator* (UMVUE) of θ and it attains the Cramér-Rao lower bound;
3. it is the maximum likelihood estimator (MLE) of θ;
4. it is *minimax*[1] optimal;
5. it is *admissible*[2].

Further, if we measure the quality of any estimator $\widehat{\theta}$ of θ by using its squared-error risk, by Rao-Blackwellization (or the sufficiency principle), it is enough to consider estimators that are just functions of $\bar{X}_m$. Thus, in essence, we have reduced the problem from m data points to just one by considering $Z := \bar{X}_m \sim N(\theta, \frac{\sigma^2}{m})$.

Now, instead of estimating one parameter θ, consider estimating $\theta_1, \dots, \theta_n \in \mathbb{R}$ from *independent* observations:

$$Z_i \sim N(\theta_i, 1), \qquad \text{for } i = 1, \dots, n, \tag{3.2}$$

[1] A proof of the minimaxity of $\bar{X}_m$ can be found in Keener (2010a, chap. 16.6).

[2] Consider the decision theoretic framework of Chapter 1. An estimator $\delta(Z)$ (where Z is the observed data) is said to be *inadmissible* if there exists another estimator $\delta'(Z)$ such that $R(\delta'(\cdot), \theta) \leq R(\delta(\cdot), \theta)$ for all $\theta \in \Theta$, and $R(\delta'(\cdot), \theta) < R(\delta(\cdot), \theta)$ for at least one θ. In this case $\delta'(\cdot)$ is a "better" estimator than $\delta(\cdot)$ (i.e., $\delta'(\cdot)$ beats $\delta(\cdot)$) and hence $\delta(\cdot)$ is *inadmissible*. An estimator $\delta'(\cdot)$ is said to be *admissible* if there exits no estimator that beats it (for all values of $\theta \in \Theta$). See Keener (2010a, chap. 11.3) for a proof of the admissibility of $\bar{X}_m$. This result is known from the works of Hodges and Lehmann (1951), Girshick and Savage (1951), and Blyth (1951).

Here, for notational simplicity, we assume that the known variance is 1; the exact same techniques would work for any known $\sigma^2 > 0$.

Succinctly, the above model can be written as $\mathbf{Z} \sim N(\boldsymbol{\theta}, I_n)$ where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$. Let $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_n)$ be any estimator of $\boldsymbol{\theta}$. A natural and intuitive estimator of $\boldsymbol{\theta}$ is $\mathbf{Z}$ — it can be shown that $\mathbf{Z}$ satisfies properties 1-4 above.

Nonetheless, it turns out that $\mathbf{Z}$ is *not admissible* for $n \geq 3$, although it is admissible for $n = 1, 2$[3,4]. C. Stein (1956) demonstrated that $\mathbf{Z}$ could be improved everywhere, but a specific "practical" form of such an improved estimator was developed in the celebrated paper James and Stein (1961).[5]

Definition 3.1 (James-Stein estimator (James and Stein 1961))**.** For $n > 2$, James and Stein (1961) proposed the following estimator

$$\hat{\boldsymbol{\theta}}_{JS} := \left(1 - \frac{n-2}{\|\mathbf{Z}\|^2}\right)\mathbf{Z}, \qquad \text{where} \qquad \|\mathbf{Z}\|^2 = \sum_{i=1}^{n} Z_i^2, \tag{3.3}$$

which is called the *James-Stein estimator*.

James and Stein (1961) showed that $\hat{\boldsymbol{\theta}}_{JS}$ is a "better" estimator than the MLE $\mathbf{Z}$.

Theorem 3.1 (James and Stein (1961))**.** *For $n \geq 3$, $\hat{\boldsymbol{\theta}}_{JS}$ dominates the MLE $\mathbf{Z}$ everywhere in terms of squared-error risk, i.e.,*

$$\mathbb{E}_{\boldsymbol{\theta}}[\|\hat{\boldsymbol{\theta}}_{JS} - \boldsymbol{\theta}\|^2] < \mathbb{E}_{\boldsymbol{\theta}}[\|\mathbf{Z} - \boldsymbol{\theta}\|^2], \qquad \textit{for all } \boldsymbol{\theta} \in \mathbb{R}^n.$$

This result is perhaps surprising because $\hat{\boldsymbol{\theta}}_{JS}$ combines information from independent observations that seemingly have nothing to do with each other. The James-Stein estimator $\hat{\boldsymbol{\theta}}_{JS}$ improves overall estimation accuracy, although it may not necessarily improve accuracy for every single θ_i. It is also important to remark that $\hat{\boldsymbol{\theta}}_{JS}$ is not admissible either.

The idea at the heart of Stein's proposal, namely that of employing *shrinkage to reduce variance, at the expense of introducing bias*, turns out to be a very powerful one that has had a huge impact on statistical methodology. This is also the key idea in nonparametric function estimation and many modern statistical models, that essentially involve estimating many parameters.

Below, we review two intuitive interpretations of the James-Stein estimator. There is also an empirical Bayes interpretation, which will be discussed in the next section.

[3] See Keener (2010a, chap. 11.3) for a proof of the admissibility of $\mathbf{Z}$ when $n = 1$. C. Stein (1956) showed the admissibility of $\mathbf{Z}$ when $p = 2$.

[4] In fact, it can be shown that $\mathbf{Z}$ is the minimum risk equivariant estimator in this problem; hence any admissible estimator for $n \geq 3$ involves an arbitrary choice.

[5] "That sensational paper had statisticians wondering and asking who W. James was." Read Everson (2007) to find out—this article also includes an account by Carl Morris.

Remark (Intuition for shrinkage — Stein's motivation (C. Stein 1956)). In Stein's original work in 1956, he argued that a good estimator $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ should obey $\hat{\theta}_i \approx \theta_i$, and therefore $\|\hat{\boldsymbol{\theta}}\|^2 \approx \|\boldsymbol{\theta}\|^2$. However, this is not true of the MLE $\mathbf{Z}$. As the coordinates of $\mathbf{Z}$ are independent, informally, we should expect $\|\mathbf{Z}\|^2$ to concentrate around $n + \|\boldsymbol{\theta}\|^2$ for large n, as

$$\mathbb{E}_{\boldsymbol{\theta}}[\|\mathbf{Z}\|^2] = \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}}[Z_i^2] = \sum_{i=1}^{n} (\theta_i^2 + 1) = n + \|\boldsymbol{\theta}\|^2.$$

Since the norm of $\|\mathbf{Z}\|^2$ is too large on average, a natural solution is to "shrink" $\|\mathbf{Z}\|^2$ towards zero, which is exactly what $\hat{\boldsymbol{\theta}}_{JS}$ does. Indeed, in James and Stein (1961), the authors consider a family of estimators indexed by c:

$$\hat{\boldsymbol{\theta}}_c := \left(1 - \frac{c}{\|\mathbf{Z}\|^2}\right) \mathbf{Z}.$$

They showed that for all $c \in (0, 2(n-2))$, $R(\hat{\boldsymbol{\theta}}_c, \boldsymbol{\theta}) < R(\mathbf{Z}, \boldsymbol{\theta})$ holds uniformly.

Remark (Another heuristic to justify "shrinkage"). Another way to interpret Stein's phenomenon is by thinking about the *winner's curse*. In particular, consider the order statistics $Z_{(1)}, \dots, Z_{(n)}$ and the order statistics $\theta_{(1)}, \dots, \theta_{(n)}$ of $\boldsymbol{\theta}$. Jensen's inequality tells us that on average, the largest coordinate of $\mathbf{Z}$ is *larger* than the largest coordinate of $\boldsymbol{\theta}$:

$$\mathbb{E}[Z_{(n)}] = \mathbb{E}[\max_{i=1,\dots,n} Z_i] \geq \max_{i=1,\dots,n} \mathbb{E}[Z_i] = \theta_{(n)},$$

and similarly, $\mathbb{E}[Z_{(1)}] < \theta_{(1)}$. As a result, it might make sense to "shrink" the order statistics (or equivalently $\mathbf{Z}$) towards the sample mean $\bar{Z} := \frac{1}{n}\sum_{i=1}^{n} Z_i$.

Note that the James-Stein estimator $\hat{\boldsymbol{\theta}}_{JS}$ is a nonlinear, biased estimator; it is not immediate how one can compute the (frequentist) risk of $\hat{\boldsymbol{\theta}}_{JS}$ to prove Theorem 1. This is the main goal of this chapter. In particular, we discuss the following: in Section 3.2 we provide an empirical Bayes interpretation for $\hat{\boldsymbol{\theta}}_{JS}$; in Section 3.3.2 we prove Theorem 3.1, after developing, in Section 3.3, a very general technique to compute the risk of any 'smooth' estimator of $\boldsymbol{\theta}$.

3.2 Empirical Bayes and the James-Stein estimator

3.2.1 The Bayes estimator under a normal prior

Let us consider the following Bayesian formulation in which the unknown parameters are taken to be random variables: suppose that

$$\theta_1, \dots, \theta_n \overset{iid}{\sim} N(0, \tau^2), \qquad \text{and} \qquad Z_i \mid \theta_i \overset{ind}{\sim} N(\theta_i, 1), \quad \text{for } i = 1, \dots, n. \tag{3.4}$$

Thus, the joint density of $(\mathbf{Z}, \boldsymbol{\theta})$ at $(z_1, \dots, z_n, \theta_1, \dots, \theta_n)$ is given by

$$\frac{1}{(2\pi\tau)^n} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(z_i - \theta_i)^2 - \frac{1}{2\tau^2}\sum_{i=1}^{n}\theta_i^2\right].$$

Consider the generalization of the above model where $Z_i \mid \theta_i \sim N(\theta_i, \sigma^2)$ are drawn independently. Show that the marginal distribution of $\mathbf{Z}$ is a product distribution, where each Z_i is marginally distributed as $N(0, \sigma^2 + \tau^2)$. Further, show that the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{Z} = \mathbf{Z}$, where $\mathbf{Z} = (z_1, \dots, z_n)$, is also a product measure where each $\theta_i \mid \mathbf{Z} = \mathbf{Z} \overset{d}{\equiv} \theta_i \mid Z_i = z_i \sim N\left(\frac{\nu}{\sigma^2} z_i, \nu\right)$; here ν is such that $\frac{1}{\nu} = \frac{1}{\tau^2} + \frac{1}{\sigma^2} = \frac{\tau^2+\sigma^2}{\tau^2\sigma^2}$.

It can be shown that the Bayes estimator (that minimizes the Bayes risk under the squared-error loss) — which is simply the mean of the posterior distribution — is given by

$$\hat{\boldsymbol{\theta}}_B := \left(1 - \frac{1}{1+\tau^2}\right)\mathbf{Z}. \tag{3.5}$$

Thus, the Bayes estimator $\hat{\boldsymbol{\theta}}_B$ shrinks the MLE $\mathbf{Z}$ towards $0 \in \mathbb{R}^n$, the prior mean. As the Z_i's are independent, each Z_i does not contain any information about θ_j, for $j \neq i$, but it DOES contain a lot of information about the parameters of the prior of the θ_i's (i.e., τ^2 in this case).

The following result gives the exact expression of the Bayes risk of $\hat{\boldsymbol{\theta}}_B$.

Assume Eq. 3.2 and Eq. 3.4. Then, the Bayes risk of $\hat{\boldsymbol{\theta}}_B$ is

$$\min_{t(\cdot)} R(t, N(0, \tau^2)) \equiv R_{\hat{\boldsymbol{\theta}}_B} = \mathbb{E}[\|\hat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}\|^2] = \frac{n\,\tau^2}{\tau^2+1} \equiv R_{\mathbf{Z}}\frac{\tau^2}{\tau^2+1},$$

where $R_{\mathbf{Z}} \equiv n$ denotes the risk of the MLE $\mathbf{Z}$.

Proof. Let us first rewrite the difference between the estimator $\hat{\boldsymbol{\theta}}_B$ and the parameter $\boldsymbol{\theta}$ as

$$\hat{\boldsymbol{\theta}}_B - \boldsymbol{\theta} = (1-\rho)(\mathbf{Z} - \boldsymbol{\theta}) - \rho\boldsymbol{\theta},$$

where $\rho = \frac{1}{1+\tau^2}$. Then, the (frequentist) risk for a fixed value of $\boldsymbol{\theta}$ is

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\|\hat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}\|^2 &= (1-\rho)^2\mathbb{E}_{\boldsymbol{\theta}}\|\mathbf{Z} - \boldsymbol{\theta}\|^2 + \rho^2\|\boldsymbol{\theta}\|^2 - 2\rho(1-\rho)\mathbb{E}_{\boldsymbol{\theta}}[(\mathbf{Z} - \boldsymbol{\theta})\boldsymbol{\theta}] \\ &= (1-\rho)^2 n + \rho^2\|\boldsymbol{\theta}\|^2.\end{aligned}$$

Taking an outer expectation and integrating over $\boldsymbol{\theta}$, we get the desired result:

$$R_{\hat{\boldsymbol{\theta}}_B} = \mathbb{E}[\|\hat{\boldsymbol{\theta}}_B - \boldsymbol{\theta}\|^2] = n(1-\rho)^2 + n\rho^2\tau^2 = n(1-\rho),$$

as $(1-\rho)^2 + \rho^2\tau^2 = 1 - 2\rho + \rho^2(1+\tau^2) = 1 - 2\rho + \rho$.

□

Clearly, $R_{\hat{\boldsymbol{\theta}}_B} < R_{\mathbf{Z}}$ always. If $\tau^2 = 1$, then the Bayes estimator has half the risk as the MLE $\mathbf{Z}$.

3.2.2 Empirical Bayes interpretation of $\hat{\boldsymbol{\theta}}_{JS}$

As we have seen before, empirical Bayes arguments combine frequentist and Bayesian elements in analyzing problems of repeated structure. In the Bayesian approach to this problem above, the choice of τ^2 is crucial (as it controls the amount of shrinkage). In an empirical Bayes approach to estimation, the data are used to estimate parameters of the prior distribution, which can then be used to approximate the Bayes estimator.

To do this in the current setting, recall that under the Bayesian model, $Z_1, \dots, Z_n$ are i.i.d. from $N(0, 1+\tau^2)$, and thus

$$\|\mathbf{Z}\|^2 = \sum_{i=1}^{n} Z_i^2 \sim (1+\tau^2)\chi_n^2,$$

where χ_n^2 is the chi-square distribution with n degrees of freedom. To approximate the Bayes estimator $\hat{\boldsymbol{\theta}}_B$ in Eq. 3.5 we need to 'estimate' $(1+\tau^2)^{-1}$.

Note that the UMVUE of $1+\tau^2$ is $\|\mathbf{Z}\|^2/n$. Thus, we may want to use $\|\mathbf{Z}\|^2$ to estimate $(1+\tau^2)^{-1}$. In fact, the following can be shown:

$$\mathbb{E}\left[\frac{n-2}{\|\mathbf{Z}\|^2}\right] = \frac{1}{1+\tau^2}.$$

Exercise 3.1. Prove the above!

Thus, an unbiased estimator for the shrinkage factor $(1+\tau^2)^{-1}$ is $\frac{n-2}{\|\mathbf{Z}\|^2}$ which when substituted in Eq. 3.5 yields the James-Stein estimator $\hat{\boldsymbol{\theta}}_{JS}$ in Eq. 3.3.

Moreover, we can derive the Bayes risk for $\hat{\boldsymbol{\theta}}_{JS}$ in this setting.

Theorem 3.2. *Assume that Eq. 3.4 holds. Then,*

$$R_{\hat{\boldsymbol{\theta}}_{JS}} \equiv R(\hat{\boldsymbol{\theta}}_{JS}, N(0,\tau^2)) = \frac{n\tau^2}{1+\tau^2} + \frac{2}{1+\tau^2}.$$

Of course this is bigger than the true Bayes risk $\frac{n\tau^2}{1+\tau^2}$, but the penalty is surprisingly modest,

$$\frac{R_{\hat{\boldsymbol{\theta}}_{JS}}}{R_{\hat{\boldsymbol{\theta}}_B}} = 1 + \frac{2}{n\tau^2}.$$

The shock the James-Stein estimator provided the statistical world didn't come from the above display. Note that the calculations above are based on the zero-centric Bayesian model Eq. 3.4, where the maximum likelihood estimator $\mathbf{Z}$, which doesn't favor values of $\boldsymbol{\theta}$ near $\mathbf{0}$, might be expected to be bested. The rude surprise came from Theorem 3.1 proved by James and Stein (1961), which we proceed to prove now.

3.3 Stein's identity and Stein's unbiased risk estimator

To prove Theorem 3.1, we will first review Stein's identity. Along the way, we will review *Stein's unbiased risk estimator* (SURE), which is of independent interest; see C. M. Stein (1981).

Consider the setting of Eq. 3.2 where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ is unknown (and fixed). Given any arbitrary estimator $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_n) =: h(\mathbf{Z})$, for some function $h : \mathbb{R}^n \to \mathbb{R}^n$, of $\boldsymbol{\theta}$, we can write

$$\begin{aligned}\|\mathbf{Z} - \hat{\boldsymbol{\theta}}\|^2 &= \|(\mathbf{Z} - \boldsymbol{\theta}) - (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})\|^2 \\ &= \|\mathbf{Z} - \boldsymbol{\theta}\|^2 + \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 - 2\,(\mathbf{Z} - \boldsymbol{\theta})^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}),\end{aligned}$$

which implies that (using the fact that $\mathbb{E}_{\boldsymbol{\theta}}\|\mathbf{Z} - \boldsymbol{\theta}\|^2 = n$)

$$\begin{aligned}R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \equiv \mathbb{E}_{\boldsymbol{\theta}}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 &= \mathbb{E}_{\boldsymbol{\theta}}\|\mathbf{Z} - \hat{\boldsymbol{\theta}}\|^2 - n + 2\,\mathbb{E}_{\boldsymbol{\theta}}[(\mathbf{Z} - \boldsymbol{\theta})^\top (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})] \\ &= \mathbb{E}_{\boldsymbol{\theta}}\|\mathbf{Z} - \hat{\boldsymbol{\theta}}\|^2 - n + 2\sum_{i=1}^n \mathrm{Cov}_{\boldsymbol{\theta}}(Z_i, \hat{\theta}_i).\end{aligned}$$

The difficulty in simplifying the expression for the risk in the last display is that we need to compute the expectations of the two terms. C. M. Stein (1981) developed an ingenious way to tackle this problem when $\hat{\boldsymbol{\theta}} \equiv h(\mathbf{Z})$ is an *almost differentiable* function (to be defined formally soon). Note that for the James-Stein estimator, $h(\mathbf{Z}) = (1 - \frac{n-2}{\|\mathbf{Z}\|^2})\mathbf{Z}$ which is differentiable.

3.3.1 Stein's lemmas

The following integration by parts identity will be an important tool in our analysis.

Lemma 3.1 (Stein's lemma[6].)**.** *Let $Z \sim N(0,1)$. Let $h : \mathbb{R} \to \mathbb{R}$ be an absolutely continuous[7] function (differentiable is sufficient) such that $\mathbb{E}[|h'(Z)|] < \infty$. Then,*

$$\mathbb{E}[h'(Z)] = \mathbb{E}[Zh(Z)]. \tag{3.6}$$

Proof. First note that if the result holds for a function h it also holds for h plus a constant, and so we can assume without loss of generality that $h(0) = 0$. Let $\phi(z) := \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$, for $z \in \mathbb{R}$, be the standard normal density. Note that $\phi(\cdot)$ satisfies the important equality

$$\phi'(z) = -z\phi(z), \qquad \text{for} \quad z \in \mathbb{R}. \tag{3.7}$$

[6]We note that this lemma has a converse, and this has become extremely important in its own right, studied and further developed in probability theory for proving convergence to normality.

[7]*A function $h : \mathbb{R} \to \mathbb{R}$ is absolutely continuous if and only if h has a derivative h' almost everywhere and $h(z) - h(a) = \int_a^z h'(t)dt$, for all $a, z \in \mathbb{R}$.*

Observe that,

$$\begin{aligned}
\int_0^\infty h'(y)\phi(y)\,dy &= \int_0^\infty h'(y)\left[\int_y^\infty z\phi(z)\,dz\right]dy \\
&= \int_0^\infty h'(y)\left[\int_0^\infty I_{(y,\infty)}(z)z\phi(z)\,dz\right]dy \\
&= \int_0^\infty \left[\int_0^\infty h'(y)zI_{(y,\infty)}(z)\phi(z)\,dy\right]dz \\
&= \int_0^\infty z\left[\int_0^\infty I_{(0,z)}(y)h'(y)\,dy\right]\phi(z)\,dz \\
&= \int_0^\infty z\left[\int_0^z h'(y)\,dy\right]\phi(z)\,dz \;=\; \int_0^\infty zh(z)\phi(z)dz,
\end{aligned}$$

where the first equality follows from Eq. 3.7 and the third equality follows from Fubini's theorem (which is justified by the assumption $\mathbb{E}[|h'(Z)|] < \infty$). A similar calculation shows that $\int_{-\infty}^0 h'(y)\phi(y)\,dy = \int_{-\infty}^0 zh(z)\phi(z)\,dz$. The desired result now follows by adding these together.

□

Suppose now that $X \sim N(\theta, \sigma^2)$. Then, the above lemma immediately yields[8]

$$\mathbb{E}[(X-\theta)h(X)] = \sigma^2\mathbb{E}[h'(X)]. \tag{3.8}$$

Remark (Some intuition for Eq. 3.8). The above result, although fairly simple, is quite remarkable. Suppose that $X \sim N(\theta, 1)$, where θ is unknown, and we had a (potentially) complicated function $h(X)$ delivering an estimate of θ. Suppose further that we wanted to estimate $\mathrm{Cov}(X, h(X)) = \mathbb{E}[(X-\theta)h(X)]$. To get an unbiased estimate of this covariance, from the definition, we'd have to either know θ, which is unknown, or we'd have to know $\mathbb{E}_\theta[h(X)]$, which again, will generically depend on the unknown θ (not to mention that it may be potentially intractable). On the other hand, Stein's lemma gives us a simple unbiased estimate: $h'(X)$! This is free from θ, and in many cases it is possible to calculate — just take the derivative of our estimator and evaluate it at the data.

Next let us indicate the regularity conditions needed for the extension of Lemma 3.1 to the multidimensional case. The following is taken from C. M. Stein (1981).

[8] Define $Z := (X-\theta)/\sigma$ and let $\tilde{h}(z) := h(\sigma z + \theta)$. Applying Stein's lemma to Z and $\tilde{h}$ now yields the desired result.

Definition 3.2 (Almost differentiable function)**.** A function $h : \mathbb{R}^n \to \mathbb{R}$ will be called *almost differentiable*[9] if there exists a function $\nabla h : \mathbb{R}^n \to \mathbb{R}^n$ such that for all $z \in \mathbb{R}^n$,

$$h(x+z) - h(x) = \int_0^1 z^\top \nabla h(x+tz)\, dt, \tag{3.9}$$

for almost all $x \in \mathbb{R}^n$. Essentially, $\nabla h(\cdot)$ is the vector differential operator of first partial derivatives[10] with i-th coordinate $\nabla_i h(\cdot) = \frac{\partial h}{\partial x_i}(\cdot)$.

Lemma 3.2. *Let $\mathbf{Z} = (Z_1, \dots, Z_n)$ where $Z_i \sim N(\theta_i, 1)$ are independent random variables, for $i = 1, \dots, n$. Let $h : \mathbb{R}^n \to \mathbb{R}$ be an almost differentiable function such that $\mathbb{E}\|\nabla h(\mathbf{Z})\| < \infty$. Then,*

$$\mathbb{E}[(\mathbf{Z} - \boldsymbol{\theta})h(\mathbf{Z})] = \mathbb{E}[\nabla h(\mathbf{Z})]. \tag{3.10}$$

Proof. W.l.o.g. let $\boldsymbol{\theta} = \mathbf{0}$. Fix some $i \in \{1, \dots, n\}$ and $\mathbf{Z}_{-i} \in \mathbb{R}^{n-1}$. Then, the function $h(\cdot, \mathbf{Z}_{-i})$ is a univariate absolutely continuous function and we can apply Stein's univariate lemma. Hence, using the independence of Z_i and $\mathbf{Z}_{-i}$,

$$\mathbb{E}\left[\frac{\partial h}{\partial z_i}(\mathbf{Z}) \mid \mathbf{Z}_{-i}\right] = \int \nabla_i h(z, \mathbf{Z}_{-i})\phi(z)\, dz = \int z h(z, \mathbf{Z}_{-i})\phi(z)\, dz = \mathbb{E}\left[Z_i h(\mathbf{Z}) \mid \mathbf{Z}_{-i}\right].$$

Taking an expectation over $\mathbf{Z}_{-i}$ now yields the desired result.

□

A function $h : \mathbb{R}^n \to \mathbb{R}^n$ is *almost differentiable* if all its coordinate functions are. Write $h = (h_1, \dots, h_n)$ for the coordinate functions, where each $h_i : \mathbb{R}^n \to \mathbb{R}$ is almost differentiable. Then, by the last result, for each $i = 1, \dots, n$,

$$\mathbb{E}[(\mathbf{Z} - \boldsymbol{\theta})h_i(\mathbf{Z})] = \mathbb{E}[\nabla h_i(\mathbf{Z})].$$

Taking the i-th equality in the above, and then summing over all $i = 1, \dots, n$ yields the following result.

Lemma 3.3. *Let $\mathbf{Z} = (Z_1, \dots, Z_n)$ where $Z_i \sim N(\theta_i, 1)$ are independent random variables, for $i = 1, \dots, n$. Let $h : \mathbb{R}^n \to \mathbb{R}^n$ be an almost differentiable function such that $\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^n |\nabla_i h_i(\mathbf{Z})|\right] < \infty$. Then,*

$$\sum_{i=1}^n \mathrm{Cov}(Z_i, h_i(\mathbf{Z})) = \sum_{i=1}^n \mathbb{E}[(Z_i - \theta_i)h_i(\mathbf{Z})] = \mathbb{E}\left[\sum_{i=1}^n \frac{\partial h_i}{\partial z_i}(\mathbf{Z})\right]. \tag{3.11}$$

[9] Observe that Eq. 3.9 indeed reduces to the notion of absolute continuity when $n = 1$.

[10] Note that an almost differentiable function h has partial derivatives almost everywhere.

Definition 3.3 (Stein's unbiased risk estimator)**.** Observe that if $\hat{\boldsymbol{\theta}} \equiv h(\mathbf{Z}) = (\hat{\theta}_1, \dots, \hat{\theta}_n)$ is an almost differentiable function of $\mathbf{Z}$ and Lemma 3.3 is applicable, then Eq. 3.6 immediately yields an *unbiased* estimator of the risk $R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$:

$$\hat{R} := -n + \|\mathbf{Z} - \hat{\boldsymbol{\theta}}\|^2 + 2\sum_{i=1}^{n} \frac{\partial \hat{\theta}_i}{\partial z_i}(\mathbf{Z}), \tag{3.12}$$

i.e.,

$$\mathbb{E}_{\boldsymbol{\theta}}[\hat{R}] = R(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}). \tag{3.13}$$

Here $\hat{R}$ is called *Stein's unbiased risk estimate* (SURE).

Of course, in order for this to be useful, we need to figure out how to compute $\sum_{i=1}^{n} \frac{\partial \hat{\theta}_i}{\partial z_i}(\mathbf{Z})$ for the estimator $\hat{\boldsymbol{\theta}}$ of interest (and, determine that $\hat{\boldsymbol{\theta}}$ is almost differentiable so that Stein's lemma is applicable). Note that many statistical estimators (e.g., projections on closed convex sets[11], etc.) satisfy the almost differentiability assumption.

Definition 3.4 (Divergence)**.** The quantity

$$D \equiv D(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^{n} \frac{\partial \hat{\theta}_i}{\partial z_i}(\mathbf{Z})$$

is called the *divergence* of the estimator $\hat{\boldsymbol{\theta}}$ and provides a measure of the effective dimension of the fit[12] $\hat{\boldsymbol{\theta}}$.

Remark (SURE for model selection)*.* SURE can be an extremely useful quantity. Aside from plainly estimating the risk of an estimator, we could also use it for model selection purposes: if our estimator depended on a tuning parameter $\lambda \in \Lambda$, denoted $\hat{\boldsymbol{\theta}}_\lambda$, then we could choose this parameter to minimize SURE:

$$\hat{\lambda} := \arg\min_{\lambda \in \Lambda} \left\{ \|\mathbf{Z} - \hat{\boldsymbol{\theta}}_\lambda\|^2 + 2\sum_{i=1}^{n} \frac{\partial \hat{\theta}_{\lambda,i}}{\partial z_i}(\mathbf{Z}) \right\}.$$

There is a considerable amount of classic literature that studies the minimization of a SURE-like risk estimate, for relatively simple procedures (such as linear smoothers) where the divergence is easily computable. Examples are: (K.-C. Li 1985, 1986, 1987; Donoho and Johnstone 1995; Meyer and Woodroofe 2000; Bradley Efron, Hastie, Johnstone, and Tibshirani 2004; Tibshirani and Taylor 2012). Xie, Kou, and Brown (2012) use SURE to estimate the shrinkage parameter in the *heteroscedastic* analogue of the Gaussian hierarchical model Eq. 3.4, and also show a kind of asymptotic optimality property for the SURE estimator.

[11] It is well-known that the projection operator onto a closed convex set in 1-Lipschitz. Further, any Lipschitz continuous function h is almost differentiable with a bounded gradient (see e.g., Meyer and Woodroofe (2000)). Thus, projections on closed convex sets are almost differentiable.

[12] To see, observe that if $\hat{\boldsymbol{\theta}}$ is a linear projection onto a space of dimension d, say $\hat{\boldsymbol{\theta}} = Q\mathbf{Z}$, where Q is a $n \times n$ projection matrix, then $D(\hat{\boldsymbol{\theta}}) = \mathrm{tr}(Q) = d$, for all $\mathbf{Z} \in \mathbb{R}^n$.

3.3.2 Risk of the James-Stein estimator

Now let us use the above results to find the risk of the James-Stein estimator $\hat{\boldsymbol{\theta}}_{JS} \equiv h(\mathbf{Z})$ (see Eq. 3.3). Here, $h_i(x) = x_i - \frac{(n-2)x_i}{x_1^2+...+x_n^2}$, for $x \in \mathbb{R}$. Then,

$$\frac{\partial h_i(x)}{\partial x_i} = 1 - \frac{n-2}{x_1^2 + ... + x_n^2} + \frac{(n-2)x_i(2x_i)}{(x_1^2 + ... + x_n^2)^2} = 1 - \frac{n-2}{\|x\|^2} + \frac{2(n-2)x_i^2}{\|x\|^4},$$

and thus,

$$\sum_{i=1}^{n} \frac{\partial h_i}{\partial z_i}(\mathbf{Z}) = n - \frac{n(n-2)}{\|\mathbf{Z}\|^2} + \frac{2(n-2)\sum_{i=1}^{n} Z_i^2}{\|\mathbf{Z}\|^4} = n - \frac{(n-2)^2}{\|\mathbf{Z}\|^2}.$$

Thus, for the James-Stein estimator, using Eq. 3.12, we have

$$\hat{R} = n + \frac{(n-2)^2}{\|\mathbf{Z}\|^2} - 2\frac{(n-2)^2}{\|\mathbf{Z}\|^2} = n - \frac{(n-2)^2}{\|\mathbf{Z}\|^2}.$$

By Eq. 3.13,

$$R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS}) = \mathbb{E}_{\boldsymbol{\theta}}[\hat{R}] = E_{\boldsymbol{\theta}}\left[n - \frac{(n-2)^2}{\|\mathbf{Z}\|^2}\right] < n \equiv R(\boldsymbol{\theta}, \mathbf{Z}).$$

Hence when $n > 2$, the James-Stein estimator always has smaller compound risk than the MLE $\mathbf{Z}$; thus $\mathbf{Z}$ is inadmissible.

When $\|\boldsymbol{\theta}\|$ is large, $\|\mathbf{Z}\|$ will be large with high probability. Then the James-Stein estimator and $\mathbf{Z}$ will be very similar and will have similar risk. But when $\|\boldsymbol{\theta}\|$ is small there can be a substantial decrease in risk using the James-Stein estimator instead of $\mathbf{Z}$. If $\boldsymbol{\theta} = \mathbf{0}$, then $\|\mathbf{Z}\|^2 = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$. We can show that $\mathbb{E}_{\boldsymbol{\theta}=\mathbf{0}}\left[\frac{1}{\|\mathbf{Z}\|^2}\right] = \frac{1}{n-2}$. Using this, we get

$$R(\mathbf{0}, \hat{\boldsymbol{\theta}}_{JS}) = \mathbb{E}_{\boldsymbol{\theta}=\mathbf{0}}[\hat{R}] = E_{\boldsymbol{\theta}}\left[n - \frac{(n-2)^2}{\|\mathbf{Z}\|^2}\right] = n - \frac{(n-2)^2}{n-2} = 2.$$

Thus, regardless of the dimension of $\boldsymbol{\theta}$ and $\mathbf{Z}$, at the origin $\boldsymbol{\theta} = \mathbf{0}$, the James-Stein estimator has risk equal to two.

In the more general compound case, we note that the risk of the James-Stein estimator depends only on n and $\|\boldsymbol{\theta}\|_2^2$. Casella and Hwang (1982) derive the following interpretable bounds valid for any $n \geq 3$:

$$\frac{n(2 + \|\boldsymbol{\theta}\|_2^2)}{n + \|\boldsymbol{\theta}\|_2^2} \leq R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS}) \leq \frac{n(2 + \|\boldsymbol{\theta}\|_2^2) - 4}{n - 2 + \|\boldsymbol{\theta}\|_2^2}.$$

Note that the above inequalities are tight for $\boldsymbol{\theta} = 0$ as shown by our previous calculations. Furthermore these inequalities demonstrate formally that $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS}) \sim n$ when $\|\boldsymbol{\theta}\|_2^2$ is large, and also that substantial risk savings are possible when $\|\boldsymbol{\theta}\|_2^2$ is small.

Since $R(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{JS})$ only depends on n and $\|\boldsymbol{\theta}\|_2^2$, we can compute it explicitly numerically. Hence we can plot a Figure analogous to Figure 1.1 in Chapter 1.

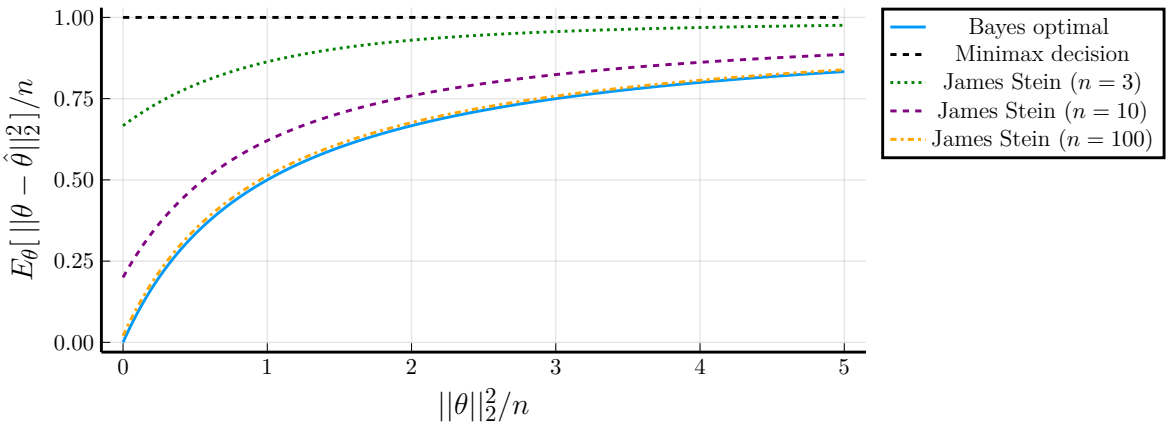


Figure 3.1: Risk functions for compound estimation of θ_i based on $Z_i \sim \mathcal{N}(\theta_i, 1)$, $i = 1, \dots, n$.

3.4 Extensions and generalizations

3.4.1 Shrinking Toward an Arbitrary Point

Thus far, we have considered estimators that shrink toward zero, but we need not do so. As it turns out, we can shrink toward an arbitrary point $\boldsymbol{\theta}_0 \in \mathbb{R}^n$. Define the estimator

$$\hat{\boldsymbol{\theta}}_{JS}^{\boldsymbol{\theta}_0} := \boldsymbol{\theta}_0 + \left(1 - \frac{n-2}{\|\mathbf{Z} - \boldsymbol{\theta}_0\|^2}\right)(\mathbf{Z} - \boldsymbol{\theta}_0).$$

Then, $\hat{\boldsymbol{\theta}}_{JS}^{\boldsymbol{\theta}_0}$ also dominates the MLE $\mathbf{Z}$ everywhere[13].

This observation turns out to be remarkably useful. Green and Strawderman (1991) build on this observation and develop a method for combining biased and unbiased measurements Z_i: they shrink the unbiased measurements toward the location given by the biased measurements. Ignatiadis and Wager (2019) build on this result and shrink toward predictions from arbitrary machine learning models.

[13]The see this consider $\mathbf{Y} := \mathbf{Z} - \boldsymbol{\theta}_0$. Then $\mathbf{Y} \sim N(\boldsymbol{\theta} - \boldsymbol{\theta}_0, I_n)$, and $\hat{\boldsymbol{\theta}}_{JS}^{\boldsymbol{\theta}_0} \equiv \hat{\boldsymbol{\theta}}_{JS}^{\boldsymbol{\theta}_0}(\mathbf{Z}) = \boldsymbol{\theta}_0 + \hat{\boldsymbol{\theta}}_{JS}(\mathbf{Y})$. Then, for any $\boldsymbol{\theta} \in \mathbb{R}^n$,

$$\begin{aligned} R(\hat{\boldsymbol{\theta}}_{JS}^{\boldsymbol{\theta}_0}, \boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{\theta}}\|\hat{\boldsymbol{\theta}}_{JS}^{\boldsymbol{\theta}_0}(\mathbf{Z}) - \boldsymbol{\theta}\|^2 = \mathbb{E}_{\boldsymbol{\theta}}\|\hat{\boldsymbol{\theta}}_{JS}(\mathbf{Y}) - (\boldsymbol{\theta} - \boldsymbol{\theta}_0)\|^2 \\ &= R(\hat{\boldsymbol{\theta}}_{JS}, \boldsymbol{\theta} - \boldsymbol{\theta}_0) < R(\mathbf{Y}, \boldsymbol{\theta} - \boldsymbol{\theta}_0) = R(\mathbf{Z}, \boldsymbol{\theta}). \end{aligned}$$

3.4.2 Shrinking towards the group mean: An empirical Bayes approach

In practice, instead of arbitrarily picking some point $\boldsymbol{\theta}_0 \in \mathbb{R}^n$, it might make sense to chose $\boldsymbol{\theta}_0 := (\bar{Z}, \bar{Z}, \dots, \bar{Z}) \in \mathbb{R}^n$ (where $\bar{Z} := \sum_{i=1}^n Z_i$ is mean of the observed Z_i's) so as to adapt to the true center of the θ_i's.

Assume the following Bayesian setup:

$$\theta_i \overset{iid}{\sim} N(\theta_0, \tau^2) \quad \text{and} \quad Z_i \mid \theta_i \overset{ind}{\sim} N(\theta_i, 1), \qquad \text{for} \quad i = 1, \dots, n,$$

and unknown $\theta_0 \in \mathbb{R}$ and $\tau > 0$. The marginal distribution of our data is

$$Z_i \overset{iid}{\sim} N(\theta_0, \tau^2 + 1),$$

and the posterior distribution of the parameters are

$$\theta_i \mid Z_i \overset{ind}{\sim} N(\theta_0 + (1-\rho)(Z_i - \theta_0), 1-\rho), \qquad \text{where} \quad \rho = (1+\tau^2)^{-1}.$$

Here $\theta_0 + (1-\rho)(Z_i - \theta_0)$ is the Bayes estimator for θ_i, but θ_0 and ρ are unknown.

Taking an empirical Bayes approach, we can use the unbiased estimator $\bar{Z}$ of θ_0 to estimate θ_0 and use $S := \sum_{i=1}^n (Z_i - \bar{Z})^2$ to estimate τ^2. In particular, note that $S \sim (1+\tau^2)\chi^2_{n-1}$, and thus, $\mathbb{E}\left[\frac{n-3}{S}\right] = (1+\tau^2)^{-1}$. This gives us $\hat{\boldsymbol{\theta}}_{JS}^{\bar{Z}\mathbf{1}}$—the empirical Bayes estimator of $\boldsymbol{\theta}$—where the i-th coordinate of $\hat{\boldsymbol{\theta}}_{JS}^{\bar{Z}\mathbf{1}}$ is

$$\hat{\boldsymbol{\theta}}_{JS}^{\bar{Z}\mathbf{1}}(i) = \bar{Z} + \left(1 - \frac{n-3}{S}\right)(Z_i - \bar{Z}), \qquad \text{for} \;\; i = 1, \dots, n.$$

If $n > 3$, this estimator dominates the MLE everywhere.

3.5 Bibliographic Remarks

The material in this chapter follows closely the beautiful lecture notes by Emmanuel Candès for the class "STATS300C: Theory of Statistics" taught at Stanford University. The lecture notes are available at the following link: https://candes.su.domains/teaching/stats300c/index.html

4 Understanding and improving James-Stein through regression

In this chapter we will explore how linear regression and the class of linear estimators can shed light into the operating mechanisms of the James-Stein procedure, and also lead to more practical and powerful methods in practice.

It is worth emphasizing that it is also fruitful to study the reverse question: can we use James-Stein to improve statistical performance when conducting linear regression. This is not the topic of this chapter, but we will pursue this question later.

4.1 James-Stein and restricted empirical Bayes

In revisiting James-Stein, it is instructive to consider the empirical Bayes motivation thereof. Recall that we seek to estimate parameters $\theta_1, \dots, \theta_n$ well in squared error loss and have have access to standard normal measurements $Z_i \mid \theta_i \sim \mathcal{N}(\theta_i, 1)$.

Consider the following class of priors:

$$\mathcal{G}_{\text{scale}} := \{\mathcal{N}(0, A) \,:\, A \geq 0\}.$$

Given any prior $G = \mathcal{N}(0, A) \in \mathcal{G}_{\text{scale}}$,

$$t_G(z) := \mathbb{E}_G\left[\theta \mid Z = z\right] = \frac{A}{A+1} Z, \text{ where } \theta \sim G,\ Z \mid \theta \sim \mathcal{N}(\theta, 1).$$

This the optimal decision for estimating θ with squared error loss. Given n parallel draws $Z_i \sim \mathcal{N}(\theta_i, 1)$, we saw that James-Stein essentially estimates a suitable prior[1]

$$\widehat{G}^{JS} = \mathcal{N}\left(0,\ \widehat{A}^{JS}\right),\ \ \widehat{A}^{JS} = 1 - \frac{(n-2)}{\|\mathbf{Z}\|^2},$$

from $\mathcal{G}_{\text{scale}}$ and then lets $\hat{\theta}_i = \mathbb{E}_{\widehat{G}^{JS}}\left[\theta_i \mid Z_i\right]$.

[1]The interpretation below fails when $\widehat{A}^{JS} < 0$. To avoid such cases one would instead use the positive-part James-Stein estimator: in the definition that follows, one replaces $\widehat{A}^{JS}$ by $\widehat{A}^{JS+} = \max\{\widehat{A}^{JS}, 0\}$. In fact it can be shown that the positive part James-Stein estimator dominates the regular James-Stein estimator.

here is another useful interpretation: any $G \in \mathcal{G}_{\text{scale}}$ defines a Bayes-optimal decision rule $t_G(\cdot)$. Scanning over all $G \in \mathcal{G}_{\text{scale}}$ we get:

$$\mathcal{L}_{\text{scale}} = \{t_G(\cdot) \ : \ G \in \mathcal{G}_{\text{scale}}\} = \{z \mapsto \lambda \cdot z \ : \ \lambda \in [0,1)\}.$$

Notice that the RHS is merely a class of estimators, and once this connection has been made, we may forget that we ever posited the empirical Bayes model with $\theta \sim G \in \mathcal{G}_{\text{scale}}$. Instead, we may change our goal post to the following task: choose the best estimator $t_\lambda(\cdot)$ in the class $\mathcal{L}_{\text{scale}}$.

The next proposition derives "oracle" choices of $t_\lambda(\cdot)$.

4.1.1 Optimal linear estimators

Proposition 4.1.

1. *(Empirical Bayes) Suppose that $\theta \sim G$ with $\mathbb{E}_G\left[\theta^2\right] < \infty$, and that $\mathbb{E}\left[Z \mid \theta\right] = \theta$, $\text{Var}\left[Z \mid \theta\right] = 1$. Then for the Bayes risk, under squared error loss, $R(t,\theta) = \mathbb{E}_G\left[(\theta - t(Z))^2\right]$ is minimized over $t \in \mathcal{L}_{scale}$ by:*

$$t_{\lambda^{*,B}}(\cdot), \ \textit{where } \lambda^{*,B} = \frac{\mathbb{E}_G\left[\theta^2\right]}{\mathbb{E}_G\left[Z^2\right]} = 1 - \frac{1}{\mathbb{E}_G\left[Z^2\right]}$$

2. *(Compound decisions) Now we assume that $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ is fixed, but continue to impose that $\mathbb{E}\left[Z_i \mid \theta_i\right] = \theta_i$, $\text{Var}\left[Z_i \mid \theta_i\right] = 1$. Then if we look at compound loss $R(t,\boldsymbol{\theta}) = \frac{1}{n}\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^n (\theta_i - t(Z_i))^2\right]$ is minimized over $t \in \mathcal{L}_{scale}$ by:*

$$t_{\lambda^{*,C}}(\cdot), \ \textit{where } \lambda^{*,C} = \frac{\|\boldsymbol{\theta}\|^2}{\mathbb{E}_{\boldsymbol{\theta}}\left[\|\mathbf{Z}\|^2\right]} = 1 - \frac{n}{\mathbb{E}_{\boldsymbol{\theta}}\left[\|\mathbf{Z}\|_2^2\right]}.$$

Proof. Let us start with the first result. We seek to minimize:

$$\mathbb{E}_G\left[(\theta - \lambda Z)^2\right] = \mathbb{E}_G\left[\theta^2\right] + \lambda^2 \mathbb{E}_G\left[Z^2\right] - 2\lambda \mathbb{E}_G\left[\theta^2\right].$$

This is a quadratic and we immediately find that:

$$\lambda^{*,B} = \frac{\mathbb{E}_G\left[\theta^2\right]}{\mathbb{E}_G\left[Z^2\right]} = \frac{\mathbb{E}_G\left[\theta^2\right]}{\mathbb{E}_G\left[\theta^2\right] + 1} = 1 - \frac{1}{\mathbb{E}_G\left[Z^2\right]}. \tag{4.1}$$

For the compound result, we instead expand:

$$\sum_{i=1}^n \mathbb{E}_{\boldsymbol{\theta}}\left[(\theta_i - \lambda Z_i)^2\right] = \sum_{i=1}^n \left(\theta_i^2 + \lambda^2 \mathbb{E}_{\theta_i}\left[Z_i^2\right] - 2\lambda\theta_i^2\right).$$

Again minimizing the above quadratic we find that:

$$\lambda^{*,C} = \frac{\|\boldsymbol{\theta}\|_2^2}{\mathbb{E}_{\boldsymbol{\theta}}\left[\|\mathbf{Z}\|^2\right]} = 1 - \frac{n}{\mathbb{E}_{\boldsymbol{\theta}}\left[\|\mathbf{Z}\|^2\right]}.$$

□

A take-home message from the above results is the following: the James-Stein estimator may be seen to approximate both oracle estimators in the proposition above. Furthermore, one may hope that James-Stein will perform well even if $Z \mid \theta$ is not Gaussian—it suffices that the specification of first two moments is correct (conditionally on θ).

4.1.2 Competing against the best linear estimator through SURE

Suppose now that we seek to match the best linear estimator in the class $\mathcal{L}_{\text{scale}}$. Recall from the previous chapter that (under Gaussian noise) we can estimate the compound mean squared error of any estimator (that satisfies almost differentiability) through SURE (Stein's Unbiased Risk estimate). For the estimator with $\hat{\theta}_i = \lambda \cdot Z_i$ we have that $\frac{\partial \hat{\theta}_i}{\partial z_i} = \lambda$ and so we see that SURE is equal to:[2]

$$\widehat{R}(t_\lambda) = -1 + (1-\lambda)^2 \|\mathbf{Z}\|_2^2 / n + 2\lambda. \tag{4.2}$$

It holds that $\mathbb{E}_{\boldsymbol{\theta}}\left[\widehat{R}(t_\lambda)\right] = R(t_\lambda, \boldsymbol{\theta})$.

Exercise 4.1. The unbiasedness of the expression in Eq. 4.2 requires only conditional moments given θ_i. In particular, prove that as long as $\mathbb{E}_{\theta_i}[Z_i] = \theta_i$ and $\mathrm{Var}_{\theta_i}[Z_i] = 1$, then:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\widehat{R}(t_\lambda)\right] = R(t_\lambda, \boldsymbol{\theta}).$$

In fact, we can prove that SURE is a good estimator of the true risk uniformly over all $\lambda \in [0, 1]$.

Proposition 4.2. *Suppose all Z_i are independent and that $\mathbb{E}_{\theta_i}[Z_i] = \theta_i$ and $\mathrm{Var}_{\theta_i}[Z_i] = 1$. If it also holds that:*

$$\frac{1}{n^2}\sum_{i=1}^n \theta_i^2 \to 0 \text{ as } n \to \infty,$$

and that:

$$\max_i \mathbb{E}_{\theta_i}\left[|Z_i - \theta_i|^4\right] < \infty,$$

[2] Compared to the previous chapter, we have rescaled SURE and the compound loss by a factor $1/n$.

then:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\sup_{\lambda\in[0,1]}\left|\widehat{R}(t_\lambda)-R(t_\lambda,\boldsymbol{\theta})\right|\right]\to 0 \text{ as } n\to\infty.$$

In fact, we also get convergence of the actual loss*:*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\sup_{\lambda\in[0,1]}\left|\widehat{R}(t_\lambda)-\ell(t_\lambda,\boldsymbol{\theta})\right|\right]\to 0 \text{ as } n\to\infty,$$

where $\ell(t_\lambda,\boldsymbol{\theta})=\frac{1}{n}\sum_{i=1}^n(\theta_i-t_\lambda(Z_i))^2$.

Proof. One can check that:

$$\widehat{R}(t_\lambda)-R(t_\lambda,\boldsymbol{\theta})=(1-\lambda)^2\frac{\|\mathbf{Z}\|_2^2}{n}-(1-\lambda)^2\frac{\mathbb{E}_{\boldsymbol{\theta}}\left[\|\mathbf{Z}\|_2^2\right]}{n}.$$

Hence:

$$\sup_{\lambda\in[0,1]}\left|\widehat{R}(t_\lambda)-R(t_\lambda,\boldsymbol{\theta})\right|=\left|\frac{1}{n}\sum_{i=1}^n(Z_i^2-\mathbb{E}_{\theta_i}\left[Z_i^2\right])\right|.$$

To study the above quantity, let us write $\varepsilon_i=Z_i-\theta_i$. Then:

$$Z_i^2-\mathbb{E}_{\theta_i}\left[Z_i^2\right]=\varepsilon_i^2-1+2\theta_i\varepsilon_i.$$

It thus suffices to prove the following. First:

$$\frac{1}{n}\sum_{i=1}^n\theta_i\varepsilon_i\overset{L_1}{\to}0 \text{ as } n\to\infty.$$

This follows e.g., by applying the Cauchy-Schwarz inequality after noting that the above has expectation equal to 0, and that

$$\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n\theta_i\varepsilon_i\right]=\frac{1}{n^2}\sum_{i=1}^n\theta_i^2,$$

where the latter converges to 0 by assumption. Furthermore:

$$\frac{1}{n}\sum_{i=1}^n\varepsilon_i^2-1\overset{L_1}{\to}0 \text{ as } n\to\infty.$$

This follows again by Cauchy-Schwarz after noting that $\mathbb{E}_{\theta_i}\left[\varepsilon_i^2\right]=1$ and that:

$$\mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^n\varepsilon_i^2\right]=\frac{1}{n^2}\sum_{i=1}^n\mathrm{Var}\left[\epsilon_i^2\right]\le\frac{1}{n^2}\sum_{i=1}^n\mathbb{E}_{\theta_i}\left[\varepsilon_i^4\right]=\frac{1}{n^2}\sum_{i=1^n}\mathbb{E}_{\theta_i}\left[(Z_i-\theta_i)^4\right],$$

and the latter converges to 0 by our second assumption.

To prove the second part of the proposition, it suffices to use the triangle inequality and to argue that:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\sup_{\lambda\in[0,1]}|\ell(t_\lambda,\boldsymbol{\theta})-R(t_\lambda,\boldsymbol{\theta})|\right]\to 0 \text{ as } n\to\infty.$$

The argument is analogous.

□

As we mentioned last week, one generally applicable model selection strategy is given by minimizing SURE. Hence (instead of doing, e.g., James-Stein) we could use the estimator $t_{\widehat{\lambda}^{\text{SURE}}}$, where:

$$\widehat{\lambda}^{\text{SURE}}\in\operatorname*{argmin}_{\lambda\in[0,\,1]}\widehat{R}(t_\lambda).$$

In the present setting we can immediately compute that:

$$\widehat{\lambda}^{\text{SURE}}=\max\left\{0,\,1-\frac{n}{\|\mathbf{Z}\|_2^2}\right\}.$$

The resulting estimator then is:

$$\widehat{\theta}_i^{\text{SURE}}=\widehat{\lambda}^{\text{SURE}}\cdot Z_i.$$

Notice that this is almost equal to the James-Stein estimator (two differences: we do not let the shrinkage factor be <0 and second, we replace the factor $n-2$ in the James-Stein estimator by n)—both are typically of little consequence in practice.

Hence one can think of SURE being applicable to James-Stein in two ways:[3]

1. James-Stein may be derived (modulo the minor caveat above) by minimizing SURE over the class of linear estimators $\mathcal{L}_{\text{scale}}$.
2. SURE can be used to estimate and theoretically assess the risk of James-Stein.

With Proposition 4.2, we can now prove the following asymptotic result for $\widehat{\boldsymbol{\theta}}^{\text{SURE}}$.

Theorem 4.1.

$$\limsup_{n\to\infty}\left(R(\widehat{\boldsymbol{\theta}}^{SURE},\boldsymbol{\theta})-\inf_{\lambda\in[0,1]}R(t_\lambda(\cdot),\boldsymbol{\theta})\right)\le 0.$$

[3]This perspective is further pursued by Tibshirani and Rosset (2019).

Proof. Let λ_n^* be such that:

$$R(t_{\lambda_n^*}, \boldsymbol{\theta}) = \inf_{\lambda \in [0,1]} R(t_\lambda, \boldsymbol{\theta}).$$

Then:

$$\begin{aligned}
&\ell(t_{\widehat{\lambda}^{\text{SURE}}}, \boldsymbol{\theta}) - \ell(t_{\lambda_n^*}(\cdot), \boldsymbol{\theta}) \\
&\quad = \left(\ell(t_{\widehat{\lambda}^{\text{SURE}}}, \boldsymbol{\theta}) - \widehat{R}(t_{\widehat{\lambda}^{\text{SURE}}})\right) - \left(\ell(t_{\lambda_n^*}, \boldsymbol{\theta}) - \widehat{R}(t_{\lambda_n^*})\right) + \underbrace{\left(\widehat{R}(t_{\widehat{\lambda}^{\text{SURE}}}) - \widehat{R}(t_{\lambda_n^*})\right)}_{\leq 0} \\
&\quad \leq 2 \sup_{\lambda \in [0,1]} \left|\widehat{R}(t_\lambda) - \ell(t_\lambda, \boldsymbol{\theta})\right|.
\end{aligned}$$

Taking expectations:

$$R(t_{\widehat{\lambda}^{\text{SURE}}}, \boldsymbol{\theta}) - R(t_{\lambda_n^*}(\cdot), \boldsymbol{\theta}) \leq 2\mathbb{E}_{\boldsymbol{\theta}} \left[\sup_{\lambda \in [0,1]} \left|\widehat{R}(t_\lambda) - \ell(t_\lambda, \boldsymbol{\theta})\right|\right],$$

and the RHS converges to 0 by Proposition 4.2.

□

Exercise 4.2. Suppose our working class of priors is given by:

$$\mathcal{G}_{\text{loc-scale}} := \{\mathcal{N}(u, A) \,:\, u \in \mathbb{R},\ A > 0\}.$$

Then the induced class of estimators is given by:

$$\mathcal{L}_{\text{loc-scale}} = \{t_G(\cdot) \,:\, \mathcal{G}_{\text{loc-scale}}\} = \{z \mapsto a + \lambda \cdot z \,:\, a \in \mathbb{R}, \lambda \in [0,1]\}.$$

Rederive/state analogues of results in the section for this broader class of estimators.

4.2 James-Stein and regression to the mean

So far, we have provided intuition and rigorous justification for the James-Stein estimator through the lens of Charles Stein, as well as Efron and Morris. Stigler (1990) asked the following question: can the James-Stein phenomenon be explained in a way that would make sense to statisticians that lived a century before Stein? Stigler's answer is affirmative.

4.2.1 The Efron-Morris baseball dataset

Before formally explaining Stigler's perspective, we informally demonstrate his ideas through a very famous application of empirical Bayes; namely to baseball statistics. Bradley Efron and Morris (1975) considered a dataset of 18 Major League baseball players during the 1970 season. At the start of the season, each of the 18 players in the dataset had 45 at bats and batting average Y_i. Efron and Morris asked the following question: how should we predict the batting average of these 18 players at the end of the season? One way to formalize the result is as follows. Each player has a true "batting rate" p_i, and we observe a binomial sample thereof with 45 trials—each at-bat is a trial:

$$45 \cdot Y_i \mid p_i \sim \text{Binomial}(45, p_i). \tag{4.3}$$

Bradley Efron and Morris (1975) further posited that at the end of the season p_i could be estimated sufficiently accurately based on the data from the remainder of the season. The first three columns of the following table show the data:

Table 4.1: Baseball dataset of Efron-Morris

Name	InitialBattingAvg	RemainingBattingAvg	Z	
Roberto Clemente	0.4	0.346	-1.351	-2.1
Frank Robinson	0.378	0.298	-1.657	-2.788
Frank Howard	0.356	0.276	-1.966	-3.11
Jay Johnstone	0.333	0.222	-2.28	-3.958
Ken Berry	0.311	0.273	-2.599	-3.166
Jim Spencer	0.311	0.27	-2.599	-3.2
Don Kessinger	0.289	0.264	-2.924	-3.29
Luis Alvarado	0.267	0.21	-3.257	-4.149
Ron Santo	0.244	0.269	-3.599	-3.228
Ron Swaboda	0.244	0.23	-3.599	-3.827
Rico Petrocelli	0.222	0.264	-3.951	-3.299
Ellie Rodriguez	0.222	0.226	-3.951	-3.894
George Scott	0.222	0.303	-3.951	-2.711
Del Unser	0.222	0.264	-3.951	-3.305
Billy Williams	0.222	0.33	-3.951	-2.329
Bert Campaneris	0.2	0.285	-4.317	-2.983
Thurman Munson	0.178	0.316	-4.698	-2.525
Max Alvis	0.156	0.2	-5.098	-4.317

Efron and Morris sought to conduct shrinkage using James-Stein. To transform Eq. 4.3 to approximate Gaussianity, they considered the variance stabilizing function

$h(p) = \sqrt{45}\arcsin(2p-1)$. Then letting $Z_i = h(Y_i)$ and $\theta_i = h(p_i)$, it approximately holds that:

$$Z_i \mid \theta_i \sim \mathcal{N}(\theta_i, 1).$$

Let us first compare the performance of the naive method in terms of the squared error loss in estimating θ_i:

```
naive_error = round(mean(abs2, tbl.θ .- tbl.Z), digits=2)
naive_error
```

```
1.11
```

What about James-Stein?

```
z_bar = mean(tbl.Z)
shrinkage_factor = 1-(18-3) / sum(abs2, tbl.Z .- z_bar)
js_fit = shrinkage_factor .* tbl.Z .+ (1 .- shrinkage_factor) .* z_bar
js_error = round(mean(abs2, tbl.θ .- js_fit), digits=2)
js_error
```

```
0.35
```

Perhaps the gains on this scale may seem to be contrived—we are interested in the p_i after all and not θ_i. We can turn the James-Stein estimates $\hat{\theta}_i^{JS}$ for θ_i into estimates for p_i by inverting the variance stabilizing transformation, that is, by estimating p_i by $h^{-1}(\hat{\theta}_i^{JS})$. What is the absolute error loss in estimating p_i by the initial batting average?

```
naive_error_batting_scale =  round(
    mean(abs, tbl.RemainingBattingAvg .- tbl.InitialBattingAvg), digits=3
)
naive_error_batting_scale
```

```
0.059
```

What is the absolute error loss after mapping the James-Stein estimates to the correct scale?

```
inv_arcsine(z) = (sin(z/sqrt(45)) + 1)/2
js_error_batting_scale =  round(
    mean(abs, tbl.RemainingBattingAvg .- inv_arcsine.(js_fit)), digits=3
)
```

```
js_error_batting_scale
```

```
0.03
```

A lovely demonstration of the practical gains that are possible through James-Stein shrinkage! However let us now turn to our initial goal: demonstrating Stigler's argument based on this dataset. His argument is captured by Figure 4.1.

There are 18 points in Figure 4.1, each one corresponds to the (Z_i, θ_i) pairs of the different players. In a practical application, we would only observe Z_i and not θ_i (we seek to estimate θ_i)—thus everything that follows is a thought experiment. In the figure we also plot three lines.

1. We plot in grey the identity line (with intercept 0 and slope 1). The points on this line correspond to estimating θ_i by Z_i (i.e., the traditional maximum likelihood approach). Under our model assumptions, it holds that $\mathbb{E}[Z \mid \theta] = \theta$, hence equivalently the identity line corresponds to the regression $Z \sim \theta$.
2. Imagine now that we actually could observe the θ_i. We seek to predict these (or get a good fit as possible) by a linear function of the Z_i. Stigler notes that this would theoretically be accomplished through the regression $\mathbb{E}[\theta_i \mid Z_i]$ and *not* the regression $\mathbb{E}[Z_i \mid \theta_i]$. In our hypothetical setting here we can estimate $\mathbb{E}[\theta_i \mid Z_i]$ by running the linear regression $\theta_i \sim Z_i$; this is what the purple line shows. Statisticians a century before Stein, e.g., Galton, already knew of the regression to the mean phenomenon: the purple line (and not the gray line) is the correct line with which to predict θ_i based on Z_i.
3. Unfortunately the OLS line discussed above is an "oracle estimator"; it requires access to the unobserved θ_i. Stigler argued that James-Stein seeks to mimic the oracle OLS line without access to θ_i. The last line of the figure, shown in green, visualizes exactly the line corresponding to James-Stein (that shrinks toward the grand mean). As we can see, James-Stein does a remarkable job of almost tracking the oracle OLS line.

4.2.2 Stigler's formal argument

Stigler (1990) turned the above conceptual argument to a formal argument as follows. For the Baseball example we applied James-Stein that shrinks toward the grand mean. Here instead for simplicity we will make this argument rigorous for James-Stein that shrinks toward 0.[4]

One way to motivate Stigler's argument is through analogy to Proposition 4.1. Therein we derived the optimal choice of λ such that $t_\lambda(\cdot)$ performs well in terms of risk (expected loss).

[4]Stigler (1990) included rigorous proofs for both cases.

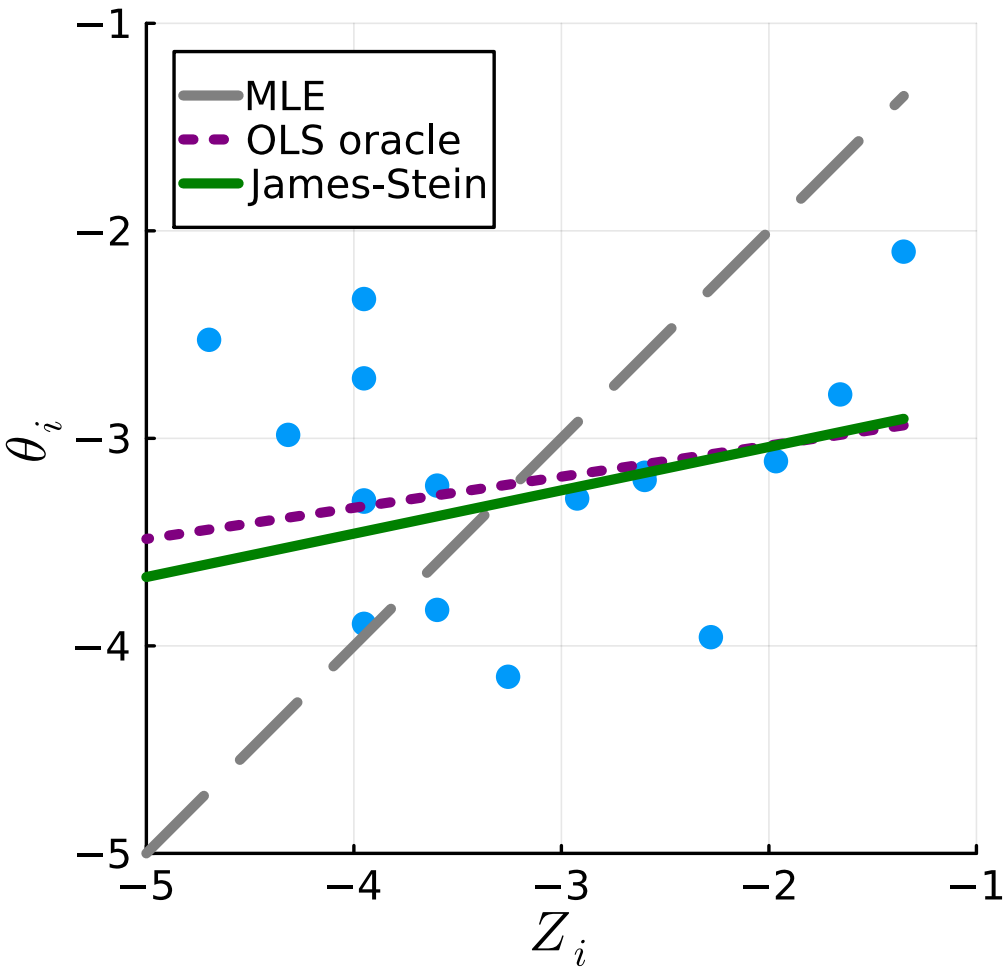


Figure 4.1: Visualization of Stigler's interpretation of James-Stein based on the Efron-Morris baseball dataset.

Instead, we may ask a more ambitious question: what is the λ that minimizes the "in-sample" loss, that is, what is the value of λ such that:

$$\hat{\lambda} \in \operatorname{argmin} \left\{ \frac{1}{n} \sum_{i=1}^{n} (\theta_i - \lambda Z_i)^2 \right\}.$$

The reader will notice that the above is an ordinary linear regression (OLS) problem without intercept, and so:

$$\hat{\beta}^{OLS} := \hat{\lambda} = \frac{\sum_{i=1}^{n} \theta_i Z_i}{\sum_{i=1}^{n} Z_i^2}. \tag{4.4}$$

Above we switched notation from $\hat{\lambda}$ to $\hat{\beta}^{OLS}$ to make the connection to linear regression more explicit. Then Stigler considers estimating $\boldsymbol{\theta}$ by:

$$\hat{\boldsymbol{\theta}}^{OLS} = \hat{\beta}^{OLS} \mathbf{Z}. \tag{4.5}$$

Eq. 4.4 fits an oracle regression; in reality we would never observe θ_i (these are the quantities we seek to estimate). However, Stigler argued that James-Stein essentially seeks to mimic Eq. 4.4.

To this end, notice that we may express the MLE that estimates θ_i by Z_i as follows:

$$\hat{\boldsymbol{\theta}}^{MLE} = b^{MLE} \mathbf{Z}, \quad b^{MLE} := 1.$$

The James-Stein estimates may be written as:

$$\hat{\boldsymbol{\theta}}^{JS} = b^{JS} \mathbf{Z}, \quad b_{JS} := 1 - \frac{n-2}{\sum_{i=1}^{n} Z_i^2}.$$

Stigler's argument boils down to the following: b^{JS} is better at approximating $\hat{\beta}^{OLS}$ than b^{MLE} is. At a heuristic level note that $\mathbb{E}_{\theta_i}[Z_i] = \theta_i$ and that $\mathbb{E}_{\theta_i}[Z_i^2] = 1 + \theta_i^2$, so that:

$$\hat{\beta}^{OLS} = \frac{\sum_{i=1}^{n} \theta_i Z_i}{\sum_{i=1}^{n} Z_i^2} \approx \frac{\sum_{i=1}^{n} \theta_i^2}{\sum_{i=1}^{n} Z_i^2} \approx 1 - \frac{n}{\sum_{i=1}^{n} Z_i^2} \approx b^{JS}.$$

Let us now turn the above intuitions into a second rigorous proof of Theorem 3.1 (in Chapter 3).

Proof. Consider any estimator of the form bZ_i for θ_i, wherein b may depend on all of $(Z_1, \dots, Z_n)$. Then:

$$\begin{aligned}
\|\boldsymbol{\theta} - b\mathbf{Z}\|^2 &= \left\|\boldsymbol{\theta} - \hat{\beta}^{OLS}\mathbf{Z} + \hat{\beta}^{OLS}\mathbf{Z} - b\mathbf{Z}\right\|^2 \\
&= \left\|\boldsymbol{\theta} - \hat{\beta}^{OLS}\mathbf{Z}\right\|^2 + \left\|\hat{\beta}^{OLS}\mathbf{Z} - b\mathbf{Z}\right\|^2 \\
&= \left\|\boldsymbol{\theta} - \hat{\beta}^{OLS}\mathbf{Z}\right\|^2 + (\hat{\beta}^{OLS} - b)^2 \|\mathbf{Z}\|^2.
\end{aligned}$$

Above we used the following two facts:

1. In the hypothetical oracle linear regression of $\theta_i \sim Z_i$, the residuals are orthogonal to the subspace generated by $\mathbf{Z}$.
2. $b\mathbf{Z}$ lies in the linear span of $\mathbf{Z}$.

Taking expectations we thus find that:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\|\boldsymbol{\theta} - b\mathbf{Z}\|^2\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\boldsymbol{\theta} - \hat{\beta}^{OLS}\mathbf{Z}\right\|^2\right] + \mathbb{E}_{\boldsymbol{\theta}}\left[(\hat{\beta}^{OLS} - b)^2 \|\mathbf{Z}\|^2\right],$$

and note that first summand on the RHS does not depend on b. Thus, to show that James-Stein dominates the MLE, it suffices to show that:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[(\hat{\beta}^{OLS} - b^{JS})^2 \|\mathbf{Z}\|^2\right] < \mathbb{E}_{\boldsymbol{\theta}}\left[(\hat{\beta}^{OLS} - 1)^2 \|\mathbf{Z}\|^2\right].$$

The above is equivalent to showing that:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[(1 - b^{JS})(2\hat{\beta}^{OLS} - 1 - b^{JS}) \|\mathbf{Z}\|^2\right] < 0.$$

In turn note that $1 - b^{JS} = (n-2)/\|\mathbf{Z}\|^2$, so that the above is equivalent to:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[2\hat{\beta}^{OLS} - 1 - b^{JS}\right] \leq 0.$$

Plugging in the definitions and rearranging, we finally can see that it suffices to prove the following:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{2\sum_{i=1}^n \theta_i Z_i \,+\, (n-2)}{\|\mathbf{Z}\|^2}\right] < 2.$$

We will show this through Stein's identity. Let $h_i(\mathbf{Z}) = \frac{Z_i}{\|\mathbf{Z}\|^2}$, so that:

$$\frac{\partial h_i(\mathbf{Z})}{\partial Z_i} = \frac{1}{\|\mathbf{Z}\|^2} - \frac{2Z_i^2}{\|\mathbf{Z}\|^4}.$$

Stein's identity yields:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[(Z_i - \theta_i)h_i(\mathbf{Z})\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial h_i(\mathbf{Z})}{\partial Z_i}\right],$$

which means that:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{Z_i^2}{\|\mathbf{Z}\|^2} - \frac{\theta_i Z_i}{\|\mathbf{Z}\|^2}\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{\|\mathbf{Z}\|^2} - \frac{2Z_i^2}{\|\mathbf{Z}\|^4}\right].$$

Summing over $i = 1, \dots, n$:

$$1 - \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\sum_{i=1}^n \theta_i Z_i}{\|\mathbf{Z}\|^2}\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{(n-2)}{\|\mathbf{Z}\|^2}\right].$$

Hence:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{2\sum_{i=1}^n \theta_i Z_i \,+\, (n-2)}{\|\mathbf{Z}\|^2}\right] < 2\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\sum_{i=1}^n \theta_i Z_i \,+\, (n-2)}{\|\mathbf{Z}\|^2}\right] = 2.$$

□

4.3 James-Stein shrinkage with side-information

So far in this chapter we have posited that $Z_i \sim \mathcal{N}(\theta_i, 1)$. Suppose that for the i-th unit we do not only observe Z_i, but also side-information $X_i \in \mathcal{X}$. In the compound setting we think of X_i as being fixed (not random), and in the (empirical) Bayes setting we think of X_i as being independent of Z_i conditionally on θ_i.

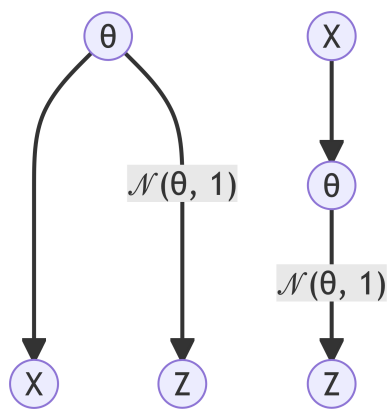


Figure 4.2: Two hypothetical DAGs representing empirical Bayes with side-information

Crucially above both Z_i and X_i can contain information about θ_i. While we understand the distribution $Z_i \mid \theta_i$ we will seek methods that work without explicit requirements on the relationship of θ_i and X_i. We mention two possible baselines:

1. **Pure empirical Bayes:** We ignore the information in X_i and apply our favorite empirical Bayes method based on $Z_1, \ldots, Z_n$ so as to match the performance of $\mathbb{E}[\theta_i \mid Z_i]$.
2. **Pure predictive modeling:** We seek the best prediction based on only covariates. This is given by $m(X_i) = \mathbb{E}[\theta_i \mid X_i]$. Note that we can estimate this based on data: $\mathbb{E}[Z_i \mid X_i] = \mathbb{E}[\theta_i \mid X_i]$ and so we could regress $Z_i \sim X_i$ using our favorite supervised learning method to learn $m(\cdot)$.

Both of the above approaches however leave information on the table. Below we will study methods that can extract information from both Z_i and X_i, but before doing so, we will consider two applications.

4.3.1 Examples of applications for shrinkage with side-information

Census bureau in the 1970s: The Census Bureau at the time was interested in estimating the per-capita income θ_i in 39,000 units of local government ("small areas") based on surveys

of 20% of the population. Let us denote the sample average income in the i-th area by Z_i. Then, for the 1970 census, the Census Bureau proceeded as follows: for any area i with a population with 500 people or more, it reported Z_i as the estimate of θ_i, while for any small area with population below 500 people, it reported the average income of the county to which the area belonged. The motivation was based on a bias-variance tradeoff consideration: with less than 500 people, Z_i would be extremely noisy, so the Bureau preferred a biased but less noisy estimate.

Fay III and Herriot (1979) realized the potential here for James-Stein to outperform the above heuristic approach. Empirical Bayes could provide a principled way for combining both sources of information Z_i and data for the county. Their principled approach did not require setting e.g., an arbitrary cutoff at a population of 500. In seeking to solve the above problem, Fay III and Herriot (1979) developed an approach to James-Stein shrinkage with side-information: they modeled $Z_i \sim \mathcal{N}(\theta_i,\ 1)$ and also sought to include side-information X_i that included the per-capita income of the whole county, IRS data, and housing data.

In fact, their method was implemented as part of the 1974 census; Fay III and Herriot (1979) write the following:

> "Because of the mathematical and logical consistency of the revised procedures, and on the basis of independent empirical evidence, the Census Bureau has used this methodology in forming the estimates for 1974 and subsequent years. To our knowledge, the Census Bureau's use is the largest application of James-Stein procedures in a federal statistical program." Fay III and Herriot (1979)

Estimating average movie ratings: Ignatiadis and Wager (2019) use the following setting as an application of empirical Bayes shrinkage with side-information: consider a dataset of movie reviews, e.g., MovieLens (Harper and Konstan 2016), where each movie ($i = 1, \dots, n$) has a given average rating (Z_i) based on a limited number of viewers. Additionally, we have access to various information about each movie (X_i), such as its genre, cast, length, etc. The objective is to estimate the "true" rating (μ_i) of each movie, meaning the average rating it would receive if it was reviewed by a larger number of similar reviewers.

4.3.2 The oracle approach

What is the best we could hope to do in the setting of this section? If we seek to estimate θ_i in mean squared error, the best we can do is of course to use the Bayes predictor:

$$\mathbb{E}\left[\theta_i \mid X_i, Z_i\right]. \tag{4.6}$$

We note in passing that the above object is nonparametrically identified. On the other hand, it is not a standard object, for example, we do not observe θ_i so we could not just apply supervised learning of $\theta_i \sim X_i, Z_i$. Furthermore, we have distributional information about $Z_i \mid \theta_i$ but do not have more information on X_i. Most fully nonparametric estimation strategies for this

problem would be fickle, suffer from the curse of dimensionality if X_i is high-dimensional or structured, and would rely substantially on any assumptions made e.g., about Gaussianity of $Z_i \mid \theta_i$.

Instead we consider an alternative.

4.3.3 A practical model: side-information that modulates the prior mean

One challenge in estimating Eq. 4.6 is that so far, we essentially allow the prior $G(\theta \mid X_i)$ to be modulated by X_i in an arbitrary way. The task simplifies substantially if X_i modulates the prior $G(\theta \mid X_i)$ only through the mean function $m(x) = \mathbb{E}[\theta \mid X = x] = \mathbb{E}[Z \mid X = x]$. To be concrete we consider the following working model that generalizes the empirical Bayes model of Efron and Morris:

$$\begin{aligned} X_i &\overset{\text{iid}}{\sim} \mathbb{P}^X, \\ \theta_i \mid X_i &\overset{\text{iid}}{\sim} \mathcal{N}(m(X_i),\ A), \\ Z_i \mid \theta_i, X_i &\overset{\text{iid}}{\sim} \mathcal{N}(\theta_i, 1). \end{aligned} \tag{4.7}$$

Under Eq. 4.7, the Bayes predictor Eq. 4.6 simplifies as follows:

$$\mathbb{E}[\theta_i \mid X_i = x, Z_i = z] \;=\; \frac{1}{1+A} m(x) \;+\; \frac{A}{1+A} z. \tag{4.8}$$

The equation above has the following elegant interpretation: we take a convex combination of the two sources of information, Z_i and X_i, wherein we use X_i through the optimal regression $m(X_i) = \mathbb{E}[\theta_i \mid X_i]$. Eq. 4.8 also suggests the following procedure:

1. Regress $Z_i \sim X_i$ using our favorite supervised learning method to learn $\widehat{m}(\cdot)$.
2. Learn $\widehat{A}$.
3. Estimate θ_i by $\frac{1}{1+\widehat{A}}\widehat{m}(X_i) + \frac{\widehat{A}}{1+\widehat{A}} Z_i$.

4.3.4 Shrinking towards linear regression

Suppose that $X_i \in \mathbb{R}^p$. We stack the $X_i^\top$ into an $n \times p$ design matrix $\mathbf{X}$.

Suppose that we posit that $m(x)$ in Eq. 4.7 is a linear function of x, that is, $m(x) = \beta^\top x$. This is the setting considered by Fay III and Herriot (1979).[5] Here's how one can proceed:

[5]They do not proceed as we do below; instead they estimated β and A by the method of moments

Let $\hat{\beta}$ be the ordinary least squares coefficients of the regression $Z_i \sim X_i$. Then we could shrink towards the OLS (ordinary least squares) predictions $X_i^\top \hat{\beta}$. This gives rise to the James-Stein-Fay-Herriot estimator:

$$\hat{\boldsymbol{\theta}}_{JS-FH} = \mathbf{X}\hat{\beta} + \left(1 - \frac{n-p-2}{\left\|\mathbf{Z} - \mathbf{X}\hat{\beta}\right\|^2}\right)(\mathbf{Z} - \mathbf{X}\hat{\beta}).$$

We have that:

Theorem 4.2. *The James-Stein-Fay-Herriot estimator dominates the MLE (for any value of* $\mathbf{X}$*) as soon as* $n \geq p+3$*, i.e.,:*

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\hat{\boldsymbol{\theta}}_{JS-FH} - \boldsymbol{\theta}\right\|^2\right] < n.$$

The expectations above hold when we treat $\mathbf{X}$ *as fixed rather than random and* $Z_i \mid \theta_i, X_i \sim \mathcal{N}(\theta_i, 1)$*.*

The following proof is suggested in passing by Jiang and Zhang (2010); in a slightly different context the argument also appears in Brown and Zhao (2009).

Proof. The idea is the following. Let Q_A be the $n \times p$ matrix generated by orthonormalizing the design matrix $\mathbf{X}$. Furthermore let Q_B be the $n \times n-p$ matrix generated by completing the column space of Q_A to all of $\mathbb{R}^n$ so that $Q = [Q_A Q_B]$ is itself orthonormal.[6]

Now let us call $\widetilde{\mathbf{Z}} = Q_B^\top \mathbf{Z} \in \mathbb{R}^{n-p}$. Then:

$$\widetilde{\mathbf{Z}} \sim \mathcal{N}(Q_B^\top \boldsymbol{\theta},\ I_{n-p}).$$

The crucial argument is as simple as follows. We apply James-Stein to $\widetilde{\mathbf{Z}}$! It takes the form:

$$\tilde{\boldsymbol{\theta}}_{JS} = \left(1 - \frac{n-p-2}{\left\|\widetilde{\mathbf{Z}}\right\|^2}\right)\widetilde{\mathbf{Z}}.$$

The dimension of the problem here is $n-p$ and we have that $n-p \geq 3$ by assumption. Thus:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\tilde{\boldsymbol{\theta}}_{JS} - Q_B^\top \boldsymbol{\theta}\right\|\right] < n-p.$$

Also let us note the following. Since $\mathbf{X}\hat{\beta}$ is the projection of $\mathbf{Z}$ to the column space of Q_A, we get:

$$\left\|\mathbf{Z} - \mathbf{X}\hat{\beta}\right\|^2 = \underbrace{\left\|Q_A^\top \mathbf{Z} - Q_A^\top \mathbf{X}\hat{\beta}\right\|^2}_{=0} + \underbrace{\left\|Q_B^\top \mathbf{Z} - Q_B^\top \mathbf{X}\hat{\beta}\right\|^2}_{=\left\|Q_B^\top \mathbf{Z}\right\|^2} = \left\|\widetilde{\mathbf{Z}}\right\|^2.$$

[6] In numerical linear algebra terms: we are taking the QR decomposition of $\mathbf{X}$.

The above shows that:

$$\tilde{\boldsymbol{\theta}}_{JS} = Q_B^\top \left(1 - \frac{n-p-2}{\left\|\mathbf{Z} - \mathbf{X}\widehat{\beta}\right\|}\right) \mathbf{Z}.$$

We are almost ready to conclude:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\hat{\boldsymbol{\theta}}_{JS-FH} - \boldsymbol{\theta}\right\|^2\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\left\|Q^\top \hat{\boldsymbol{\theta}}_{JS-FH} - Q^\top \boldsymbol{\theta}\right\|^2\right] \\
&= \mathbb{E}_{\boldsymbol{\theta}}\left[\left\|Q_A^\top \hat{\boldsymbol{\theta}}_{JS-FH} - Q_A^\top \boldsymbol{\theta}\right\|^2\right] + \mathbb{E}_{\boldsymbol{\theta}}\left[\left\|Q_B^\top \hat{\boldsymbol{\theta}}_{JS-FH} - Q_B^\top \boldsymbol{\theta}\right\|^2\right] \\
&= \underbrace{\mathbb{E}_{\boldsymbol{\theta}}\left[\left\|Q_A^\top \mathbf{Z} - Q_A^\top \boldsymbol{\theta}\right\|^2\right]}_{=p} + \underbrace{\mathbb{E}_{\boldsymbol{\theta}}\left[\left\|\tilde{\boldsymbol{\theta}}_{JS} - Q_B^\top \boldsymbol{\theta}\right\|^2\right]}_{n-p} < n.
\end{aligned}$$

□

4.3.5 Shrinking towards an arbitrary machine learning model

The estimator and theoretical result of Theorem 4.2 is elegant. However, it relies substantially on the fact that we take $m(x)$ to be a linear function of a low-dimensional x. What if we fit $\widehat{m}(\cdot)$ through boosting or through a neural network and e.g., $\mathcal{X}$ consists e.g., of images? Is it possible to come up with a procedure that dominates Z_i (in the sense of James-Stein) when we shrink toward an arbitrary machine learning model (that we may not be able to handle theoretically)?

The starting point of the idea of Ignatiadis and Wager (2019) is the following: suppose we are given a fixed regression function $\widetilde{m}(\cdot)$ that may be misspecified, that is, $\widetilde{m}(\cdot) \neq m(\cdot)$ and we seek to linearly combine $\widetilde{m}(X_i)$ with Z_i to estimate θ_i. How should we proceed? In other words, consider the following class of estimators:

$$\mathcal{L}(\widetilde{m}) := \left\{\hat{\boldsymbol{\theta}}(\lambda) = (\hat{\theta}_1(\lambda), \dots, \hat{\theta}_n(\lambda)),\ \hat{\theta}_i(\lambda) := (1-\lambda)\widetilde{m}(X_i) + \lambda Z_i \ :\ \lambda \in [0,1]\right\}. \tag{4.9}$$

Notice that under Eq. 4.7 and if $\widetilde{m}(\cdot) = m(\cdot)$, then the choice:

$$\lambda^*(A, m) = \frac{A}{A+1},$$

in fact leads to the Bayes decision Eq. 4.8 and so this choice of λ must also give the optimal estimator within the class $\mathcal{L}(m)$; Eq. 4.9. But what if $\widetilde{m}(\cdot) \neq m(\cdot)$?

Proposition 4.3. *Suppose the triples (θ_i, X_i, Z_i) are generated according to Eq. 4.7 (and take all expectations that follow with respect to the randomness in the triple.) Then the following optimization problem over estimators $\hat{\boldsymbol{\theta}}(\lambda) \in \mathcal{L}(\widetilde{m})$ Eq. 4.9,*

$$\min_{\lambda \in [0,1]} \left\{ \mathbb{E}\left[\left\| \hat{\boldsymbol{\theta}}(\lambda) - \boldsymbol{\theta} \right\|^2 \right] \right\},$$

is solved by:

$$\lambda^*(A, \tilde{m}) = \frac{A + \mathbb{E}\left[(m(X) - \widetilde{m}(X))^2\right]}{A + \mathbb{E}\left[(m(X) - \widetilde{m}(X))^2\right] + 1} = \lambda^*(A + \mathbb{E}\left[(m(X) - \widetilde{m}(X))^2\right], m).$$

Proof. Similar to the proof of Proposition 4.1, so omitted.

□

This result says the following: even if knew A in model Eq. 4.7, if $\widetilde{m}(\cdot) \neq m(\cdot)$ then we should still prefer to use a different choice of "A" and not the true A: we should inflate the true prior variance to also account for the out-of-sample mean squared error $\mathbb{E}\left[(m(X) - \widetilde{m}(X))^2\right]$! Furthermore, there is one more upshot:

$$A + \mathbb{E}\left[(m(X) - \widetilde{m}(X))^2\right] = \mathbb{E}\left[(Z - \widetilde{m}(X))^2\right] - 1.$$

The LHS is what we need to know to emulate the optimal rule in $\mathcal{L}(\widetilde{m})$. Furthermore, we may directly estimate the RHS, since it is the out-of-sample prediction error for Z based on $\widetilde{m}$, which we can assess e.g., using cross-validation or a test sample.

This leads to the following algorithm (presented with 2 folds for simplicity but generalizes to K folds), which is called "empirical Bayes with cross-fitting" (ECDF):

1. Form a partition of $\{1, \ldots, n\}$ into two folds I_1 and I_2.
2. Use observations in I_1, to estimate the regression $m(x) = \mathbb{E}\left[Z_i \mid X_i = x\right]$ by $\widehat{m}_{I_1}(\cdot)$ (using any machine learning model).
3. Using observations in I_2, compute:

$$\widehat{A}_{I_2} = \frac{1}{|I_2| - 2} \sum_{i \in I_2} \left(\widehat{m}_{I_1}(X_i) - Z_i \right)^2 - 1.$$

4. For all $i \in I_2$, estimate θ_i by:

$$\hat{\theta}_i^{EBCF} = \frac{1}{\widehat{A}_{I_2} + 1} \widehat{m}_{I_1}(X_i) + \frac{\widehat{A}_{I_2}}{\widehat{A}_{I_2} + 1} Z_i.$$

5. Repeat with folds I_1 and I_2 flipped to also get estimates of $\hat{\theta}_i^{EBCF}$ for $i \in I_1$.

Ignatiadis and Wager (2019) show that the above procedure satisfies the James-Stein property.

Theorem 4.3. *Suppose that conditionally on $(X_i, \theta_i)_i$, the Z_i are jointly independent and are distributed as $\mathcal{N}(\theta_i, 1)$. Suppose further that $|I_1| \geq 3$ and $|I_2| \geq$, then:*

$$\mathbb{E}_{\boldsymbol{\theta}, \mathbf{X}} \left[\left\| \hat{\boldsymbol{\theta}}^{EBCF} - \boldsymbol{\theta} \right\|^2 \right] < n.$$

Proof. The idea is the following. It suffices to prove that:

$$\sum_{i \in I_j} \mathbb{E}_{\boldsymbol{\theta}, \mathbf{X}} \left[\left(\hat{\boldsymbol{\theta}}_i^{EBCF} - \theta_i \right)^2 \right] < \left| I_j \right|, \; j = 1, 2. \tag{4.10}$$

This follows directly from our existing results! For example, take the second fold I_2. Then $(\hat{\theta}_i^{EBCF})_{i \in I_2}$ is equal to the following estimator: we apply James-Stein to $(Z_i)_{i \in I_2}$ but shrink toward the location vector $(\widehat{m}_{I_1}(X_i))_{i \in I_2}$. The latter depends on $(X_i)_{1 \leq i \leq n}$ and $(Z_i)_{i \in I_2}$, hence if we condition on all the X_i as well as the Z_i in I_1, then we may treat $\widehat{m}_{I_1}(X_i))_{i \in I_2}$ as a fixed location vector. Hence, since $|I_2| \geq 3$, Eq. 4.10 for $j = 2$ follows from the results in Section 3.4.1, Chapter 3.

□

4.4 Shrinkage in the heteroscedastic problem

Throughout these notes—even when we set forth the definition of parallel statistical decision procedures in Definition 1.2 (Chapter 1)—we have assumed that the likelihood, $p(\cdot \mid \theta_i)$ remains the same for all simple statistical problems under consideration. For example, in the Gaussian problem we have assumed that $Z_i \mid \theta_i \sim \mathcal{N}(\theta_i, 1)$. In many applications, it is more reasonable to suppose instead that:

$$Z_i \mid \theta_i \sim \mathcal{N}(\theta_i, \sigma_i^2), \tag{4.11}$$

where σ_i^2 varies with i.[7] As one example, in the Baseball example of Efron and Morris, we only sought to predict the batting average of players with 45 at bats. If we want to make predictions for all players, then they will have different at bats, and so, the precision of our initial measurement for each player will vary.

In general, there are a few more subtleties involved in the heteroscedastic problem. For example, suppose one is interested in compound estimation of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$. What compound loss function should one use? One could use e.g., squared error as before,

$$\ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} (\hat{\theta}_i - \theta_i)^2, \tag{4.12}$$

[7] Or more generally, to suppose that $Z_i \mid \theta_i \sim p_i(\cdot \mid \theta_i)$.

or one could also weight each error by the corresponding precisions, that is:

$$\ell(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\sigma_i^2} (\hat{\theta}_i - \theta_i)^2. \tag{4.13}$$

The latter loss represents the desire to downweight errors for units i that were observed with a lot of noise.

4.4.1 Precision-weighted squared error loss

The loss in Eq. 4.13 turns out to be easier to handle! Let $\tilde{Z}_i = Z_i/\sigma_i$ and $\tilde{\theta}_i = \theta_i/\sigma_i$. Then:

$$\tilde{Z}_i \mid \tilde{\theta}_i \sim \mathcal{N}(\tilde{\theta}_i, 1),$$

i.e., we are back to the homoscedastic setting. If we apply e.g., James-Stein on $\tilde{Z}_i$ and then transform back to the original scale (by multiplication by σ_i), then we target precisely the loss in Eq. 4.13.

4.4.2 Squared error loss

Eq. 4.12 is more difficult to handle—and in fact even in the simple Gaussian setting there is no clear cut answer how to best conduct empirical Bayes shrinkage. One proposal that works well is due to Xie, Kou, and Brown (2012). They generalize the ideas we set forth in Section 4.1 to the heteroscedastic setting.[8]

The idea is the following. First suppose that we posit that:

$$\theta_i \sim \mathcal{N}(0,\, A),\;\; Z_i \mid \sigma_i \sim \mathcal{N}(\theta_i, 1).$$

Then:

$$\mathbb{E}\left[\theta_i \mid \sigma_i, Z_i\right] = \frac{A}{A + \sigma_i^2} Z_i. \tag{4.14}$$

Xie, Kou, and Brown (2012) use the above equation to define the class of estimators that we get as we vary A in Eq. 4.14. Hence:

$$\mathcal{L} := \left\{ \hat{\boldsymbol{\theta}}(A) \,:\, A \geq 0 \right\}, \;\; \hat{\theta}_i(A) := \frac{A}{A + \sigma_i^2} Z_i. \tag{4.15}$$

They then choose an estimator from the class $\mathcal{L}$ by minimizing SURE. This requires a slight modification of SURE that we learned in last class to Gaussian noise with variance σ^2 that is not necessarily equal to 1.

[8] In fact the causality is backwards: our treatment in these lecture notes for the homoscedastic case was inspired by Xie, Kou, and Brown (2012) (as well as Kou and Yang (2017)).

Definition 4.1 (Stein's unbiased risk estimator (heteroscedastic case))**.** Suppose that $\hat{\boldsymbol{\theta}} \equiv h(\mathbf{Z}) = (\hat{\theta}_1, \dots, \hat{\theta}_n)$ is an almost differentiable function of $\mathbf{Z}$. Further suppose that Z_i are independent and $Z_i \sim \mathcal{N}(\theta_i, \sigma_i^2)$. Then:

$$\hat{R} := \sum_{i=1}^{n} \left(-\sigma_i^2 + (Z_i - \hat{\theta}_i)^2 + 2\sigma_i^2 \frac{\partial \hat{\theta}_i}{\partial z_i}(\mathbf{Z}) \right), \tag{4.16}$$

is unbiased for the mean squared error in estimating θ_i, that is,

$$\mathbb{E}_{\boldsymbol{\theta}}[\hat{R}] = \mathbb{E}_{\boldsymbol{\theta}} \left[\left\| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right\|^2 \right]. \tag{4.17}$$

Xie, Kou, and Brown (2012) apply the above to choose an estimator from the class Eq. 4.15. Concretely, they let:

$$\widehat{A} \in \operatorname*{argmin}_{A \geq 0} \left\{ \sum_{i=1}^{n} \left(\sigma_i^2 + \frac{\sigma_i^4}{(A + \sigma_i^2)^2} Z_i^2 - 2 \frac{\sigma_i^4}{A + \sigma_i^2} \right) \right\},$$

and then estimate $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}(\widehat{A})$.

5 Empirical Bayes intervals and confidence sets

In this chapter we explain how to quantify uncertainty in empirical Bayes settings by reporting "empirical Bayes confidence intervals." This is a task that has received less attention compared to empirical Bayes for e.g., shrinkage estimation.

5.1 Preliminaries

5.1.1 Intervals for simple statistical decision problems

To introduce definitions, we start by studying a generic simple statistical decision problem (e.g., as set forth in Definition 1.1 in Chapter 1). We have a parameter $\theta \in \Theta \subset \mathbb{R}$ and observe a single $Z \mid \theta \sim p(\cdot \mid \theta)$. Our decision space $\mathcal{T}$ consists of outputting an interval $\mathcal{I} := \mathcal{I}(Z)$ or a set $\mathcal{S}$.[1]

Let us also suppose that $\theta \sim G$. A notion of coverage is given as follows:

Definition 5.1 (Marginal coverage interval)**.** An interval $\mathcal{I} := \mathcal{I}(Z)$ is said to have marginal coverage for the parameter θ when:

$$\mathbb{P}_G\left[\theta \in \mathcal{I}(Z)\right] \geq 1 - \alpha. \tag{5.1}$$

It is important to note that randomness in Eq. 5.1 is with respect to both the randomness in θ and in $Z \mid \theta$.

Below we provide two examples of intervals satisfying Eq. 5.1.

Definition 5.2 (Frequentist confidence interval)**.** An interval $\mathcal{I}$ is called a frequentist $(1-\alpha)$-confidence interval for θ if:

$$\mathbb{P}_\theta\left[\theta \in \mathcal{I}\right] \geq 1 - \alpha \quad \text{for all} \quad \theta \in \Theta.$$

[1] We will typically present results for $\mathcal{I}$ but most definitions and concepts translated directly to more general sets $\mathcal{S}$.

Proposition 5.1. *A frequentist confidence interval is also an interval with marginal coverage as in Eq. 5.1.*

Proof. Let $\mathcal{J}$ be a frequentist $(1-\alpha)$-confidence interval. Then:

$$\mathbb{P}_G\left[\theta \in \mathcal{J}\right] = \mathbb{E}_G\left[\mathbb{P}_G\left[\theta \in \mathcal{J} \mid \theta\right]\right] = \mathbb{E}_G[\underbrace{\mathbb{P}_\theta\left[\theta \in \mathcal{J}\right]}_{\geq 1-\alpha}] \geq 1-\alpha.$$

□

Example 5.1 (Confidence interval for a Gaussian mean)**.** Suppose that $Z \sim \mathcal{N}(\theta, 1)$. Then the "textbook" $1-\alpha$ confidence interval for θ is given by:

$$\mathcal{J}(z) = z \pm q_{1-\alpha/2},$$

where $q_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

Definition 5.3 (Credible interval)**.** An interval $\mathcal{J}$ is called an $(1-\alpha)$-credible interval for θ if:

$$\mathbb{P}_G\left[\theta \in \mathcal{J} \mid Z\right] \geq 1-\alpha \text{ almost surely.}$$

Proposition 5.2. *A credible interval is also an interval with marginal coverage as in Eq. 5.1.*

Proof. Let $\mathcal{J}$ be an $(1-\alpha)$-credible interval. Then:

$$\mathbb{P}_G\left[\theta \in \mathcal{J}\right] = \mathbb{E}_G[\underbrace{\mathbb{P}_G\left[\theta \in \mathcal{J} \mid Z\right]}_{\geq 1-\alpha \text{ a.s.}}] \geq 1-\alpha.$$

□

Example 5.2 (Credible interval for a Gaussian mean)**.** Suppose that $Z \sim \mathcal{N}(\theta, 1)$ and further suppose that $\theta \sim \mathcal{N}(0, A)$. Then recall that

$$\theta \mid Z \sim \mathcal{N}\left(\frac{A}{A+1}Z, \frac{A}{A+1}\right),$$

and so a credible intervals may be constructed by forming the interval whose left, (resp. right) end-point is the $\alpha/2$ (resp. $1-\alpha/2$) quantile of the posterior distribution, i.e.,

$$\mathcal{J}(z) = \frac{A}{A+1}z \pm \sqrt{\frac{A}{A+1}} \cdot q_{1-\alpha/2},$$

where $q_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution.

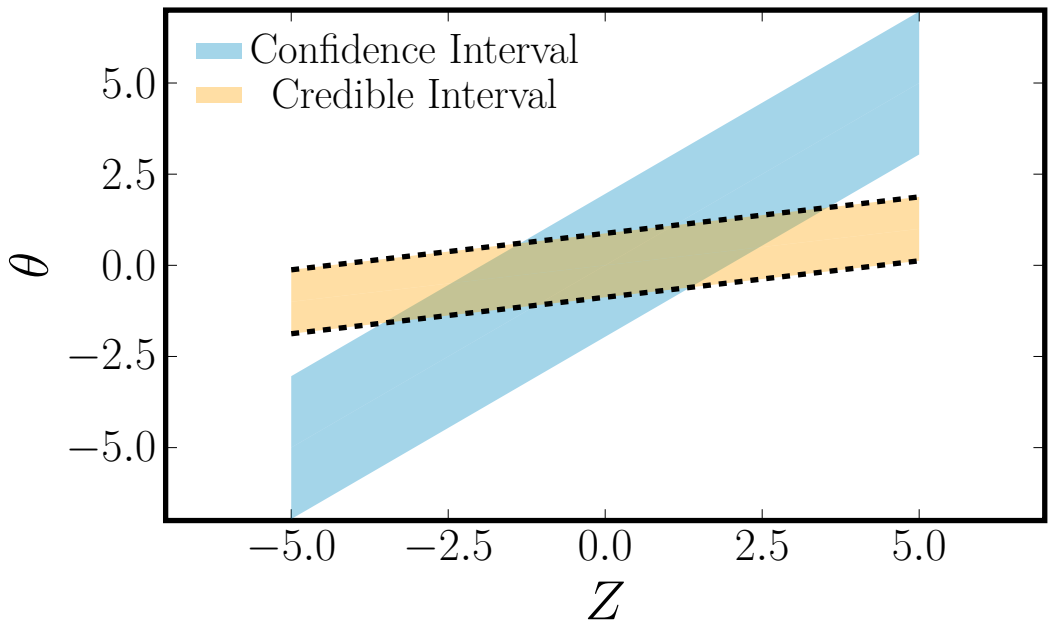


Figure 5.1: Confidence interval and credible interval for θ when $Z \sim \mathcal{N}(\theta, 1)$ and $\theta \sim \mathcal{N}(0,\, 0.25)$.

We illustrate the ideas above with a simple example. We consider the Gaussian setting considered in the above examples and take the prior $G = \mathcal{N}(0,\, 0.25)$.

Both types of intervals satisfy $\mathbb{P}_G\left[\theta \in \mathcal{I}(Z)\right] = 1 - \alpha$. However, their frequentist coverage $\mathbb{P}_\theta\left[\theta \in \mathcal{I}(Z)\right]$ and $\mathbb{P}_G\left[\theta \in \mathcal{I}(Z) \mid Z\right]$ can be substantially different as a function of θ, resp. Z.

Which of these two cases are we happier with? It depends!

> "This suggests caution when using empirical Bayes posterior intervals: While these intervals approximately maintain a target frequentist coverage rate on average across groups, the coverage can be quite poor for outlying groups, which in the examples considered include students with low test-taking ability, or counties with high levels of household radon, which are likely the groups of highest concern." Hoff (2022)

> "Applying our method, we find considerable overestimation of the effect and undercoverage of the confidence interval when the z-value exceeds 1.5. These are serious issues which are undoubtedly a part of the explanation for the phenomenon of poor replication." Zwet, Schwab, and Senn (2021)

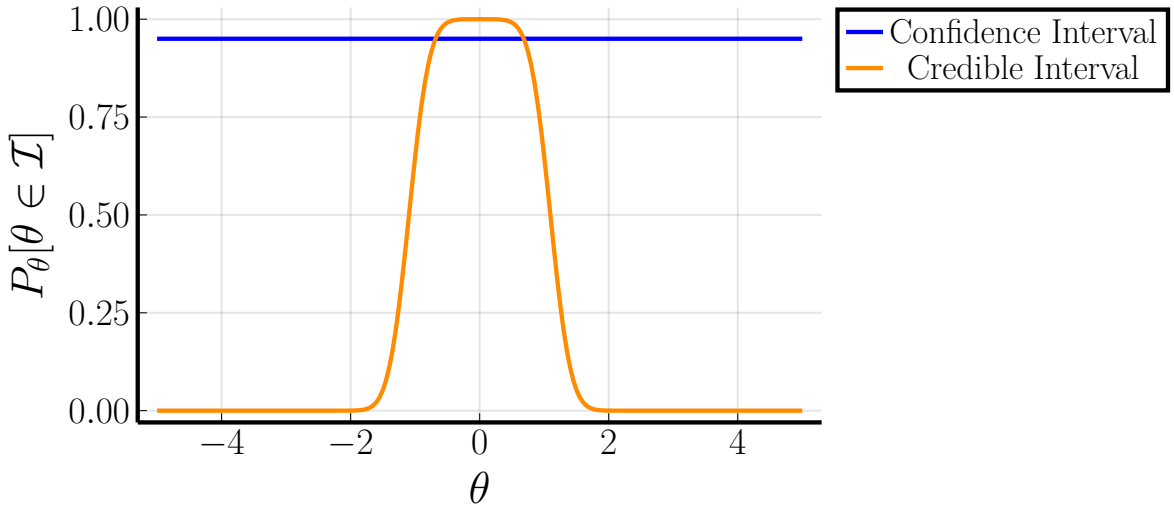


Figure 5.2: Frequentist coverage of confidence and credible intervals as a function of θ.

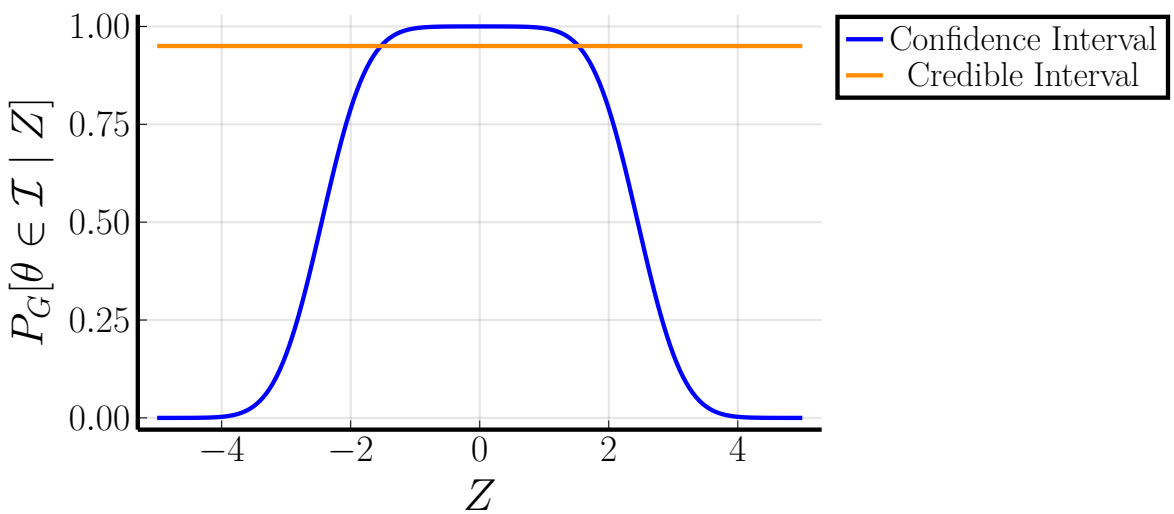


Figure 5.3: Coverage conditionally on Z of confidence and credible intervals.

5.1.2 Optimal Bayesian intervals for simple statistical decisions

In analogy to our discussion Chapter 1, we may seek to construct confidence intervals with some notion of optimality depending on our choice of loss function. Throughout we assume that $\theta \sim G$ and we take the loss of the interval to be the interval length:

$$\ell(\mathcal{I}, \theta) = |\mathcal{I}| \,.$$

We seek to construct $\mathcal{I}$ such that $\mathbb{E}_G\left[|\mathcal{I}|\right]$ is minimal. We also need to put an additional constraint on $\mathcal{I}$.

Optimality among credible interval: One option would be to require that $\mathcal{I}$ is a credible interval as in Definition 5.3. In that case we could seek to find the interval $\mathcal{I}$ such that:

$$\begin{aligned} \mathcal{I}^*(\cdot) \in \operatorname*{argmin}_{\mathcal{I}(\cdot)} \quad & \mathbb{E}_G\left[|\mathcal{I}(Z)|\right] \\ \text{s.t.} \quad & \mathbb{P}_G\left[\theta \in \mathcal{I}(Z) \mid Z = z\right] \geq 1 - \alpha \;\text{ for all }\; z \in \mathcal{Z}. \end{aligned} \tag{5.2}$$

Note that by an argument analogous to that of Proposition 1.2 in Chapter 1, the above interval can be computed pointwise for each z by finding $\mathcal{I}(z)$ with the following properties:

$$\mathcal{I}(z) \in \operatorname*{argmin}_{\mathcal{I} \text{ interval}} \left\{\mathbb{E}_G\left[|\mathcal{I}| \mid Z = z\right] \;:\; \mathbb{P}_G\left[\theta \in \mathcal{I} \mid Z = z\right] \geq 1 - \alpha\right\}.$$

Exercise 5.1 (Optimality of Gaussian-Gaussian credible interval)**.** Prove that the credible interval in Example 5.2 is optimal in the sense of Eq. 5.2.

Optimality among marginal coverage intervals: Next we may instead only constrain $\mathcal{I}$ to have marginal coverage as in Eq. 5.1. By Proposition 5.2, this constraint is less restrictive than requiring $\mathcal{I}$ to be a credible interval and so this assumption could in principle lead to smaller optimal expected interval length.

$$\begin{aligned} \mathcal{I}^*(\cdot) \in \operatorname*{argmin}_{\mathcal{I}(\cdot)} \quad & \mathbb{E}_G\left[|\mathcal{I}(Z)|\right] \\ \text{s.t.} \quad & \mathbb{P}_G\left[\theta \in \mathcal{I}(Z)\right] \geq 1 - \alpha. \end{aligned} \tag{5.3}$$

Proposition 5.3. *Fix $\lambda > 0$. Suppose that for any $z \in \mathcal{Z}$, $\mathcal{I}(z)$ solves the following optimization problem:*

$$\mathcal{I}(z) \in \operatorname*{argmin}_{\mathcal{I} \text{ interval}} \left\{|\mathcal{I}| - \lambda \mathbb{P}_G\left[\theta \in \mathcal{I} \mid Z = z\right]\right\}. \tag{5.4}$$

Then, $\mathcal{I}(\cdot)$ is the optimal marginal coverage interval in the sense Eq. 5.3 for,

$$1 - \alpha = \mathbb{P}_G\left[\theta \in \mathcal{I}(Z)\right].$$

Proof. First, by iterated expectation we have by Eq. 5.4 that $\mathcal{J}(\cdot)$ solves the following optimization problem:

$$\mathcal{J}(\cdot) \in \operatorname*{argmin}_{\mathcal{J}(\cdot)} \left\{ \mathbb{E}_G \left[|\mathcal{J}| \right] - \lambda \mathbb{P}_G \left[\theta \in \mathcal{J} \right] \right\}.$$

But the objective is merely the Lagrangian of the constrained objective Eq. 5.3.

□

Example 5.3 (Optimal Gaussian-Gaussian marginal coverage interval)**.** Consider the credible interval described in Example 5.2. In the setting of the same example, this interval in fact is also the optimal marginal coverage interval in the sense of Eq. 5.3.

Proof. Let us fix $\lambda > 0$ and let us solve Eq. 5.4 for any fixed value of z. It will be convenient to parameterize the interval $\mathcal{J}$ as $(\mathbb{E}_G [\theta \mid Z = z] + c) \pm \chi$ for some $c \in \mathbb{R}$, $\chi \geq 0$. Then we seek to minimize over c, χ the objective:

$$\Psi(c, \chi) := 2\chi - \lambda \mathbb{P}_G \left[|\theta - \mathbb{E}_G [\theta \mid Z = z] - c| \leq \chi \mid Z = z \right].$$

In fact, let us attempt to first optimize over c (for fixed χ). That is, we seek to maximize, the following:

$$\begin{aligned} \mathbb{P}_G \left[|\theta - \mathbb{E}_G [\theta \mid Z = z] - c| \leq \chi \mid Z = z \right] &= \mathbb{P}_G \left[-\chi + c \leq \theta - \mathbb{E}_G [\theta \mid Z = z] \leq \chi + c \mid Z = z \right] \\ &= \Phi_{A/(A+1)}(\chi - c) - \Phi_{A/(A+1)}(-\chi - c), \end{aligned}$$

where we write Φ_u for the CDF of a Gaussian with variance u and φ_u for the pdf. By taking derivatives, we immediately find that the above is optimized for $c = 0$. Hence it remains to find χ that minimizes:

$$2\chi - \lambda \left(\Phi_{A/(A+1)}(\chi) - \Phi_{A/(A+1)}(-\chi) \right),$$

and so it suffices to minimize:

$$2\chi - \lambda \cdot 2\Phi_{A/(A+1)}(\chi).$$

By first order optimality, the optimal χ is either equal to 0, or satisfies:

$$2 = 2\lambda \varphi_{A/(A+1)}(\chi).$$

Hence let us take

$$\lambda = 1 \Big/ \varphi_{A/(A+1)}(\sqrt{A/(A+1)} q_{1-\alpha/2}),$$

then the optimal $\chi = \sqrt{A/(A+1)} q_{1-\alpha/2}$ and so:

$$\mathcal{J}(z) = \mathbb{E}_G [\theta \mid Z = z] \pm \sqrt{\frac{A}{A+1}} q_{1-\alpha/2}.$$

By Proposition Proposition 5.3, this interval is the optimal $1 - \mathbb{P}_G [\theta \in \mathcal{J}(Z)]$ marginal coverage interval. However, note that, $1 - \mathbb{P}_G [\theta \in \mathcal{J}(Z)] = 1 - \alpha$ and so we conclude.

□

5.1.3 Intervals with parallel simple decision problems

Now let us turn to the setting of parallel simple decisions, that is, for $i = 1, \dots, n$, we observe $Z_i \mid \theta_i \sim p(\cdot \mid \theta_i)$.

Definition 5.4. An empirical Bayes interval is defined as any interval-valued mapping

$$\mathcal{J} := \mathcal{J}(Z_1, \dots, Z_n) \equiv \mathcal{J}(\mathbf{Z}).$$

Intervals as in Definition 5.4 are formed for three main purposes.

Empirical Bayes intervals for individual latent parameters: First, one may be interested in covering the (latent) parameter θ_i. If that is the case, we write $\mathcal{J}_i$ for the interval; the subscript i explicitly denotes that our objective is to construct $\mathcal{J}_i$ such that $\theta_i \in \mathcal{J}_i$.

What should our notion of coverage be? Below we provide three notions that have appeared in the literature.

1. **Empirical Bayes coverage for parameter θ_i:** We seek the coverage property

$$\mathbb{P}_G \left[\theta_i \in \mathcal{J}_i(\mathbf{Z})\right] \geq 1 - \alpha. \tag{5.5}$$

   This definition is similar to Definition 5.1, however, randomness is also taken with respect to θ_j, Z_j for $j \neq i$.

2. **Compound (average) coverage for parameters θ_i:** Suppose we form an interval $\mathcal{J}_i$ for each parameter θ_i. A compound (frequentist) notion of coverage is that of compound–average coverage:

$$\frac{1}{n} \sum_{j=1}^{n} \mathbb{P}_{\boldsymbol{\theta}} \left[\theta_j \in \mathcal{J}_j(\mathbf{Z})\right] \geq 1 - \alpha. \tag{5.6}$$

   Note that the above statement treats $\boldsymbol{\theta}$ as fixed and is analogous, to e.g., compound results we have shown for the James-Stein estimator.

3. **Frequentist coverage for parameter θ_i:** Finally, we may require frequentist coverage of a single parameter θ_i, that is, we may require:

$$\mathbb{P}_{\boldsymbol{\theta}} \left[\theta_i \in \mathcal{J}_i(\mathbf{Z})\right] \geq 1 - \alpha. \tag{5.7}$$

   Similar to Eq. 5.6 we treated $\boldsymbol{\theta}$ as fixed, however, instead of requiring coverage on average, we require coverage for θ_i.

Empirical Bayes sets for all latent parameters: One may be interested in constructing a set $\mathcal{S}(\mathbf{Z}) \in \Theta^n$ such that:

$$\mathbb{P}_G \left[\boldsymbol{\theta} \in \mathcal{S}(\mathbf{Z})\right] \geq 1 - \alpha, \quad \text{or} \quad \mathbb{P}_{\boldsymbol{\theta}} \left[\boldsymbol{\theta} \in \mathcal{S}(\mathbf{Z})\right] \geq 1 - \alpha.$$

Can empirical Bayes help for this task?

Intervals for empirical Bayes estimands and Bayes decisions: A further purpose for forming empirical Bayes intervals is to cover an empirical Bayes estimand that takes the form of Bayes optimal decision $t_G(z)$, e.g., $t_G(z) = \mathbb{E}_G[\theta \mid Z = z]$ for fixed z. In this case we may require coverage of $t_G(z)$:

$$\mathbb{P}_G[t_G(z) \in \mathcal{I}(\mathbf{Z})] \geq 1 - \alpha.$$

5.2 Cox's empirical Bayes confidence intervals for a latent parameter

For simplicity, let us revisit the Gaussian-Gaussian example wherein: $\theta_i \overset{\text{iid}}{\sim} \mathcal{N}(0, A)$ for $A > 0$ and $Z_i \mid \theta_i \sim \mathcal{N}(\theta_i, 1)$.

Cox (1975) proposed the following empirical Bayes confidence intervals for θ_i:

$$\mathcal{I}_i^{Cox}(\mathbf{Z}) = \frac{\hat{A}}{\hat{A}+1} Z_i \pm \sqrt{\frac{\hat{A}}{\hat{A}+1}} \cdot q_{1-\alpha/2}, \ \hat{A} = \max\left\{\frac{1}{n}\sum_{i=1}^n Z_i^2 - 1, 0\right\}. \tag{5.8}$$

Theorem 5.1. *Under the above model assumptions, i.e., $\theta_i \overset{iid}{\sim} \mathcal{N}(0, A)$ for $A > 0$ and $Z_i \mid \theta_i \sim \mathcal{N}(\theta_i, 1)$, it holds for any fixed i that:*

$$\lim_{n\to\infty} \mathbb{P}\left[\theta_i \in \mathcal{I}_i^{Cox}(\mathbf{Z}_{1:n})\right] = 1 - \alpha.$$

Proof. For simplicity we prove coverage for a slight variant of Eq. 5.8, which is also due to Cox (1975). We estimate $\hat{A}$ in Eq. 5.8 in a leave-one-out-fashion, that is:

$$\mathcal{I}_i^{Cox}(\mathbf{Z}) = \frac{\hat{A}_{-i}}{\hat{A}_{-i}+1} Z_i \pm \sqrt{\frac{\hat{A}_{-i}}{\hat{A}_{-i}+1}} \cdot q_{1-\alpha/2}, \ \hat{A}_{-i} = \max\left\{0, \frac{1}{n-1}\sum_{j\neq i} Z_j^2 - 1\right\}. \tag{5.9}$$

Let us introduce the notation $w = A/(A+1)$ and also let $v \geq 0$ be another fixed number. Then:

$$\begin{aligned}
\mathbb{P}\left[\theta_i \in vZ_i \pm \sqrt{v} q_{1-\alpha/2}\right] &= \mathbb{P}\left[-\sqrt{v} q_{1-\alpha/2} \leq \theta_i - vZ_i \leq \sqrt{v} q_{1-\alpha/2}\right] \\
&= \mathbb{P}\left[-\sqrt{\frac{v}{w}} q_{1-\alpha/2} + \frac{v-w}{\sqrt{w}} Z_i \leq \frac{\theta_i - wZ_i}{\sqrt{w}} \leq \sqrt{\frac{v}{w}} q_{1-\alpha/2} + \frac{v-w}{\sqrt{w}} Z_i\right] \\
&= \Phi\left(\sqrt{\frac{v}{w}} q_{1-\alpha/2} + \frac{v-w}{\sqrt{w}} Z_i\right) - \Phi\left(\sqrt{\frac{v}{w}} q_{1-\alpha/2} + \frac{v-w}{\sqrt{w}} Z_i\right).
\end{aligned}$$

Hence, returning to our original goal, let us call $\hat{w}_{-i} = \hat{A}_{-i}/(1+\hat{A}_{-i})$.

$$\begin{aligned}
\mathbb{P}\left[\theta_i \in \mathcal{I}_i^{Cox}(\mathbf{Z})\right] &= \mathbb{E}\left[\mathbb{P}\left[\theta_i \in \mathcal{I}_i^{Cox}(\mathbf{Z}) \mid \hat{A}_{-i}\right]\right] \\
&= \mathbb{E}\left[\Phi\left(\sqrt{\frac{\hat{w}_{-i}}{w}}q_{1-\alpha/2} + \frac{\hat{w}_{-i}-w}{\sqrt{w}}Z_i\right) - \Phi\left(\sqrt{\frac{\hat{w}_{-i}}{w}}q_{1-\alpha/2} + \frac{\hat{w}_{-i}-w}{\sqrt{w}}Z_i\right)\right] \\
&\to \mathbb{E}\left[\Phi(q_{1-\alpha/2}) - \Phi(-q_{1-\alpha/2})\right] = 1-\alpha.
\end{aligned}$$

In the last line, we used the fact that for fixed i, $\hat{A}_{-i}$ converges to A almost surely, i.e., $\hat{w}_{-i}$ converges to w almost surely, along with dominated convergence.

□

We make the following remarks.

1. The above result is only asymptotic. For small n it may not perform well because it is essentially ignoring the uncertainty in estimating A. A lot of the literature focused on improving the properties of Eq. 5.8 for small n. Solutions have included analytic corrections (Cox 1975; Yoshimori and Lahiri 2014), approximations inspired by hierarchical Bayes (Morris 1983; N. M. Laird and Louis 1987), as well as bootstrap methods (N. M. Laird and Louis 1987; B. Efron 1987; Carlin and Gelfand 1991). For example, Yoshimori and Lahiri (2014) construct a second-order efficient interval that satisfies $\mathbb{P}\left[\theta_i \in \mathcal{I}_i(\mathbf{Z})\right] = 1-\alpha+O(n^{-3/2})$.

2. Our result depended on the parametric assumption that $\theta_i \overset{\text{iid}}{\sim} \mathcal{N}(0, A)$. What about the case wherein this assumption is violated? For example, suppose that $\theta_i \sim G$ where G is such that $\mathbb{E}_G\left[\theta_i^2\right] = A$. Do intervals such as Eq. 5.8 have any guarantees in that case? As a comparison, recall that in the case of shrinkage, even though James-Stein still had the interpretation of mimicking the best linear estimator of θ_i. Unfortunately, the answer is negative. As an example, take $\alpha = 0.05$, then the coverage can be as low as 74%.[2]

5.3 Robust empirical Bayes confidence intervals

As mentioned in the last remark above, the Cox-EB intervals rely heavily on the normality of the prior. Can we modify these to lead to valid marginal coverage when $\theta_i \overset{\text{iid}}{\sim} G$ with $\mathbb{E}_G\left[\theta_i^2\right] = A$? Armstrong, Kolesár, and Plagborg-Møller (2022) present an ingenious construction with this property.

[2] More generally, Armstrong, Kolesár, and Plagborg-Møller (2022) prove that at a fixed α, the coverage can be as low as $1 - 1/\max(q_{1-\alpha/2}^2, 1)$.

Their idea is the following. Instead of seeking to mimick the oracle credible interval in Example 5.2, they propose to consider "oracle" intervals of the following form:

$$\mathcal{J}(Z;\chi,A) = \frac{A}{A+1}Z + \chi\sqrt{\frac{A}{A+1}}.$$

The above interval is parameterized by χ. Note that if we take $\chi = q_{1-\alpha/2}$, then we recover the oracle $(1-\alpha)$-credible intervals. Armstrong, Kolesár, and Plagborg-Møller (2022) recommend a principled way to pick $\chi \equiv \chi(\alpha) > q_{1-\alpha/2}$. Their proposal is to pick $\chi(\alpha)$ as follows:

$$\chi(A) = \inf\left\{\chi > 0 \,:\, \inf\left\{\mathbb{P}_G\left[\theta \in \mathcal{J}(Z;\chi,A)\right] \,:\, \mathbb{E}_G\left[\theta^2\right] = A\right\} \geq 1-\alpha\right\}. \tag{5.10}$$

Then, by definition (and a continuity argument) we will have that:

$$\mathbb{P}_G\left[\theta \in \mathcal{J}(Z;\chi(A),A)\right] \geq 1-\alpha, \tag{5.11}$$

for any G with $\mathbb{E}_G\left[\theta^2\right] = A$.

Furthermore, this leads to the following empirical Bayes interval. Let $\hat{A}$ be an estimate of $\mathbb{E}_G\left[\theta^2\right]$. Then Armstrong, Kolesár, and Plagborg-Møller (2022) propose the following confidence interval,

$$\theta_i \in \mathcal{J}(Z;\chi(\hat{A}),\hat{A}).$$

Asymptotic coverage analogous to Theorem 5.1 can also be established.

Let us try to take a closer look at Eq. 5.10. First note that for $G = \mathcal{N}(0,A)$:

$$\mathbb{P}_G\left[\theta \in \mathcal{J}(Z;\chi,A)\right] = 2\Phi(\chi) - 1.$$

Let us compute this for general G. We will do this by iterated expectation. Also let us write $w = A/(A+1)$ and $b = \theta/A$. Then:

$$\begin{aligned}
\mathbb{P}_\theta\left[\theta \in \mathcal{J}(Z;\chi,A)\right] &= \mathbb{P}_\theta\left[|wZ-\theta| \leq \chi\sqrt{w}\right] \\
&= \mathbb{P}_\theta\left[|Z-\theta-b| \leq \chi/\sqrt{w}\right] \\
&= \Phi(b+\chi/\sqrt{w}) - \Phi(b-\chi/\sqrt{w}) \\
&=: r(b,\chi/\sqrt{w}).
\end{aligned}$$

The prior G on θ with $\mathbb{E}_G\left[\theta^2\right] = A$ induces the prior H on b with:

$$\mathbb{E}_H\left[b^2\right] = \mathbb{E}_H\left[\frac{\theta^2}{A^2}\right] = \frac{1}{A}.$$

Hence:

$$\mathbb{P}_G\left[\theta \in \mathcal{J}(Z;\chi,A)\right] = \mathbb{E}_H\left[r(b,\chi/\sqrt{w})\right].$$

Consider the following optimization problem:

$$\text{minimize } \mathbb{E}_H\left[r(b,\chi/\sqrt{w})\right] \text{ s.t. } \mathbb{E}_H\left[b^2\right]. \tag{5.12}$$

The above is a linear program in H and can be computed numerically. Hence to compute Eq. 5.10 it suffices to solve Eq. 5.12 for multiple values of χ and pick the smallest χ so that the minimum value of the optimization problem is at least $1-\alpha$.

5.3.1 Average coverage in the compound decision problem

In the introduction of this chapter, we also discussed the plausible goal of constructing intervals with the property:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\theta_i}\left[\theta_i \in \mathcal{J}_i(\mathbf{Z})\right] \geq 1-\alpha.$$

The construction of Armstrong, Kolesár, and Plagborg-Møller (2022) also enables such a guarantee asymptotically. The argument has a similar flavour as the compound decision results we have already discussed.

Proposition 5.4. *Consider the compound oracle intervals,*

$$\theta_i \in \mathcal{J}^{AKP}(Z_i; \chi(\|\boldsymbol{\theta}\|^2/n), \|\boldsymbol{\theta}\|^2/n).$$

It holds that:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\theta_i}\left[\theta_i \in \mathcal{J}^{AKP}(Z_i; \chi(\|\boldsymbol{\theta}\|^2/n), \|\boldsymbol{\theta}\|^2/n)\right] \geq 1-\alpha.$$

Proof. Take the prior $G = \frac{1}{n}\sum_{i=1}^{n}\delta_{\theta_i}$. This prior has second moment $\mathbb{E}_G\left[\theta^2\right] = \|\boldsymbol{\theta}\|^2/n$. Next notice that:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\theta_i}\left[\theta_i \in \mathcal{J}^{AKP}(Z_i; \chi(\|\theta\|^2/n), \|\theta\|^2/n)\right] = \mathbb{P}_G\left[\theta \in \mathcal{J}^{AKP}(Z; \chi(\|\theta\|^2/n), \|\theta\|^2/n)\right] \geq 1-\alpha$$

The first equality follows from the compound argument, and the inequality follows from Eq. 5.11.

□

Next note that we can consistently estimate (under some regularity) $\left\|\boldsymbol{\theta}^2\right\|_n$ by $\hat{A} = \max\left\{\frac{1}{n}\sum_{i=1}^{n} Z_i^2 - 1, 0\right\}$ and so Armstrong, Kolesár, and Plagborg-Møller (2022) provide conditions such that:

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\theta_i}\left[\theta_i \in \mathcal{J}^{AKP}(Z_i; \chi(\hat{A}), \hat{A})\right] \geq 1-\alpha.$$

5.4 FAB—Frequentist Assisted by (empirical) Bayes intervals

In this section we will explain how to form confidence intervals/sets with the property Eq. 5.7.

It will be convenient to start again by looking only at the n-th simple decision problem (and momentarily ignoring the data Z_j for $j \neq n$). Suppose we are interested in forming a frequentist confidence set $\mathcal{S}$ for θ_n. Suppose further that we have a hunch that $\theta \dot{\sim} G$ approximately holds. We may then seek to choose our confidence set subject to the following criteria: first, it is a frequentist confidence set, i.e., it has the coverage property in Definition 5.2 for all values $\theta_n \in \Theta$, and second, it has the smallest possible expected volume (Lebesgue mass) when we also integrate with respect to the randomness $\theta \sim G$:

$$\begin{aligned} \mathcal{S}^*(\cdot; G) \in \operatorname*{argmin}_{\mathcal{S}(\cdot)} \quad & \mathbb{E}_G \left[\lambda^{Leb}(\mathcal{S}(Z_n))\right] \\ \text{s.t.} \quad & \mathbb{P}_{\theta_n} \left[\theta_n \in \mathcal{S}(Z_n)\right] \geq 1 - \alpha \ \text{ for all } \ \theta_n \in \Theta. \end{aligned} \tag{5.13}$$

We make the dependence on G of the optimal confidence set in Eq. 5.13 explicit by writing $\mathcal{S}^*(\cdot; G)$. The above sets are called "FAB"—Frequentist Assisted by Bayes— by Yu and Hoff (2018). We will explain how we can compute $\mathcal{S}^*(\cdot; G)$ below.

First, however, we turn to our original tasks, i.e., the task of computing an empirical Bayes interval for θ_n based on $Z_1, \dots, Z_n$ such that Eq. 5.7 holds.

The idea is the following. $\mathcal{S}^*(\cdot; G)$ depends on the choice of G: for any G our confidence set has frequentist coverage, however, it will have most power when indeed θ_n is approximately distributed as G. Hence, instead of taking G to be a hunch/guess, we may instead estimate it through empirical Bayes:

1. Let $\widehat{G}_{-n} := \widehat{G}(Z_1, \dots, Z_{n-1})$ be an estimate of G based on $Z_1, \dots, Z_{n-1}$ but *not* Z_n.
2. Report the empirical Bayes confidence set

$$\mathcal{S}_n(\mathbf{Z}) := \mathcal{S}^*(Z_n; \widehat{G}_{-n}). \tag{5.14}$$

This leave-one-out construction enables us to verify Eq. 5.7.

Theorem 5.2. *The confidence set $\mathcal{S}_n(\mathbf{Z})$ Eq. 5.14 has frequentist coverage as in Eq. 5.7, that is,*

$$\mathbb{P}_{\boldsymbol{\theta}} \left[\theta_n \in \mathcal{S}_n(\mathbf{Z})\right] \geq 1 - \alpha \ \textit{ for all } \ \boldsymbol{\theta} \in \Theta_n.$$

Proof.

$$\begin{aligned}\mathbb{P}_{\boldsymbol{\theta}}\left[\theta_n \in \mathcal{S}_n(\mathbf{Z})\right] &= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{P}_{\boldsymbol{\theta}}\left[\theta_n \in \mathcal{S}^*(Z_n; \widehat{G}_{-n}) \mid Z_1, \dots, Z_{n-1}\right]\right] \\ &\geq \mathbb{E}_{\boldsymbol{\theta}}\left[1-\alpha\right] = 1-\alpha.\end{aligned}$$

In going from the first to the second line, we used the fact that Z_n is independent of $Z_1, \dots, Z_{n-1}$, that $\widehat{G}_{-n}$ is a function of $Z_1, \dots, Z_{n-1}$ (and so may be treated as deterministic conditionally on the latter), and the fact that for any fixed G the set $\mathcal{S}^*(Z_n; G)$ is a confidence set for θ_n.

□

5.4.1 Constructing optimal FAB confidence sets

It remains to explain how we can compute optimal confidence sets as in Eq. 5.13. The basic idea, which is due to Pratt (1961), and Pratt (1963), is to build on duality of confidence sets and hypothesis testing. That is, for all θ we may define a a binary indicator function $\phi_\theta(z) \in \{0, 1\}$ such that

$$\theta \in \mathcal{S}(z) \iff \phi_\theta(z) = 0. \tag{5.15}$$

For fixed θ_0, $\phi_{\theta_0}(Z)$ is a test of the null hypothesis $H_0 : \theta = \theta_0$: we reject H_0 when $\phi_{\theta_0}(Z) = 1$. If $\mathcal{S}(z)$ is a $1-\alpha$ confidence set for θ, then $\phi_{\theta_0}(Z)$ is a size α test.[3]

Now fix a confidence $\mathcal{S}(\cdot)$ and its associated testing function $(\theta, z) \mapsto \phi_\theta(z)$. Fix any θ.

$$\begin{aligned}\mathbb{E}_G\left[\lambda^{Leb}(\mathcal{S}(Z))\right] &= \mathbb{E}_G\left[\int \mathbf{1}(\theta_0 \in \mathcal{S}(Z)) d\theta_0\right] \\ &= \int \mathbb{P}_G\left[\theta_0 \in \mathcal{S}(Z_n)\right] d\theta_0 \\ &= \int \left(1 - \mathbb{P}_G\left[\phi_{\theta_0} = 1\right]\right) d\theta_0.\end{aligned}$$

Thus to compute the optimal confidence set $\mathcal{S}$. It suffices to construct for each θ_0 a test ϕ_{θ_0} that solves the following optimization problem:

$$\text{maximize} \;\; \mathbb{P}_G\left[\phi_{\theta_0}(Z) = 1\right] \;\; \text{s.t.} \;\; \mathbb{P}_{\theta_0}\left[\phi_{\theta_0}(Z) = 1\right] \leq \alpha. \tag{5.16}$$

This family of tests then yields an optimal confidence set via Eq. 5.15.

Example 5.4 (FAB with Gaussian noise and point mass at 0 (Pratt 1961))**.** Suppose we take $G = \delta_0$, a Dirac mass at 0 and that $Z \mid \theta \sim \mathcal{N}(\theta, 1)$. Then the confidence set $\mathcal{S}^*(z; \delta_0)$ (which in fact is a confidence interval) is the following:

$$\mathcal{S}^*(z; \delta_0) = \left[\min\left\{0,\, Z - q_{1-\alpha}\right\},\, \max\left\{0,\, Z + q_{1-\alpha}\right\}\right].$$

[3]Since:

$$\mathbb{P}_{\theta_0}\left[\phi_{\theta_0}(Z) = 1\right] = \mathbb{P}_{\theta_0}\left[\theta_0 \notin \mathcal{S}(z)\right] = 1 - \mathbb{P}_{\theta_0}\left[\theta_0 \in \mathcal{S}(z)\right] \leq \alpha.$$

Before delving into the proof, let us make the following remark. Compared to the textbook interval in Example 5.1, when $\theta \approx 0$, the above interval essentially avoids the "Bonferroni" adjustment for looking at both tails. In such cases, this interval in Example 5.4 has approximately 85% of the length of the interval in Example 5.1.

Essentially this is the scope for adaptation in the above problem, i.e., one cannot do better than 85% compared to Example 5.1 while maintaining a valid confidence interval (set). When intervals are constructed by studentization (unknown variance), then the gains due to FAB can be substantially larger than 85%, see Yu and Hoff (2018).

Proof. We may follow the recipe above. It will be convenient in this case to note the following.

$$\mathbb{E}_0\left[\lambda^{Leb}(\mathcal{S}(Z))\right] = \int \left(1 - \mathbb{P}_G\left[\phi_{\theta_0} = 1\right]\right) d\theta_0 = \int_{\theta_0 \neq 0} \left(1 - \mathbb{P}_G\left[\phi_{\theta_0} = 1\right]\right) d\theta_0.$$

Hence we only need to construct tests as in Eq. 5.16 for $\theta_0 \neq 0$. The objective Eq. 5.16 then states that we seek to find the most powerful test for the following comparison:

$$H_0 : \theta = \theta_0 \ \text{ vs. } H_A : \theta = 0.$$

An optimal test is given by the Neyman-Pearson lemma.[4] Hence for any $\theta_0 \neq 0$, an optimal test is given by:

$$\phi_{\theta_0}(z) = \mathbf{1}\left(\frac{p(z \mid 0)}{p(z \mid \theta_0)} > k(\theta_0)\right), \ \text{ with } k(\theta_0) \text{ s.t. } \mathbb{E}_{\theta_0}\left[\phi_{\theta_0}(z)\right] = \alpha.$$

Suppose first that $\theta_0 < 0$. In that case, by Neyman-Pearson, we reject for large values of Z, i.e.,

$$\phi_{\theta_0}(Z) = \mathbf{1}(Z > \theta_0 + q_{1-\alpha}).$$

Analogously, for $\theta_0 > 0$, Neyman-Pearson rejects for small values of Z, i.e.,

$$\phi_{\theta_0}(Z) = \mathbf{1}(Z < \theta_0 - q_{1-\alpha}).$$

We seek to turn this into a confidence set via Eq. 5.15. Hence fix Z. We seek to find all values of θ_0 such that the tests constructed above do not reject. Hence:

Fix $\theta_0 < 0$. We will not reject it if $Z \leq \theta_0 + q_{1-\alpha}$, i.e., if $\theta_0 \geq Z - q_{1-\alpha}$. Similarly, for $\theta_0 > 0$, we will not reject it if $Z \geq \theta_0 - q_{1-\alpha}$, i.e., if $\theta_0 \leq Z + q_{1-\alpha}$. We thus get the confidence interval that we claimed above.

□

[4] We assume here that the distribution of Z is absolutely continuous with respect to the Lebesgue measure so that we need not deal with randomized tests/confidence sets.

5.5 Confidence sets for all latent parameters

Let us turn out attention to constructing confidence sets such that $\boldsymbol{\theta} \in \mathcal{S}(\mathbf{Z})$. The typical frequentist confidence set is the following sphere:

$$\mathcal{S}(\mathbf{Z}) = \left\{ \boldsymbol{\theta} : \|\boldsymbol{\theta} - Z\|^2 \leq \chi^2_{n,1-\alpha} \right\},$$

where $\chi^2_{n,1-\alpha}$ is the $1-\alpha$ quantile of the χ^2-distribution with n degrees of freedom. It holds that:

$$\mathbb{P}_{\boldsymbol{\theta}} \left[\boldsymbol{\theta} \in \mathcal{S}(\mathbf{Z}) \right] = 1 - \alpha.$$

It turns out that for $n > 3$, the above interval is inadmissible (just as $\hat{\boldsymbol{\theta}} = \mathbf{Z}$ is inadmissible in mean squared error for estimating $\boldsymbol{\theta}$). Let:

$$\hat{\boldsymbol{\theta}}^{JS,a,+} = \left(1 - \frac{a}{\|\mathbf{Z}\|^2} \right)_+ \mathbf{Z}.$$

Then consider the sphere that is centered not at $\mathbf{Z}$, but instead at the (modified) James-Stein estimator above:

$$\mathcal{S}^{JS,a}(\mathbf{Z}) = \left\{ \boldsymbol{\theta} : \left\| \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{JS,a,+} \right\|^2 \leq \chi^2_{n,1-\alpha} \right\}.$$

The following theorem verifies a result conjectured by C. M. Stein (1962):

Theorem 5.3 (Hwang and Casella (1984))**.** *Let $n \geq 3$ and let a be sufficiently small (e.g., $a \lesssim 0.8(n-2)$). Then:*

$$\mathbb{P}_{\boldsymbol{\theta}} \left[\boldsymbol{\theta} \in \mathcal{S}^{JS,a}(\mathbf{Z}) \right] > 1 - \alpha,$$

for all $\boldsymbol{\theta} \in \Theta^n$.

5.6 Confidence intervals for empirical Bayes estimands

We now turn to the last task we discussed in Section 5.1.3, namely that of forming confidence intervals about empirical Bayes estimands. For simplicity, we consider the Bayes decision $t_G(z) = \mathbb{E}_G \left[h(\theta) \mid Z = z \right]$, where z is fixed. How can we construct a confidence interval around $t_G(z)$? We provide a general solution below that works for any choice of $p(\cdot \mid \theta)$, as long as $Z_i \in \mathbb{R}$.

Our starting point is to recall the Kolmogorov-Smirnov minimum distance estimator that Robbins (1964) proposed to estimate $t_G(z)$ nonparametrically using G-modeling. First, let

$$\widehat{G} \in \operatorname{argmin} \left\{ d_{\mathrm{KS}}(F_{\widetilde{G}}, \widehat{F}_n) \,:\, \widetilde{G} \in \mathcal{G} \right\}, \tag{5.17}$$

where $F_G(\cdot)$ is the marginal distribution of Z, $\widehat{F}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(Z_i \leq t)$ is the empirical distribution of the Z_i and d_{KS} is the Kolmogorov-Smirnov distance:

$$d_{\mathrm{KS}}(F_1, F_2) = \sup_{z \in \mathbb{R}} |F_1(z) - F_2(z)|$$

for any two distribution functions F_1, F_2 on $\mathbb{R}$. We already discussed this construction in Eq. 1.16 from Chapter 1 for the special case where $\mathcal{G}$ consists of all distributions supported on Θ. Here Eq. 5.17 is slightly more general and allows the data analyst to specify a smaller class of distributions $\mathcal{G}$ on Θ.

Robbins (1964) used Eq. 5.17 to estimate $t_G(z)$ based on the plug-in principle, $\hat{t}_G(z) = t_{\widehat{G}}(z)$. Here we will use an analogous construction for our goal of **inference** for $t_G(z)$. The idea is the following. We have a strong probabilistic understanding of $d_{\mathrm{KS}}(F_G, \widehat{F}_n)$. In particular, by Massart's tight constant for the Dvoretzky–Kiefer–Wolfowitz (DKW) inequality (Massart 1990), it holds that:

$$\begin{aligned} &\mathbb{P}_G\left[F_G \in \mathcal{F}_n^{\mathrm{DKW}}(\alpha)\right] \geq 1 - \alpha, \text{ where} \\ &\mathcal{F}_n^{\mathrm{DKW}}(\alpha) := \left\{F \text{ distribution } : d_{\mathrm{KS}}(F, \widehat{F}_n) \leq \sqrt{\log\left(2/\alpha\right)/(2n)}\right\}. \end{aligned} \tag{5.18}$$

Ignatiadis and Wager (2019) use the term "F-Localization" for a confidence set of marginal distribution such as $\mathcal{F}_n^{\mathrm{DKW}}(\alpha)$. Eq. 5.18 can be used to form a confidence interval for $t_G(z)$ as follows. We then form the F-Localization confidence interval

$$\mathcal{I}_\alpha(z) = [\hat{t}_\alpha^-(z), \hat{t}_\alpha^+(z)],$$

where:

$$\begin{aligned} &\hat{t}_\alpha^-(z) = \inf\left\{t_G(z) \mid G \in \mathcal{G}\left(\mathcal{F}_n(\alpha)\right)\right\}, \hat{t}_\alpha^+(z) = \sup\left\{t_G(z) \mid G \in \mathcal{G}\left(\mathcal{F}_n(\alpha)\right)\right\}, \\ &\text{where } \mathcal{G}(\mathcal{F}) = \left\{G \in \mathcal{G} \mid F_G \in \mathcal{F}\right\}. \end{aligned} \tag{5.19}$$

What does this construction say? While the point estimate of Robbins (1964) only considers the prior $\hat{G}$ that minimizes the Kolmogorov-Smirnov distance, in Eq. 5.19 we account for all plausible priors G with a marginal distribution that is statistically plausible based on the uncertainty characterization in Eq. 5.18.

It immediately follows that the F-Localization intervals constructed above have finite-sample frequentist coverage for $t_G(z)$:

It holds that:

$$\mathbb{P}_G\left[t_G(z) \in \mathcal{I}_\alpha(z)\right] \geq 1 - \alpha.$$

Proof.

$$\mathbb{P}_G\left[t_G(z) \in \mathcal{I}_\alpha(z)\right] \geq \mathbb{P}_G\left[\mathcal{F}_n^{\mathrm{DKW}}(\alpha)\right] \geq 1 - \alpha.$$

The first inequality follows from the definition of the F-Localization interval, and the second inequality from Eq. 5.18.

□

5.7 Further bibliographic remarks

Cox (1975) develops results for intervals with the property Eq. 5.1 by treating them as a special case of prediction intervals. Morris (1983) uses the term "empirical Bayes intervals" also for intervals with the property Eq. 5.1 (as well as intervals with the property Eq. 5.5). We prefer to reserve the terminology only for the latter, since the definition in Eq. 5.1 pertains to a simple statistical decision problem—and does not require, e.g., borrowing information from parallel related statistical decision problems. Further results and practical constructions for such intervals with an emphasis on parametric problems are given by N. M. Laird and Louis (1987), Carlin and Gelfand (1990), and Carlin and Gelfand (1991). Section 5.1.2 closely follows and elaborates on Jiang (2019), which is a discussion paper of Bradley Efron (2019). Koenker (2020) compares some more nonparametric approaches for the construction of empirical Bayes intervals for individual latent parameters.

We refer to Casella and Hwang (2012) for a review of the historical developments of empirical Bayes confidence sets for all latent parameters that elaborates on Section 5.5. Some important references include Samworth (2005) and Bradley Efron (2006).

Ignatiadis and Wager (2022) develop general methods for confidence intervals of empirical Bayes estimands including the F-Localization method described in Section 5.6. In particular, they describe how the optimization problems in Eq. 5.19 can be solved efficiently using convex programming when $\mathcal{G}$ is convex, and they also develop alternative intervals based on affine minimax estimators. The idea of leveraging the Kolmogorov-Smirnov band around the marginal distribution for finite-sample inference (as in the approach of Section 5.6) has a long history in statistics. For example, Anderson (1969) suggested to use the Kolmogorov-Smirnov band to form confidence intervals for the mean of a $[0, 1]$-valued random variable, as follows: one takes the minimum, resp. maximum of $\int z dF(z)$ subject to $F \in \mathcal{F}_n^{\mathrm{DKW}}(\alpha)$ and F supported on $[0, 1]$; also see Romano and Wolf (2000). Furthermore, one may construct other F-Localizations, i.e., confidence sets of distributions, beyond Eq. 5.18. We refer to Lord and Cressie (1975), Lord and Stocking (1976), Greenshtein and Itskov (2018), and Ignatiadis and Wager (2022) for such constructions and their application to empirical Bayes problems.

6 Exponential Families, Tweedie's Formula, and F-modeling

6.1 Preliminaries: Exponential families

Consider a random vector $Z \in \mathbb{R}^p$ having a density, with respect to (w.r.t.) a dominating measure λ, parametrized by $\theta := (\theta_1, \dots, \theta_s) \in \mathbb{R}^s$ and expressible as:

$$p(z \mid \theta) := \exp\Big[\sum_{j=1}^{s} \theta_j T_j(z) - A(\theta)\Big] h(z), \qquad \text{for } \ z \in \mathbb{R}^p. \tag{6.1}$$

Here $h : \mathbb{R}^p \to \mathbb{R}$ is a nonnegative function, $T = (T_1, \dots, T_s)$ is a measurable function from $\mathbb{R}^p$ to $\mathbb{R}^s$, and the parameter space is the set

$$\Xi := \{\theta \in \mathbb{R}^s : A(\theta) < \infty\}, \tag{6.2}$$

where the function $A : \Xi \to \mathbb{R}$ (sometimes referred to as the *log-partition function* or the *cumulant function*) is defined as

$$A(\theta) := \log \int \exp\Big[\sum_{j=1}^{s} \theta_j T_j(z)\Big] h(z) d\lambda(z). \tag{6.3}$$

We will assume that Ξ is a non-empty open set (in $\mathbb{R}^s$).

In this case, Z is said to belong to a *regular s-parameter exponential family*, and θ is the natural or canonical parametrization.

There are many examples of parametric families belonging to an exponential family, e.g., Gaussian, binomial, multinomial, Poisson, gamma, and beta distributions, as well as many others. Here are some examples.

Example 6.1 (Poisson distribution)**.** Consider the Poisson distribution parametrized by $\mu \in (0, \infty)$:

$$p_\mu(z) = \frac{\mu^z e^{-\mu}}{z!}, \qquad \text{for } \ z = 0, 1, 2, \dots . \tag{6.4}$$

The above family is indeed a 1-parameter exponential family with natural parameter $\theta := \log \mu$ and $\Xi = \mathbb{R}$. Here λ is the counting measure on the nonnegative integers, $T(z) = z$, $h(z) = \frac{1}{z!}$ and $A(\theta) = \log \sum_{z=0}^{\infty} e^{\theta z} \frac{1}{z!} = \log(e^{e^\theta}) = e^\theta$.

Example 6.2 (Chi-square distribution)**.** Suppose that $Z \equiv S^2 \mid \sigma^2 \sim \frac{\sigma^2}{\nu}\chi^2_\nu$ where χ^2_ν is the chi-squared distribution with ν degrees of freedom, i.e.,

$$p_{\sigma^2}(s^2) = \frac{\nu^{\nu/2}}{(\sigma^2)^{\nu/2}\, 2^{\nu/2}\Gamma(\nu/2)} \left(s^2\right)^{\nu/2-1} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right), \qquad \text{for} \;\; s^2 > 0. \tag{6.5}$$

The above family is a 1-parameter exponential family with natural parameter $\theta := -\frac{\nu}{2\sigma^2}$ and $\Xi = (-\infty, 0)$. Here $T(s^2) = s^2$ $(T(z) = z)$,

$$h(s^2) = \frac{1}{\Gamma(\nu/2)} \left(s^2\right)^{\nu/2-1} \mathbf{1}_{(0,\infty)}(s^2) \quad \text{and} \quad A(\theta) = \frac{\nu}{2}\log(-\theta).$$

Example 6.3 (Multivariate normal)**.** Consider the family of multivariate normal distributions on $\mathbb{R}^p$ with a fixed known nonsingular covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ and unknown mean vector $\mu = (\mu_1, \dots, \mu_p) \in \mathbb{R}^p$, i.e., $Z \sim N_p(\mu, \Sigma)$ has density given by

$$p_\mu(z) = \frac{e^{-\frac{1}{2}(z-\mu)^\top \Sigma^{-1}(z-\mu)}}{\sqrt{(2\pi)^p|\Sigma|}}, \qquad \text{for} \;\; z \in \mathbb{R}^p. \tag{6.6}$$

It is easy to check that Eq. 6.6 can be expressed in the form Eq. 6.1 where we take

$$\theta := \Sigma^{-1}\mu, \qquad T(z) := z, \qquad A(\theta) := \frac{1}{2}\theta^\top \Sigma \theta \quad \text{and} \quad h(z) \equiv p_0(z) = \frac{e^{-\frac{1}{2}z^\top \Sigma^{-1} z}}{\sqrt{(2\pi)^p|\Sigma|}}.$$

Suppose that $Z \sim f_\theta$ as in Eq. 6.1. Here are some important properties of exponential families.

1. The support of Z (i.e., $p(z \mid \theta) > 0$) does not depend on θ. Let $\mathcal{Z} \subset \mathbb{R}^p$ denote the support of Z.
2. It is clear that the statistic $T(Z)$ is a *sufficient statistic* for this family. It can be shown that[1]
$$\mathbb{E}_\theta\left[T_j(Z)\right] = \frac{\partial A(\theta)}{\partial \theta_j}, \qquad \text{for} \;\; j = 1, \dots, s. \tag{6.7}$$
3. The natural parameter space Ξ is a *convex set* and the cumulant function $A(\cdot)$ is a *convex function.*

[1]A proof of this can be obtained as follows. Recall Eq. 6.3. Thus,

$$e^{A(\theta)} = \int e^{\theta^\top T(z)} h(z) d\lambda(z).$$

Differentiating this expression with respect to θ_j, which can be done under the integral if $\theta \in \Xi^o$, gives

$$e^{A(\theta)}\frac{\partial A(\theta)}{\partial \theta_j} = \int T_j(z) e^{\theta^\top T(z)} h(z)\, d\lambda(z) \quad \Rightarrow \frac{\partial A(\theta)}{\partial \theta_j} = \int T_j(z) p(z \mid \theta)\, d\lambda(z) = \mathbb{E}_\theta\left[T_j(Z)\right].$$

4. The *moment generating function* of $T \equiv (T_1(Z), \dots, T_s(Z))$, for $u \in \mathbb{R}^s$ such that $u+\theta \in \Xi$, is

$$\begin{aligned} M_T(u) := \mathbb{E}\left[e^{u^\top T}\right] &= \int e^{u^\top T(z)} e^{\theta^\top T(z) - A(\theta)} h(z)\, d\lambda(z) \\ &= e^{A(u+\theta)-A(\theta)} \int p(z \mid u+\theta)\, d\lambda(z) = e^{A(u+\theta)-A(\theta)}. \end{aligned}$$

Noting that if $M_T(\cdot)$ is finite in some neighborhood of the origin, then M_T has continuous derivatives of all orders at the origin, and for $r_j \geq 0$, for $j = 1, \dots, s$,

$$\alpha_{r_1,\dots,r_s} := \mathbb{E}\left[T_1^{r_1}(Z) \times \cdots \times T_s^{r_s}(Z)\right] = \frac{\partial^{r_1}}{\partial u_1^{r_1}} \cdots \frac{\partial^{r_s}}{\partial u_s^{r_s}} M_T(u)\Big|_{u=0}.$$

Thus, when $r_j = 1$ and $r_k = 0$ for all $k \neq j$, we obtain Eq. 6.7.

5. The *cumulant generating function* is

$$K_T(u) := \log M_T(u) = A(u+\theta) - A(\theta). \tag{6.8}$$

See Keener (2010b, chap. 10) for a more detailed study of exponential families.

6.2 Tweedie's formula

Now suppose that Θ is assumed to have a prior distribution G (on $\mathbb{R}^s$). Thus our model becomes:

$$\Theta \sim G \qquad \text{and} \qquad Z \mid \Theta = \theta \sim p(\cdot \mid \theta), \tag{6.9}$$

where we assume that $p(\cdot \mid \theta)$ comes from the exponential family Eq. 6.1. Then the marginal density of Z (w.r.t.$\sim\lambda$) is

$$f_G(z) := \int p(z \mid \theta)\, dG(\theta), \qquad \text{for } \ z \in \mathbb{R}^p.$$

Let $\mathcal{Z}$ be the support of the marginal distribution of Z. Now Bayes rule provides the posterior density of Θ given Z. Suppose that Θ has density $g(\cdot)$, w.r.t. a dominating measure ξ, with support $\Omega \subset \Xi$. Then, the posterior density of Θ given $Z = z$ (w.r.t.$\sim\xi$) is given by, for $\theta \in \Omega$ and $z \in \mathcal{Z}$,

$$p_{\Theta|Z}(\theta \mid z) = \frac{p(z \mid \theta) g(\theta)}{f_G(z)} = \frac{e^{\theta^\top T(z) - A(\theta)} h(z) g(\theta)}{f_G(z)} = e^{\theta^\top T(z) - \kappa(z)} e^{-A(\theta)} g(\theta), \tag{6.10}$$

where

$$\kappa(z) := \log\left(\frac{f_G(z)}{h(z)}\right), \qquad \text{for} \quad z \in \mathcal{Z}. \tag{6.11}$$

This implies that $\Theta \mid Z = z$ is also an *exponential family* with canonical parameter $T(z)$, sufficient statistic Θ, and log-partition function $\kappa(z)$. Thus, the cumulant generating function is

$$\log \mathbb{E}\left[e^{\Theta^\top t} \mid Z = z\right] = \kappa(t + z) - \kappa(z) \tag{6.12}$$

for $z \in \mathcal{Z}$ such that $t + z \in \mathcal{Z}$.

Tweedie's formula, given below, calculates the posterior expectation of Θ given $Z = z$ in the setting Eq. 6.9.

Lemma 6.1 (Tweedie's formula)**.** *For $z \in \mathcal{Z}$, we have*

$$\mathbb{E}\left[\Theta \mid Z = z\right] = \nabla\kappa(z) = \frac{\nabla f_G(z)}{f_G(z)} - \frac{\nabla h(z)}{h(z)}. \tag{6.13}$$

Proof. The result is a direct consequence of the fact that the distribution of $\Theta \mid Z = z$ is an s-parameter exponential family with log-partition function $\kappa(\cdot)$ defined via Eq. 6.11: By property 1. above the expectation of the sufficient statistic Θ can then be expressed as the gradient of the log-partition function.

□

For $p = s = 1$, the above formula for the Gaussian case was given in Robbins (1956). Bradley Efron (2011) calls this Tweedie's formula since Robbins attributes it to M.C.K. Tweedie; however it appears earlier in Dyson (1926) who credits it to the English astronomer Arthur Eddington.

Lemma 6.2. *Consider model Eq. 6.9 where we assume that $p(\cdot \mid \theta)$, for $\theta \in \Xi$, is a member of an exponential family of distributions as in Eq. 6.1 with $T(z) = z$ and $s = p$. Suppose further that $h(\cdot)$ in Eq. 6.1 integrates to 1 (w.r.t.~λ). Then $\kappa(\cdot)$, as defined in Eq. 6.11, is a convex function. As a consequence, $\mathbb{E}\left[\Theta \mid Z = z\right]$ is the gradient of a convex function.*

Proof. Observe that under the assumptions of the lemma, from Eq. 6.11 we see that the distribution of $\Theta \mid Z = z$ is an s-parameter exponential family with log-partition function $\kappa(\cdot)$ defined via Eq. 6.11. As the log-partition function $\kappa(\cdot)$ is known to be convex, the result follows.

□

6.2.1 Tweedie's formula for multivariate normal distribution

Suppose now that Z has multivariate normal distribution with known covariance matrix as in Example 6.3. Then, for $z \in \mathbb{R}^p$,

$$\mathbb{E}\left[\Theta \mid Z = z\right] = \nabla\kappa(z) = \Sigma^{-1}z + \frac{\nabla f_G(z)}{f_G(z)},$$

where the last equality follows from Eq. 6.13 and the fact that $\nabla h(z) = -h(z)(\Sigma^{-1}z)$. Thus, the Bayes estimator of mean μ in Eq. 6.6 is

$$\mathbb{E}\left[\mu \mid Z = z\right] = z + \Sigma\frac{\nabla f_G(z)}{f_G(z)}. \tag{6.14}$$

6.2.2 A Tweedie-like formula for the χ^2-distribution

Now we consider the χ^2-distribution from Example 6.2. As we already explained, the above lies in an exponential family with natural parameters $\theta = -\frac{\nu}{2\sigma^2}$. In fact, we could have equivalently parameterized it as an exponential family with natural parameter $\tau^2 := 1/\sigma^2$, that is, the precision.

Exercise 6.1. Use Tweedie's formula to derive a formula for $\mathbb{E}\left[\frac{1}{\sigma^2} \mid Z = z\right]$.

The χ^2-case is another example where we have access to an F-formula that does not follow from Tweedie's result. The following result is due to Robbins (1982) (also see Gu and Koenker (2017)).

$$\mathbb{E}\left[\sigma^2 \mid S^2 = s^2\right] = \frac{\nu}{2}\frac{\left(s^2\right)^{\nu/2-1}}{f_G(s^2)}\int_{s^2}^{\infty}\left(t^2\right)^{1-\nu/2} f_G(t^2)dt^2.$$

Proof. We will use the following observation:

$$\int_{s^2}^{\infty}\exp\left(-\frac{\nu t^2}{2\sigma^2}\right)dt^2 = -\left[\frac{2\sigma^2}{\nu}\exp\left(-\frac{\nu t^2}{2\sigma^2}\right)\right]_{s^2}^{\infty} = \frac{2\sigma^2}{\nu}\exp\left(-\frac{\nu s^2}{2\sigma^2}\right).$$

Then, letting $C := \frac{(\nu/2)^{\nu/2}}{\Gamma(\nu/2)}$,

$$
\begin{aligned}
&\int_0^\infty \sigma^2 p(s^2 \mid \sigma^2) dG(\sigma^2) \\
&= \int_0^\infty \frac{\nu}{2} \frac{2\sigma^2}{\nu} p(s^2 \mid \sigma^2) dG(\sigma^2) \\
&= C \int_0^\infty \frac{\nu}{2} \cdot \left(s^2\right)^{\nu/2-1} \left(\sigma^{-2}\right)^{\nu/2} \cdot \frac{2\sigma^2}{\nu} \exp\left(-\frac{\nu s^2}{2\sigma^2}\right) dG(\sigma^2) \\
&= C \int_0^\infty \frac{\nu}{2} \left(s^2\right)^{\nu/2-1} \left(\sigma^{-2}\right)^{\nu/2} \left(\int_{s^2}^\infty \exp\left(-\frac{\nu t^2}{2\sigma^2}\right) dt^2\right) dG(\sigma^2) \\
&= \int_{s^2}^\infty \frac{\nu}{2} \left(s^2\right)^{\nu/2-1} \left(t^2\right)^{1-\nu/2} \int_0^\infty C \left(t^2\right)^{\nu/2-1} \left(\sigma^{-2}\right)^{\nu/2} \exp\left(-\frac{\nu t^2}{2\sigma^2}\right) dG(\sigma^2) dt^2 \\
&= \frac{\nu}{2} \left(s^2\right)^{\nu/2-1} \int_{s^2}^\infty \left(t^2\right)^{1-\nu/2} f_G(t^2) dt^2.
\end{aligned}
$$

□

6.3 Compound decisions and F-modeling

As we have seen before, e.g., in Section 1.4.1 and Section 1.4.2 of Chapter 1, empirical Bayes approaches typically follow one of two strategies: F-modeling or G-modeling. F-modeling is not always possible, however it is, when a Tweedie-type formula is available, as in this chapter. In the remainder of this chapter, we will present an example of an application of the F-modeling approach for the compound estimation of normal means. We will present a G-modeling strategy for the same problem in Chapter 7.

Before providing results for the compound estimation of normal means, we first present more general results on compound estimation.

6.3.1 Symmetric decisions

Consider the setting of Definition 1.2 in Chapter 1 where we have unknown parameters $\theta_1, \dots, \theta_n$ and we observe independent random variables

$$Z_i \sim p(\cdot \mid \theta_i), \qquad \text{for} \quad i = 1, \dots, n. \tag{6.15}$$

A natural class of decision functions is the class of *simple symmetric estimators.[2] This is the class of estimators $\mathbf{\Delta}(\mathbf{Z}) \equiv \mathbf{\Delta}(\mathbf{Z}|t(\cdot))$ of the form

$$\mathbf{\Delta}(\mathbf{Z}) := (t(Z_1), \dots, t(Z_n)) \tag{6.16}$$

for some function $t : \mathbb{R} \to \mathbb{R}$. Given $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$, let

$$t^*_{\boldsymbol{\theta}} := \operatorname*{argmin}_{t:\mathbb{R}\to\mathbb{R}} R(\mathbf{\Delta}(\cdot|t), \boldsymbol{\theta}) \tag{6.17}$$

where the infimum above is over all measurable functions $t : \mathbb{R} \to \mathbb{R}$; we explicitly denote the dependence of the above estimator on $\boldsymbol{\theta}$.

Consider an oracle that knows the value of the vector $\boldsymbol{\theta}$ but must use a simple symmetric estimator. Such an oracle would use the estimator $\mathbf{\Delta}(\mathbf{Z}|t^*_{\boldsymbol{\theta}})$, where $t^*_{\boldsymbol{\theta}}$ is defined in Eq. 6.16. The goal of compound decision theory is to achieve nearly the risk obtained by such an oracle, but by using a "legitimate" estimator, one that may involve the entire vector of observations $\mathbf{Z}$ but does not involve knowledge of the parameter vector $\boldsymbol{\theta}$.

Empirical Bayes: The compound estimation of the vector $\boldsymbol{\theta}$ is closely related to the Bayes estimation of a single random observation. In this Bayes problem, we estimate a random parameter M based on Z such that:

$$M \sim G, \qquad Z \mid M \sim p(\cdot \mid M), \tag{6.18}$$

where G is the unknown prior distribution. Here, the target is the Bayes procedure t_G, i.e., the estimator that minimizes the expected average (Bayes) risk under G:

$$R(t(\cdot), G) = \mathbb{E}_G\left[\ell(t(Z), M)\right] = \int \mathbb{E}_\theta\left[\ell(t(Z), \theta)\right] \, dG(\theta).$$

The goal is to find a procedure $t(\cdot)$ whose expected risk under G is suitably near that of t_G as $n \to \infty$, when G is unknown.

6.3.2 Connection between the compound and the empirical Bayes settings

The prior distribution G which naturally matches the unknown parameters $\{\theta_i : 1 \le i \le n\}$ in Eq. 6.15 in the compound setting is the empirical distribution G_n of the θ_i's:

$$G_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{\theta_i}. \tag{6.19}$$

The *fundamental theorem of compound decisions* (Robbins (1951)) asserts that the compound risk of a simple symmetric (aka separable) estimator $\mathbf{\Delta}(\mathbf{Z}) \equiv \mathbf{\Delta}(\mathbf{Z}|t)$ (see Eq. 6.16) in the

[2] Such an estimator is also sometimes called a 'separable' estimator as one uses a fixed deterministic function of the i-th observation to estimate the i-th mean.

multivariate model Eq. 6.15 is identical to the Bayes risk of the same rule $t(Z)$ under the prior Eq. 6.19 in the univariate model Eq. 6.18, i.e.,

$$R(\boldsymbol{\Delta}(\cdot|t), \boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\theta_i}\left[\ell(t(Z_i), \theta_i)\right] = \mathbb{E}_{G_n}\left[\ell(t(Z), M)\right]. \tag{6.20}$$

As a consequence of the above formal relationship, we can connect the Bayes procedure (under squared error loss) $\Delta_G(\mathbf{Z}) = (t_G(Z_1), \dots, t_G(Z_n))$, for a specified prior G, which is, of course, the posterior mean given by the formula:

$$t_G(z) := \mathbb{E}\left[M \mid Z = z\right] = \frac{\int u\, p\left(z \mid u\right) dG(u)}{\int p\left(z \mid u\right) dG(u)}, \tag{6.21}$$

to the compound setting to obtain an explicit form for $t^*_{\boldsymbol{\theta}}$ (see Eq. 6.17):

$$t^*_{\boldsymbol{\theta}}(u) := \frac{\sum_{i=1}^{n} \theta_i\, p\left(u \mid \theta_i\right)}{\sum_{i=1}^{n} p\left(u \mid \theta_i\right)}. \tag{6.22}$$

6.3.3 Compound estimation of normal means

In this subsection we take $p(\cdot \mid \theta_i) = \phi(\cdot - \theta_i)$ corresponding to the *normal means* problem (here ϕ is the standard normal density function).

Consider this problem for a general variance σ^2; that is, suppose $Z \mid M \sim N(M, \sigma^2)$, where $M \sim G$. Let f_{G,σ^2} be the marginal density of Z, i.e.,

$$f_{G,\sigma^2}(z) := \int \frac{1}{\sigma}\phi\left(\frac{z-\theta}{\sigma}\right) dG(\theta), \qquad \text{for } z \in \mathbb{R}. \tag{6.23}$$

By Tweedie's formula (see Eq. 6.14) the Bayes estimator (under squared error loss) is given by

$$\begin{aligned} t_{G,\sigma^2}(z) := \mathbb{E}\left[M \mid Z = z\right] &= \frac{1}{f_{G,\sigma^2}(z)} \int u\, \frac{1}{\sigma}\phi\left(\frac{z-u}{\sigma}\right) dG(u) \\ &= z + \sigma^2 \frac{f'_{G,\sigma^2}(z)}{f_{G,\sigma^2}(z)}. \end{aligned} \tag{6.24}$$

Here $f'_{G,\sigma^2}(z)$ is the derivative of $f_{G,\sigma^2}(z)$.

In this section we consider estimation of $t_G(\cdot) \equiv t_{G,1}(\cdot)$ and $t^*_{\boldsymbol{\theta}}$ (see Eq. 6.22). Our approach will take advantage of Tweedie's formula above which directly relates the quantity of interest $t_G(\cdot)$ to $f_G \equiv f_{G,1}$, which is the marginal density of the observations $Z_1, \dots, Z_n$. Moreover, it suggests a natural estimator for $t_G(\cdot) \equiv t_{G,1}(\cdot)$ of the form:

$$\hat{t}(z) := z + \frac{\hat{f}'(z)}{\hat{f}(z)}, \tag{6.25}$$

where $\hat{f}'(\cdot)$ and $\hat{f}(\cdot)$ are appropriate estimators for the marginal density (of the data) $f_{G,1}(\cdot) \equiv f_G(\cdot)$ and its derivative $f'_G(\cdot)$.

Remark. All the equations obtained so far for the empirical Bayes setup have a parallel derivation and presentation in the compound decision setup for a given $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$, where $G_n = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i}$ — the empirical distribution determined by the components of $\boldsymbol{\theta}$ — plays the role of G; for example, comparing Eq. 6.22 and Eq. 6.24 we get

$$t^*_{\boldsymbol{\theta}}(z) := t_{G_n,1}(z).$$

Similarly, we denote

$$t^*_{\boldsymbol{\theta},\sigma^2}(z) := t_{G_n,\sigma^2}(z) = z + \sigma^2 \frac{f'_{G_n,\sigma^2}(z)}{f_{G_n,\sigma^2}(z)}, \quad \text{where } f_{G_n,\sigma^2}(z) = \frac{1}{n}\sum_{i=1}^n \frac{1}{\sigma}\phi\left(\frac{z-\theta_i}{\sigma}\right).$$

In the following we will consider *kernel estimators* of f_G and f'_G and study the risk consistency of $\hat{t}$ as in Brown and Greenshtein (2009). Define the kernel density estimator (obtained from the data $\mathbf{Z}$) as

$$\hat{f}(z) \equiv \hat{f}_h(z) := \frac{1}{nh}\sum_{i=1}^n \phi\left(\frac{z-Z_i}{h}\right), \tag{6.26}$$

and its derivative as

$$\hat{f}'(z) \equiv \hat{f}'_h(z) := \frac{1}{nh}\sum_{i=1}^n \left(\frac{Z_i - z}{h^2}\right)\phi\left(\frac{z-Z_i}{h}\right),$$

where $\phi(\cdot)$ is the normal kernel (i.e., the standard normal density). The subscript $h > 0$ is the bandwidth for the kernel estimator. Typically, $h \equiv h_n$ will depend on n, and $\lim_{n\to\infty} h_n = 0$.

Let

$$v_n \equiv v = 1 + h^2.$$

The following simple lemma establishes that $\hat{f}_h(z)$ and $\hat{f}'_h(z)$ are unbiased estimates of $f_{G_n,v}$, and $f'_{G_n,v}$ in the compound setting. It also further interprets their form. Let F_n denote the empirical distribution determined by $Z_1, \dots, Z_n$, i.e.,

$$F_n := \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}.$$

Lemma 6.3. *Let $h > 0$ and $v = 1 + h^2$. Suppose that $Z_i \sim N(\theta_i, 1)$, for $i = 1, \dots, n$, are independent. Then, we have the following relationships: for $z \in \mathbb{R}$,*

$$\hat{f}_h(z) = f_{F_n,h}(z), \qquad \hat{f}'_h(z) = f'_{F_n,h}(z),$$

and

$$\mathbb{E}\left[\hat{f}_h(z)\right] = f_{G_n,v}(z), \qquad \mathbb{E}\left[\hat{f}'_h(z)\right] = f'_{G_n,v}(z),$$

Proof. We write

$$\hat{f}_h(z) = \int \frac{1}{h}\phi\left(\frac{z-u}{h}\right)\, dF_n(u) = f_{F_n,h}(z),$$

since $h^{-1}\phi(\cdot/h)$ is the normal density with variance h^2. Similarly, we can derive the expression for $\hat{f}'_h(z)$.

Note that

$$\begin{aligned}\mathbb{E}\left[\hat{f}_h(z)\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{1}{h}\phi\left(\frac{z-Z_i}{h}\right)\right] &= \frac{1}{n}\sum_{i=1}^{n}\left\{\int \frac{1}{h}\phi\left(\frac{z-y}{h}\right)\phi\left(y-\theta_i\right)\, dy\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{v}\phi\left(\frac{z-\theta_i}{v}\right) = f_{G_n,v}(z),\end{aligned}$$

where the last step can be quickly seen by observing that the term within the braces is the density function of Z where $Y \sim N(\theta_i, 1)$ and $Z \mid Y \sim N(Y, h^2)$; thus, $Z \sim N(\theta_i, v^2)$ where $v^2 = 1 + h^2$. The expression for $\mathbb{E}\left[\hat{f}'_h(z)\right]$ can also be derived analogously.

□

In light of the above lemma, our estimator Eq. 6.25 may be written as

$$\hat{t}(z) = z + \frac{\hat{f}'_h(z)}{\hat{f}_h(z)} \approx z + \frac{f'_{G_n,v}(z)}{f_{G_n,v}(z)} \approx t^*_{\boldsymbol{\theta},v}(z) \approx t^*_{\boldsymbol{\theta}}(z), \tag{6.27}$$

where in the first approximation step we have replaced the terms by their expectations, as in Lemma 6.3, and in the second and third approximation steps we have used the fact that $v \equiv v_n = 1 + h_n^2 \to 1$ (or $h_n \to 0$).

A formal justification of the above heuristic is a bit more technical, that we only discuss briefly; see Theorem 1 and its proof in Brown and Greenshtein (2009). To state the main theoretical result here we consider the setup of a triangular array, where, at stage n, the parameter space, denoted Θ_n, is of dimension n. We write $\boldsymbol{\theta}^n = (\theta_1^n, \dots, \theta_n^n) \in \Theta_n$.

For every $\epsilon > 0$, we assume

$$|\theta_i^n| \le C_n = o(n^\epsilon), \qquad \text{for} \quad i = 1, \dots, n. \tag{6.28}$$

Such configurations include the interesting cases where $\theta_i^n = O(\sqrt{\log n})$.

We introduce the following slight modification for $\hat{t}(\cdot)$: a truncated estimator $\tilde{t}(\cdot)$ which at stage n is of the form

$$\tilde{t}(z) := sign(\hat{t}(z)) \times \min\{|\hat{t}(z)|, C_n\}, \qquad \text{for} \;\; z \in \mathbb{R}. \tag{6.29}$$

Note that we chose to truncate $\hat{t}(\cdot)$, so that $|\tilde{t}(\cdot)| \le C_n$. We can now state the main result on the risk consistency of $\tilde{t}$.

Theorem 6.1 (Brown and Greenshtein (2009))**.** *Consider a triangular array with Θ_n, as above, and sequences $\boldsymbol{\theta} \equiv \boldsymbol{\theta}^n \in \Theta_n$ satisfying Eq. 6.28. Let $v \equiv v_n = 1 + h_n^2 \to 1$ as $n \to \infty$ ($v_n > 1$ for every n), be any sequence such that:*

i. $h_n^2 \log n = o(1)$;
ii. $h_n^2 n^{\epsilon'} \to \infty$ *as* $n \to \infty$, *for every* $\epsilon' > 0$

Assume that, for some $\epsilon > 0$ and n_0,

$$R(\boldsymbol{\Delta}(\mathbf{Z}|t^*_{\boldsymbol{\theta}}), \boldsymbol{\theta})) > n^{\epsilon}, \qquad \text{for } n \geq n_0. \tag{6.30}$$

Then,

$$\limsup_{n\to\infty} \frac{R(\boldsymbol{\Delta}(\mathbf{Z}|\tilde{t}), \boldsymbol{\theta})}{R(\boldsymbol{\Delta}(\mathbf{Z}|t^*_{\boldsymbol{\theta}}), \boldsymbol{\theta})} = 1. \tag{6.31}$$

Remark (On the conditions in the theorem). Theorem 6.1} states that, in situations which are not too advantageous for the oracle so that its risk is of an order larger than n^{ϵ} for some $\epsilon > 0$, we may asymptotically do as well as the oracle by letting $v \equiv v_n$ approach 1 in the right way. Doing as well as the oracle means that the ratio of the risks approaches 1. Note that some condition resembling Eq. 6.30 is needed; if, for example, $\boldsymbol{\theta} = (0, \dots, 0) \in \mathbb{R}^n$, $n = 1, 2, \dots$, then the corresponding risk of the oracle is identically 0, and we can obviously not achieve such a risk by our estimator.

Note that condition (i) above means that h_n must go to zero as $n \to \infty$. But condition (ii) makes sure that h_n cannot go to zero fast; in fact it must converge to 0 very slowly (e.g., $h_n = (\log n)^{-\alpha}$ where $\alpha > 1/2$).

Proof. We do not give a complete proof of the result here but just indicate the general idea; see Brown and Greenshtein (2009) for the proof. The proof is divided into the two following steps.

Step 1: It can be shown that under assumption (i) we have (see [Brown and Greenshtein (2009); Lemma 2])

$$\lim_{n\to\infty} \frac{R(\boldsymbol{\Delta}(\mathbf{Z}|t^*_{\boldsymbol{\theta}}),\, \boldsymbol{\theta})}{R(\boldsymbol{\Delta}(\mathbf{Z}|t^*_{\boldsymbol{\theta},v}),\, \boldsymbol{\theta})} = \lim_{n\to\infty} \frac{\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^n \left(t^*_{\boldsymbol{\theta}}(Z_i) - \theta_i\right)^2\right]}{\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^n \left(t^*_{\boldsymbol{\theta},v}(Z_i) - \theta_i\right)^2\right]} = 1.$$

Step 2: It can be shown that for any $\epsilon'' > 0$ (arbitrarily small), under assumption (ii), we have

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\sum_{i=1}^n \left(t^*_{\boldsymbol{\theta},v}(Z_i) - \tilde{t}(Z_i)\right)^2\right] = o(n^{\epsilon''});$$

see Lemma 3 in Brown and Greenshtein (2009).

Then, we can write $R(\boldsymbol{\Delta}(\mathbf{Z}|\tilde{t}), \boldsymbol{\theta})$ as

$$
\begin{aligned}
& \mathbb{E}_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \left(\tilde{t}(Z_i) - \theta_i \right)^2 \right] \\
\leq \quad & (1 + \frac{1}{\eta}) \, \mathbb{E}_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \left(t^*_{\boldsymbol{\theta},v}(Z_i) - \tilde{t}(Z_i) \right)^2 \right] + (1+\eta) \, \mathbb{E}_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \left(t^*_{\boldsymbol{\theta},v}(Z_i) - \theta_i \right)^2 \right] \\
= \quad & (1 + \frac{1}{\eta}) o(n^{\epsilon''}) + (1+\eta) \left\{ 1 + o(1) \right\} \mathbb{E}_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \left(t^*_{\boldsymbol{\theta}}(Z_i) - \theta_i \right)^2 \right]
\end{aligned}
$$

for every $\eta > 0$. for every $\eta > 0$. As $t^*_{\boldsymbol{\theta}}$ minimizes the risk Eq. 6.17, we have

$$
1 \leq \frac{\mathbb{E}_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \left(\tilde{t}(Z_i) - \theta_i \right)^2 \right]}{\mathbb{E}_{\boldsymbol{\theta}} \left[\sum_{i=1}^{n} \left(t^*_{\boldsymbol{\theta}}(Z_i) - \theta_i \right)^2 \right]} \leq (1 + \frac{1}{\eta}) o(n^{\epsilon'' - \epsilon}) + (1+\eta) \left\{ 1 + o(1) \right\}.
$$

As ϵ'' can be chosen to be smaller than ϵ (recall assumption Eq. 6.30), we have

$$
1 \leq \limsup_{n \to \infty} \frac{R(\boldsymbol{\Delta}(\mathbf{Z}|\tilde{t}), \boldsymbol{\theta})}{R(\boldsymbol{\Delta}(\mathbf{Z}|t^*_{\boldsymbol{\theta}}), \boldsymbol{\theta})} \leq 1 + \eta.
$$

As $\eta > 0$ is arbitrary, we have the desired result.

□

7 G-modeling

In this section we study the G-modeling approach to empirical Bayes estimation. Here the idea is to directly estimate the prior G from the data first and then apply the learned Bayes rule. Recall, that in the Bayesian setting, our goal is to "estimate" the random parameter M_i based on $\mathbf{Z} = (Z_1, \dots, Z_n)$ where:

$$M_i \overset{iid}{\sim} G, \qquad Z_i \mid M_i = \theta \overset{ind}{\sim} p(\cdot \mid \theta), \tag{7.1}$$

and G is the unknown prior distribution supported on $\Theta \subset \mathbb{R}$.

Here, the target is the Bayes procedure t_G — the estimator that minimizes the Bayes risk under G.

7.1 General maximum likelihood empirical Bayes (GMLEB)

The first task in G-modeling is to estimate G directly from the data $\mathbf{Z}$. In this chapter we will consider using the *nonparametric maximum likelihood estimator* (NPMLE) of G to accomplish this task; cf. Section 1.4.2 in Chapter 1 where we minimized an appropriate Kolmogorov-Smirnov distance to obtain an estimator of G.

The NPMLE is any $\widehat{G}_n \in \mathcal{P}(\Theta)$—the class of all Borel probability distributions on Θ—that maximizes the marginal likelihood of the observations $\mathbf{Z}$ drawn from Eq. 7.1. Note that marginally, the observations are i.i.d., and the ith observation Z_i is distributed according to the mixture model with density

$$f_G(z) := \int p(z \mid \theta)\, dG(\theta), \qquad \text{for } z \in \mathbb{R}. \tag{7.2}$$

Formally, an NPMLE is any maximizer

$$\widehat{G}_n \in \operatorname*{argmax}_{G \in \mathcal{P}(\Theta)} \sum_{i=1}^{n} \log f_G(Z_i). \tag{7.3}$$

The idea of finding the NPMLE of a latent distribution is an old one. The idea was suggested in an abstract by Robbins (1950) (also see Robbins (1951)), and later received substantial theoretical development by Kiefer and Wolfowitz (1956). In Section 7.2 we provide some basic

characterizations/properties of the NPMLE $\widehat{G}_n$ in Eq. 7.3, where we maximize the (marginal) log-likelihood over the infinite dimensional space of all distributions.

Jiang and Zhang (2009) proposed the general maximum likelihood EB (GMLEB) method in which one first estimates G by the NPMLE $\widehat{G}_n$ and then one plugs this estimator into the oracle general EB rule t_G, i.e., one replaces the unknown prior G in the oracle Bayes rule t_G by $\widehat{G}_n$. Thus, the GMLEB estimates the unobserved latent variable M_i by

$$\hat{\theta}_i = t_{\widehat{G}_n}(Z_i), \qquad \text{for } \ i = 1, \dots, n. \tag{7.4}$$

Clearly, the GMLEB estimator is completely nonparametric and does not require any restriction, regularization, bandwidth selection or other forms of tuning. Further, the GMLEB procedure is applicable to any model of the form Eq. 7.1 and any Bayes decision t_G; compare this with the F-modeling approach which can be quite restrictive as it crucially needs Tweedie-type formulas to express posterior quantities (e.g., posterior mean) in terms of the marginal f_G. The GMLEB is also appealing since the function $t_{\widehat{G}_n}(\cdot)$ enjoys all analytical properties of the Bayes rule; e.g., in the normal means problem under squared error loss, $t_{\widehat{G}_n}(\cdot)$ is monotonic, infinitely differentiable and more.

Further, the theoretical results of Jiang and Zhang (2009), in the normal means problem, affirmed that by aiming at the minimum risk of all separable estimators, this greedier general EB approach realizes significant risk reduction over linear (and threshold methods) for a wide range of the unknown signal vectors for moderate and large samples. The authors prove that the risk of the GMLEB estimator is within an infinitesimal fraction of the general EB benchmark when the risk is of the order $n^{-1}(\log n)^5$ or greater, depending on the magnitude of the unknown means (see Theorem 7.3 in Section 7.4).

7.2 Characterization and basic properties of the NPMLE

In this section, we establish some basic properties of solutions to the nonparametric maximum likelihood problem Eq. 7.3, including: (i) *existence*, (ii) *uniqueness*, (iii) *discreteness* of solutions $\widehat{G}_n$, and (iv) *bounds on the support* of $\widehat{G}_n$. These results provide a foundation both for computing $\widehat{G}_n$ (Section 7.3) and for understanding its statistical properties (Section 7.4). Our treatment here is adopted from B. G. Lindsay (1995); see e.g.,Theorems 18-21.

Before starting the formal result, let us briefly discuss the main ideas in studying the solution(s) of Eq. 7.3. We first introduce some notation. Let m denote the number of distinct values in the set $\{Z_1, \dots, Z_n\}$, and let the distinct values be denoted as $y_1 < \dots < y_m$. Let $n_j := \#\{i : Z_i = y_j\}$, for $1 \le j \le m$, denote the number of observations Z_i that are equal to y_j. Then, problem Eq. 7.3 can be thought of as maximizing the objective function

$$\ell(G) := \sum_{j=1}^{m} n_j \log f_G(y_j), \tag{7.5}$$

over the class of all probability distributions G on $\Theta \subset \mathbb{R}$, where f_G is as defined in Eq. 7.2.

Reformulating the problem. The key to putting this problem into a standard optimization framework is to recognize that the objective function $\ell(G)$ above depends directly on the possible values of the *mixture likelihood* vector

$$\mathbf{L}(G) := (f_G(y_1), \dots, f_G(y_m)) \in \mathbb{R}^m.$$

We change our perspective on this problem from maximizing the log-likelihood Eq. 7.5 over all latent distributions G into the problem of determining which of the eligible classes of mixture likelihood vectors $\mathbf{L}(G)$ gives the largest value to the log-likelihood.

Step 1. Construct the *feasible region* of $\mathbb{R}^m$. It will be the set of all possible fitted values of the likelihood vector:

$$\mathcal{M} := \Big\{\mathbf{L}(G) = (f_G(y_1), \dots, f_G(y_m)) : G \text{ is a probability distribution}\Big\} \subset [0, \infty)^m.$$

First, note that $\mathcal{M}$ is a convex set in $\mathbb{R}^m$. As we shall see, this together with the concavity of the objective function $\ell(G)$, ensures that Eq. 7.3 is in a class of nice optimization problems.

Step 2. We can now redefine the maximization problem:

$$\max_{\mathbf{p} \in \mathcal{M}} \sum_{i=1}^{m} n_j \log p_j =: \ell(\mathbf{p}) \tag{7.6}$$

where the *objective function* here is $\ell(\mathbf{p})$—a strictly concave function on the positive orthant.[1] Thus, under appropriate conditions, we can expect a unique maximizer $\hat{\mathbf{p}} \in \mathcal{M}$.

Step 3. Solve for the NPMLE $\widehat{G}_n$ by solving from the known $\hat{\mathbf{p}}$ for the latent distribution $\widehat{G}_n$ via the m equations

$$\mathbf{L}(\widehat{G}_n) = \hat{\mathbf{p}}.$$

The following result is from B. G. Lindsay (1995) (also see Bruce G. Lindsay (1983)).

Theorem 7.1 (Existence and support size)**.** *Suppose that: (i) $p(z \mid \theta)$ is a continuous function in θ and $\lim_{|\theta| \to \infty} p(z \mid \theta) = 0$, for every $z \in \mathbb{R}$, (ii) $p(\cdot \mid \cdot)$ is upper bounded, and (iii) $\mathcal{M}$ contains at least one point with positive likelihood.*[2] *Then there exists unique $\hat{\mathbf{p}} \in \partial\mathcal{M}$, the boundary of $\mathcal{M}$, such that $\hat{\mathbf{p}}$ maximizes $\ell(\mathbf{p})$ (see Eq. 7.6) over $\mathcal{M}$. Further, the point $\hat{\mathbf{p}}$ can be expressed as $(f_{\widehat{G}_n}(y_1), \dots, f_{\widehat{G}_n}(y_m))$ where $\widehat{G}_n$ has m or fewer points of support.*

[1] **Exercise**: Show this.

[2] If no point in $\mathcal{M}$ has positive likelihood, then the uniqueness of the maximum must fail, because then all elements have likelihood zero.

Proof. Consider the following set where G varies over all sub-probability distributions [3]:

$$\mathcal{M}_{sub} := \Big\{\mathbf{L}(G) = (f_G(y_1), \dots, f_G(y_m)) : G \text{ is a sub-probability distribution}\Big\} \supseteq \mathcal{M}.$$

Moreover, if $p(\cdot \mid \cdot)$ is upper bounded, then $\mathcal{M}_{sub}$ is a compact set.[4]

Let $\mathbf{p}_0 \in \mathcal{M}$ be one point with positive likelihood as stated in the assumption), and say $\ell(\mathbf{p}_0) = c_0 > -\infty$. Then the set

$$\mathcal{L} := \{\mathbf{p} \in \mathcal{M}_{sub} : \ell(\mathbf{p}) \geq c_0\} \subset \mathcal{M}_{sub}$$

is compact[5] (i.e., closed and bounded). Note that $\ell(\cdot)$ is a continuous function on the compact set $\mathcal{L}$, and hence attains its maximum on $\mathcal{L}$. Further, $\max_{\mathbf{p}\in\mathcal{M}_{sub}} \ell(\mathbf{p}) = \max_{\mathbf{p}\in\mathcal{L}} \ell(\mathbf{p})$.

We next show that any maximizer $\hat{\mathbf{p}} = (f_G(y_1), \dots, f_G(y_m))$ of $\ell(\cdot)$ over $\mathcal{M}_{sub}$ must lie in $\mathcal{M}$. Suppose not, i.e., suppose that G is not a proper probability distribution (i.e., $0 < G(\Theta) < 1$). Define $\tilde{G}(A) := G(A)/G(\Theta)$, for any Borel $A \subset \Theta$. Then $\tilde{G}$ is a valid probability distribution and $q_j := f_{\tilde{G}}(y_j) = f_G(y_j)/G(\Theta)$, for all $j = 1, \dots, m$. Moreover as the function log is strictly increasing, $\ell(q_1, \dots, q_m) > \ell(\hat{\mathbf{p}})$, yielding a contradiction. Therefore, $\max_{\mathbf{p}\in\mathcal{M}_{sub}} \ell(\mathbf{p}) = \max_{\mathbf{p}\in\mathcal{M}} \ell(\mathbf{p}) = \ell(\hat{\mathbf{p}})$, thereby showing the existence of a maximizer $\hat{\mathbf{p}}$ of the objective $\ell(\cdot)$ over $\mathcal{M}$.

Further, as $\mathcal{M}$ is convex, and the objective function $\ell(\cdot)$ is strictly concave, it takes on a unique maximum value.

To see this, suppose that $\ell(\cdot)$ does not have a unique maximum over $\mathcal{M}$. Then there exists $\hat{\boldsymbol{\pi}}_1 \neq \hat{\boldsymbol{\pi}}_2 \in \mathcal{M}$ such that

$$\ell(\hat{\boldsymbol{\pi}}_1) = \ell(\hat{\boldsymbol{\pi}}_2) = \max_{\mathbf{p}\in\mathcal{M}} \ell(\mathbf{p}).$$

Define $\hat{\boldsymbol{\pi}} := (\hat{\boldsymbol{\pi}}_1 + \hat{\boldsymbol{\pi}}_2)/2$. Then $\hat{\boldsymbol{\pi}} \in M$, as $\mathcal{M}$ is a convex set, and by the strictly concavity of $\ell(\cdot)$ we get

$$\ell(\hat{\boldsymbol{\pi}}) > \frac{1}{2} f(\hat{\boldsymbol{\pi}}_1) + \frac{1}{2} f(\hat{\boldsymbol{\pi}}_2) = \max_{\mathbf{p}\in\mathcal{M}} \ell(\mathbf{p}),$$

which is a contradiction. Thus, $\ell(\cdot)$ has a unique maximum over $\mathcal{M}$.

Further, as $\ell(\cdot)$ is a strictly coordinate-wise increasing function, this unique maximum value $\hat{\mathbf{p}}$ has to be at the boundary of $\mathcal{M}$ (otherwise we could increase the objective value $\ell(\hat{\mathbf{p}})$ to $\ell(\hat{\mathbf{p}} + \epsilon \mathbf{1}_m)$ for some small enough $\epsilon > 0$).

The last part of the result follows from Carathéodory's theorem for convex sets:

Theorem (Carathéodory): Let S be a set in the m-dimensional Euclidean space $\mathbb{R}^m$. Every element $x \in \text{conv}(S)$ can be expressed as a convex combination of at most $m+1$ elements of S. If x is in the boundary of $\text{conv}(S)$, then $m+1$ can be replaced by m.

[3] Thus G is a measure with total mass at most 1.

[4] **Exercise**: Show this.

[5] **Exercise**:Show this

□

Remark (On the assumptions of Theorem 7.1). The requirement that the set $\mathcal{L}$ be closed is more of a technical requirement to make the theory simple[6] (note that the boundedness of the curve $\mathcal{L}$ is essential, because if the likelihood vectors have unbounded components, then one can construct unbounded likelihoods). There will be cases, such as the normal location mixture example, where the parameter θ can vary over the entire $\mathbb{R}$. To ensure closedness of $\mathcal{L}$, we must include the left- and right-hand limits; in the normal example, the likelihood vector $\mathbf{L}_\theta$ converges to $\mathbf{0} \in \mathbb{R}^m$ in both directions, as $\theta \to \pm\infty$. We can include this limit point in the set $\mathcal{L}$ without real consequence because it can never appear in the maximizing mixture.[7]

Existence and discreteness. The first statement of Theorem 7.1 guarantees the existence of a solution; in particular, the existence of a discrete $\widehat{G}_n$ with no more than m support points—the number of distinct data points $Z_1, \dots, Z_n$. Thus we typically write:

$$\widehat{G}_n = \sum_{j=1}^{\hat{k}} \hat{w}_j \delta_{\hat{a}_j} \qquad \text{where} \quad \hat{w}_j \geq 0, \ \sum_{j=1}^{\hat{k}} \hat{w}_j = 1 \ \text{ and } \ \hat{a}_j \in \Theta, \tag{7.7}$$

with $\hat{k} \leq m$ providing an upper bound on the complexity of at least one solution.[8] This implies that $\widehat{G}_n$ may be taken to be the maximum likelihood solution to a $\hat{k}$-component mixture model where $\hat{k}$ is selected in a data dependent manner. Since finite mixture models are nested by the number of components and $\hat{k} \leq m$, we may also say in general that $\widehat{G}_n$ is the maximum likelihood solution to an m-component mixture model.

However, in practice, the number of components $\hat{k}$ is typically much smaller than m. For instance, in the univariate Gaussian location mixture model Polyanskiy and Wu (2020) establish a much stronger bound of $\hat{k} = O_P(\log n)$ under certain conditions on the prior distribution G.

Remark (On the uniqueness of $\widehat{G}_n$). We note that while the uniqueness of $\hat{\mathbf{p}}$ is guaranteed by the above lemma, there may be more than one $\widehat{G}_n$ which satisfies $\mathbf{L}(\widehat{G}_n) = \hat{\mathbf{p}}$. The uniqueness of $\widehat{G}_n$ is a more delicate issue. For example, it is known that in the univariate normal location

[6] If $\mathcal{L}$ is not closed, the theorem can be applied to the closure; one must then determine if the limit points show up in the maximizing mixture and, if so, how to interpret them.

[7] Define $\mathcal{M} := \big\{(f_G(y_j))_{j=1}^m : G \in \mathcal{P}(\Theta)\big\} \cup \{\mathbf{0}\}$. Observe that

$$\mathcal{M} = \operatorname{conv}(\mathcal{L}), \qquad \text{where} \qquad \mathcal{L} := \big\{(p(y_j \mid \vartheta))_{j=1}^m : \vartheta \in \mathbb{R}\big\} \cup \{\mathbf{0}\}.$$

Since $\vartheta \mapsto (p(y_j \mid \vartheta))_{j=1}^m$ is continuous and $\lim_{|\vartheta| \to \infty} (p(y_j \mid \vartheta))_{j=1}^m = \mathbf{0}$, the set $\mathcal{L}$ is closed, and by boundedness of the Gaussian likelihood, $\mathcal{L}$ is compact. Hence $\mathcal{M} \subset \mathbb{R}^m$ is convex and compact, and $\ell(\mathbf{p}) := \sum_{j=1}^m n_j \log p_i$ is strictly concave over $\mathcal{M}$.

[8] The bound $\hat{k} \leq m$ is tight: for each $m \geq 1$, there are sequences of observations $(y_j)_{j=1}^m$ such that the smallest number of components $\hat{k}$ of any solution to Eq. 7.3 is precisely m, see e.g., p. 116 in B. G. Lindsay (1995). In particular, in a normal location mixture model, one can construct sets of data for which the bound m is attained simply by spreading the observations far apart.

mixture model, the uniqueness holds; see Bruce G. Lindsay (1983) and Bruce G. Lindsay and Roeder (1993). However, it does not hold true for the multivariate version of the problem; see e.g., Soloff, Guntuboyina, and Sen (2021). We will come back to this later, if time permits.

Gradient characterization. Next we consider a way of characterizing whether a given latent distribution, say G_0, is the NPMLE.

To do this we form a path in the space of distributions from G_0 to any other distribution, say G_1, by letting $G_\alpha := (1-\alpha)G_0 + \alpha G_1$, for $\alpha \in [0,1]$. For every α, this generates an intermediate distribution, with $\alpha = 0$ and 1 corresponding to the original two distributions of interest.

Next, we compute the log-likelihood along this path, obtaining a one parameter log-likelihood function

$$\ell(\alpha) := \sum_{j=1}^m n_j \log f_{G_\alpha}(y_j).$$

The derivative of $\ell(\alpha)$ at $\alpha = 0$ is the *directional derivative* corresponding to this path from G_0 to G_1 and it has the simple form

$$\begin{aligned} D(G_0, G_1) := \left.\frac{\partial \ell(\alpha)}{\partial \alpha}\right|_{\alpha=0} &= \lim_{\alpha \downarrow 0} \frac{\sum_{j=1}^m n_j [\log f_{(1-\alpha)G_0+\alpha G_1}(y_j) - \log f_{G_0}(y_j)]}{\alpha} \\ &= \sum_{j=1}^m n_j \lim_{\alpha \downarrow 0} \frac{[\log f_{(1-\alpha)G_0+\alpha G_1}(y_j) - \log f_{G_0}(y_j)]}{\alpha} \\ &= \sum_{j=1}^m \frac{n_j}{f_{G_0}(y_j)} \int p(y_j \mid \theta)\, d(G_1 - G_0)(\theta) \\ &= \sum_{i=1}^m n_j \left\{ \frac{f_{G_1}(y_j)}{f_{G_0}(y_j)} - 1 \right\} \end{aligned}$$

where we note that

$$f_{(1-\alpha)G_0+\alpha G_1}(y_j) = f_{G_0}(y_j) + \alpha \int p(y_j \mid \theta)\, d(G_1 - G_0)(\theta)$$

and we have used the fact that $\lim_{\beta \downarrow 0} \frac{\log(x+\beta) - \log x}{\beta} = \frac{1}{x}$.

Next, it is clear that if the gradient function $D(G_0, G_1)$ takes on positive values at any G_1, then the likelihood along the path from G_0 in the direction of G_1 is increasing at G_0, so that G_0 cannot be the NPMLE.

Theorem 7.2. *$\widehat{G}_n \in \mathcal{P}(\Theta)$ solves Eq. 7.3 if and only if*

$$D(\widehat{G}_n, \vartheta) \le 0 \text{ for all } \vartheta \in \Theta, \qquad \text{where} \quad D(G, \vartheta) := \sum_{j=1}^m n_j \left\{ \frac{p(y_j \mid \vartheta)}{f_G(y_j)} - 1 \right\}. \tag{7.8}$$

Further, the support of any NPMLE $\widehat{G}_n$ is contained in the zero set $\Theta_0 := \{\vartheta \in \Theta : D(\widehat{G}_n, \vartheta) = 0\}$.

Proof. The following uses similar techniques as Section 5.2 of B. G. Lindsay (1995). By convexity, the first-order optimality condition for $\widehat{G}_n$ is

$$D(\widehat{G}_n, G) \leq 0 \qquad \text{for all } G \in \mathcal{P}(\Theta) \tag{7.9}$$

where

$$\begin{aligned} D(\widehat{G}_n, G) &:= \lim_{\alpha \downarrow 0} \frac{\sum_{j=1}^m n_j [\log f_{(1-\alpha)\widehat{G}_n + \alpha G}(y_j) - \log f_{\widehat{G}_n}(y_j)]}{\alpha} \\ &= \sum_{j=1}^m \frac{n_j}{f_{\widehat{G}_n}(y_j)} \left(f_G(y_j) - f_{\widehat{G}_n}(y_j) \right) = \sum_{j=1}^m n_j \left\{ \frac{f_G(y_j)}{f_{\widehat{G}_n}(y_j)} - 1 \right\}. \end{aligned}$$

When $G = \delta_\vartheta$ is a point mass we write $D(\widehat{G}_n, \vartheta)$ instead of $D(\widehat{G}_n, G)$. It suffices to check $D(\widehat{G}_n, \vartheta) \leq 0$ for all $\vartheta \in \mathbb{R}$ because $D(\widehat{G}_n, G) = \int D(\widehat{G}_n, \vartheta)\, dG[\vartheta]$.

Note that from Eq. 7.8 we obviously have, by integrating, $\int D(\widehat{G}_n, \theta)\, d\widehat{G}_n(\theta) \leq 0$. However,

$$\int D(\widehat{G}_n, \theta)\, d\widehat{G}_n(\theta) = \sum_{j=1}^m n_j \left\{ \frac{\int p(y_j \mid \theta)\, d(\theta)}{f_G(y_j)} - 1 \right\} = 0.$$

Thus, letting $\hat{\theta}_n \sim \widehat{G}_n$, the random variable $D(\widehat{G}_n, \hat{\theta}_n) \leq 0$ but $\mathbb{E}\left[D(\widehat{G}_n, \hat{\theta}_n) \mid \widehat{G}_n \right] = 0$. Therefore, $D(\widehat{G}_n, \hat{\theta}_n) = 0$ a.s., and hence the support of any NPMLE $\widehat{G}_n$ is contained in the zero set $\Theta_0 := \{\vartheta \in \Theta : D(\widehat{G}_n, \vartheta) = 0\}$.

□

Support point properties. The second part of the lemma characterizes the location of the support points $\hat{a}_j$ of $\widehat{G}_n$. The result is that if ξ is a support point for any NPMLE $\widehat{G}_n$, then $D(\widehat{G}_n, \xi) \equiv D(\widehat{G}_n, \delta_\xi) = 0$. Together with the gradient inequality in Eq. 7.8 this implies that the support points will be local maxima of the gradient function $D(\widehat{G}_n, \vartheta)$.[9] This result is also very useful in proofs of the uniqueness of the NPMLE $\widehat{G}_n$.

7.3 Computation

Since the NPMLE $\widehat{G}_n$ is completely nonparametric, the support points $\{a_j, 1 \leq j \leq \hat{k}\}$ and weights $\{w_j, 1 \leq j \leq \hat{k}\}$ in Eq. 7.7 are selected (or computed) solely to maximize the log-likelihood in Eq. 7.5. There are quite a few possible algorithms for solving Eq. 7.3 to compute

[9] One of the consequences of this result is that a gradient-based algorithm need not keep track of the support points in $\widehat{G}_n$ because they can be recovered from the gradient function at the end of the algorithm.

the NPMLE $\widehat{G}_n$. The most classical approach to solving Eq. 7.3 is to use the expectation-maximization (EM) algorithm; see N. Laird (1978). However, different EM-initializations can lead to different versions of $\widehat{G}_n$, which can then result in different values of $t_{\widehat{G}_n}(Z_i)$. Further, the EM may converge very slowly as is well-documented in the literature; see e.g.,Koenker and Mizera (2014).

As the "nonparametric" domain (or constraint set) $\mathcal{P}(\Theta)$ is convex and the objective $\ell(G)$ is concave, the NPMLE $\widehat{G}_n$ solves a convex optimization problem, and tools from convex optimization may be leveraged to find principled approximations[10] to $\widehat{G}_n$. See Böhning (1999) for a book length treatment of different computational algorithms that can be used to solve Eq. 7.3.

A natural strategy, that has become quite popular in the last decade (advocated by Koenker and Mizera (2014)), is to discretize the space Θ (or a subset thereof) where $\widehat{G}_n$ can put mass.[11] For example, we can consider an appropriate compact interval and fix an equispaced grid $c_1 < c_2 < \ldots < c_N$ in this interval, for some $N \geq 1$. We can then maximize the marginal log-likelihood over all distributions supported on the aforementioned grid, which leads to the following finite-dimensional convex optimization problem:

$$\max_{(\pi_1,\ldots,\pi_N)\in\mathbb{R}^N} \sum_{j=1}^{m} n_j \log\left(\sum_{i=1}^{N} L_{ji}\pi_i\right) \qquad \text{s.t.} \quad \pi_i \geq 0, \;\; \forall\, i = 1,\ldots,N, \;\; \sum_{i=1}^{N} \pi_i = 1, \tag{7.10}$$

where $L_{ji} := p(y_j \mid c_i)$, for all i, j; here $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_N)$, belonging to the probability simplex in $\mathbb{R}^N$, is the variable of interest. In particular, Koenker and Mizera (2014) proposed solving the dual formulation of Eq. 7.10 using off-the-shelf interior point based solvers. In fact, the routine `KWDual` in the `R` package `REBayes` (Koenker and Gu 2017) adopts the commercial interior point solver Mosek to solve the dual of Eq. 7.10. Compared to the EM, modern convex optimization methods can be more efficient and stable. See Kim, Carbonetto, Stephens, and Anitescu (2020) and Y. Zhang, Cui, Sen, and Toh (2022) for some recent algorithms that can solve Eq. 7.10 for larger sample sizes (e.g., $m \approx 10^6$). It can be shown that as the grid becomes more dense, the NPMLE $\widehat{G}_n$ computed from these discrete approximations converge to a solution of the infinite dimensional problem Eq. 7.3; we refer to Soloff, Guntuboyina, and Sen (2021) for an analysis that quantify the discretization error in the normal location mixture model.

7.4 Theoretical Properties

In this section we study some of the optimality properties of the GMLEB procedure. We focus on the Gaussian sequence model in the compound setting (under quadratic loss), where we

[10] As the set $\mathcal{P}(\Theta)$ is infinite-dimensional, we need to approximate it by a finite-dimensional set.

[11] For example, it can be shown that the $\widehat{G}_n$ in the Gaussian location mixture only puts mass on points between the minimum and maximum of the observations.

observe

$$Z_i \overset{ind}{\sim} N(\theta_i, 1), \qquad \text{for } i = 1, \dots, n,$$

and the unknown parameter is $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n$. From the fundamental theorem of compound decisions, we have (recall that $G_n = \frac{1}{n}\sum_{i=1}^n \delta_{\theta_i}$)

$$R^*(G_n) := \mathbb{E}_{G_n}\left[(t_{G_n}(Z) - M)^2\right] = \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{n}\sum_{i=1}^n (t^*_{\boldsymbol{\theta}}(Z_i) - \theta_i)^2\right] = \min_{t(\cdot)} \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{n}\sum_{i=1}^n (t(Z_i) - \theta_i)^2\right].$$

Thus, the best separable estimator $t(\cdot)$ minimizes the Bayes risk under prior G_n. The general EB approach seeks procedures which approximate this Bayes rule $t^*_{\boldsymbol{\theta}}(\cdot) \equiv t_{G_n}(\cdot)$ or approximately achieve the risk benchmark $R^*(G_n)$ above. Further, the Bayes rule $t^*_{\boldsymbol{\theta}}$ provides a natural benchmark against which we can compare the performance of the GMLEB estimator $t_{\widehat{G}_n}(\cdot)$.

There are at least three ways of quantifying the performance of the GMLEB method:

(1) We can study the risk behavior of the estimated Bayes procedure $t_{\widehat{G}_n}$, and compare it to the Bayes optimal procedure $t^*_{\boldsymbol{\theta}}$.

(2) The NPMLE $\widehat{G}_n$ immediately yields an estimator of the average mixing density[12] f_{G_n}, namely $f_{\widehat{G}_n}$; we can quantify this estimation accuracy.

(3) In the compound setting the NPMLE $\widehat{G}_n$ can be thought of as estimating the unknown G_n. We can quantify the accuracy of this estimator directly.

We first discuss 1. above, i.e., the performance of $t_{\widehat{G}_n}$ in estimating $t^*_{\boldsymbol{\theta}}$. Jiang and Zhang (2009) has a comprehensive study of this problem under various assumptions on G_n, e.g., G_n is light-tailed, heavy-tailed, or *sparse*.[13] The following is one such result; its proof is quite involved and we skip it here (also see Saha and Guntuboyina (2020) for similar results in the multivariate Gaussian sequence model with detailed proofs).

Theorem 7.3 (Theorem 1 in Jiang and Zhang (2009))**.** *Let* $\mathbf{Z} \sim N(\boldsymbol{\theta}, I_n)$ *where*

$$\boldsymbol{\theta} \in \Theta_n := \{\boldsymbol{\theta} = (\theta_1, \dots, \theta_n) \in \mathbb{R}^n : \exists\, b_n \ s.t. \ \max_{1 \le i \le n} |\theta_i - b_n| = O(\sqrt{\log n})\}.$$

Then,

$$\lim_{n\to\infty} \sup_{\boldsymbol{\theta}\in\Theta_n} \frac{\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{n}\sum_{i=1}^n (t_{\widehat{G}_n}(Z_i) - \theta_i)^2\right]}{R^*(G_n)} \le 1,$$

provided $nR^*(G_n)/(\log n)^5 \to \infty$.

[12] See Eq. 7.11 below for more details on this.

[13] The study of EB methods for sparse priors has received quite a bit of attention mainly because of the connections to multiple hypothesis testing.

The above result says that the GMLEB estimator $t_{\widehat{G}_n}$ has risk which is *uniformly* close to the risk of the oracle for a large set of parameter vectors, namely Θ_n. Note that although $R^*(G_n) \leq 1$, for "sparse" sequences it can be much smaller (close to 0; e.g., for $\theta_1 = ... = \theta_n = 0$). The condition $nR^*(G_n)/(\log n)^5 \to \infty$ in Theorem 7.3 shows that $t_{\widehat{G}_n}$ has near optimal performance for a broad range of "sparse" parameters (as $R^*(G_n)$ can be as small as $1/n$, up to logarithmic factors).

Although the above result showcases the remarkable performance of the GMLEB estimator $t_{\widehat{G}_n}$ for a broad range of settings, if $R^*(G_n) \approx O(1)$, then the above result does not provide any rates as to how fast the risk of $t_{\widehat{G}_n}$ approaches that of the oracle t_{G_n}. To formalize this, we can study he *regret* of $t_{\widehat{G}_n}$ in estimating the general EB oracle rule $t^*_{G_n}(\cdot)$:

$$r_{n,\boldsymbol{\theta}}(t^*_{G_n}) := \mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{n}\sum_{i=1}^n (t_{\widehat{G}_n}(Z_i) - \theta_i)^2\right] - R^*(G_n).$$

Jiang and Zhang (2009) prove rates for the regret; in particular, it shows *almost parametric rates* for the difference between the square roots of the risks above. Also see Saha and Guntuboyina (2020) for results of a similar flavor.

In regard to point 2 above, in Section 7.4.1 below, we study the accuracy of $f_{\widehat{G}_n}$ for estimating f_{G_n} in the Hellinger metric; see C.-H. Zhang (2009) and Saha and Guntuboyina (2020) for more general results in this direction.

Let us now discuss point 3 above. Results that study the consistency of $\widehat{G}_n$ (for estimating G_n) date back to the seminal work of Kiefer and Wolfowitz (1956). In Section 7.4.2 below we state and prove such a result that is applicable for many symmetric location mixtures beyond the Gaussian model (under the Bayesian setting). In the recent paper Soloff, Guntuboyina, and Sen (2021), the authors provide finite sample rates for $\widehat{G}_n$, in the 2-Wasserstein metric. Direct estimation of G_n is related to deconvolution problems in statistics where the minimax convergence rates can be very slow, that is, polynomial in $1/\log n$ (see e.g.,C.-H. Zhang (1990), Fan (1991)). Although this shows that direct estimation of G_n is a hard problem, the results of Jiang and Zhang (2009) show that EB estimation (in particular, estimating $t^*_{\boldsymbol{\theta}}$) is not impacted by this slow rate.

7.4.1 The Hellinger accuracy of $f_{\widehat{G}_n}$

Observe that once the NPMLE $\widehat{G}_n$ is computed, we can also obtain $f_{\widehat{G}_n}$ — a natural estimator of the marginal density f_G in the Bayesian setting. In the compound setting, $f_{\widehat{G}_n}$ is really estimating the *average mixing density* f_{G_n}. To see this, observe that the expected value of the log-likelihood is

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{1}{n}\sum_{i=1}^n \log f_G(Z_i)\right] = \frac{1}{n}\sum_{i=1}^n \int \log f_G(z)\phi(z-\theta_i)\,dz = \int \log f_G(z) f_{G_n}(z)\,dz \tag{7.11}$$

which is uniquely maximized at $f_G = f_{G_n}$ (as the right side of the above display is equivalent to Kullback-Leibler divergence between f_G and f_{G_n}).

The following result quantifies the estimation accuracy of the NPMLE $f_{\widehat{G}_n}$ as an estimator of f_{G_n} in the simple setting where the θ_i's lie in a fixed compact interval.

Theorem 7.4. *Suppose that G_n is supported on a compact interval $[-R, R]$, for all $n \geq 1$. Then,*[14]

$$\mathbb{E}\left[\mathfrak{h}^2(f_{\widehat{G}_n}, f_{G_n})\right] \lesssim_R \frac{(\log n)^2}{n},$$

where $\mathfrak{h}^2(f_1, f_2) := \frac{1}{2}\int(\sqrt{f_1} - \sqrt{f_2})^2$ is the squared Hellinger distance between the densities f_1 and f_2.

Theorem Theorem 7.4 shows that estimation of f_{G_n} in the Hellinger distance is a relatively easy statistical task that may be achieved at the parametric rate $1/\sqrt{n}$, up to logarithmic factors, by the NPMLE $f_{\widehat{G}_n}$; see C.-H. Zhang (2009) and Saha and Guntuboyina (2020) for more results of a similar flavor under less restrictive assumptions on G_n.

Proof. The general theory of the rates of convergence of maximum likelihood estimators from, say e.g., van der Vaart and Wellner (1996),can be used to bound $\mathfrak{h}^2(f_{\widehat{G}_n}, f_{G_n})$. This general theory requires bounds on the covering numbers [15] of the underlying class of densities. In our context, we need to bound covering numbers of the class

$$\mathcal{F} := \{f_G : G \in \mathcal{P}(\mathbb{R})\}. \tag{7.12}$$

Our main covering number result for $\mathcal{F}$ is stated next (we do not prove it here).

Lemma 7.1 (Lemma 2 in C.-H. Zhang (2009))**.** *There exists a universal constant $C^* > 0$ such that, for all $0 < \epsilon \leq \frac{1}{\sqrt{2\pi}}$ and $B > 0$, we have*

$$\log N(\epsilon, \mathcal{F}, \|\cdot\|_B^\infty) \leq C^*(\log \epsilon)^2 \max\left\{\frac{B}{\sqrt{|\log \epsilon|}}, 1\right\}.$$

Here $N(\epsilon, \mathcal{F}, \|\cdot\|_B^\infty)$ is the ϵ-cover of the set $\mathcal{F}$ under the pseudometric $\|\cdot\|_B^\infty$, where $\|f\|_B^\infty := \sup_{x\in[-B,B]} |f(x)|$.

We will also make use of the following lemma (proved later).

[14] By $A \lesssim_R B$ we mean that there exists a constant $C = C(R)$ that depends only on R s.t. $A \leq C \cdot B$.
[15] Covering numbers are formally defined as the number of balls needed to cover the underlying space.

Lemma 7.2. *Let $Z_i \overset{ind}{\sim} N(\theta_i, 1)$, for $i = 1, \dots, n$, where we assume that $|\theta_i| \le R$ for all i. For $M \ge \sqrt{8 \log n} \ge 2R$, and $0 < \lambda \le 1$, we have, for any $a \in \mathbb{R}$,*

$$\mathbb{E}\left[\left(\prod_{i=1}^{n} |aZ_i|^{\mathbf{1}\{|Z_i|>M\}}\right)^{\lambda}\right] \le \exp\left\{\frac{|a|^{\lambda} 4M^{\lambda-1}}{\sqrt{2\pi}}\right\}.$$

For notational simplicity, let us write $f_* := f_{G_n}$. We will establish a (finite sample) large deviation inequality of the form: for all $t \ge 1$ and

$$\gamma_n^2 := C\frac{(\log n)^2}{n},$$

for some constant $C > 0$ to be chosen later, for all $n \ge n^*$,

$$\mathbb{P}\left[\mathfrak{h}(f_{\widehat{G}_n}, f_*) \ge t\gamma_n\right] \le 3n^{-t^2}. \tag{7.13}$$

Then, as for any nonnegative r.v. $X \sim F_X$, $\mathbb{E}[X] = \int_0^\infty (1 - F_X(x))\, dx$, we have

$$\begin{aligned} &\frac{1}{\gamma_n^2}\mathbb{E}\left[\mathfrak{h}^2(f_{\widehat{G}_n}, f_*)\right] \le 1 + 3\int_1^\infty n^{-u} du \le 1 + \frac{3}{n\log n} \le 4 \qquad (\text{for } n \ge 3) \\ \Rightarrow \quad & \mathbb{E}\left[\mathfrak{h}^2(f_{\widehat{G}_n}, f_*)\right] \le 4\gamma_n^2. \end{aligned}$$

Therefore, it suffices to just prove Eq. 7.13.

Note that,

$$\mathbb{P}[A_n] := \mathbb{P}\left[\mathfrak{h}(f_{\widehat{G}_n}, f_*) \ge t\gamma_n\right] = \mathbb{P}\left[\mathfrak{h}(f_{\widehat{G}_n}, f_*) \ge t\gamma_n, \prod_{i=1}^{n} \frac{f_{\widehat{G}_n}(Z_i)}{f_*(Z_i)} \ge 1\right]. \tag{7.14}$$

Our strategy is as follows. Let $M := \sqrt{8\log n}$. We shall work with the set $[-B, B]$, where $B := R + M$, and the pseudometric given by the pseudonorm $\|\cdot\|_B^\infty$.

Consider the following class of marginal densities:

$$\mathcal{F}(t\gamma_n) := \{f_G : \mathfrak{h}(f_G, f_*) \ge t\gamma_n\} \subset \mathcal{F}. \tag{7.15}$$

This is essentially the class of mixture densities in Eq. 7.12 subject to the additional constraint that their Hellinger distance to f_* is sufficiently large. For $\epsilon > 0$, let $\mathcal{S} := \{f_j : j \in \mathcal{J}\} \subset \mathcal{F}$ (here $\mathcal{J} = \{1, \dots, J\}$, $J = \#\mathcal{S}$) be a *proper* $(\|\cdot\|_B^\infty, \epsilon)$-cover of $\mathcal{F}(t\gamma_n)$[16], i.e.,

$$\sup_{f \in \mathcal{F}(t\gamma_n)} \inf_{1 \le j \le J} \|f - f_j\|_\infty^B \le \epsilon.$$

[16]By "proper cover", we mean that the centers of the cover are themselves elements of Eq. 7.15.

Hence, on the event A_n Eq. 7.14, there must exist $\hat{j} \in \mathcal{J}$ such that $\|f_{\hat{j}} - f_{\widehat{G}_n}\|_\infty^B \leq \epsilon$. This further implies that on A_n:

$$f_{\widehat{G}_n}(z) \leq f_{\hat{j}}(z) + \epsilon \leq \max_{j \in \mathcal{J}} f_j(z) + \epsilon \qquad \text{for all } z \in [-B, B].$$

We introduce the following function $v \equiv v_B : \mathbb{R} \to (0, \infty)$ via:

$$v(z) := \epsilon\, \mathbf{1}\{|z| \leq B\} + \epsilon\, \frac{B^2}{z^2} \mathbf{1}\{|z| > B\}, \text{ for } z \in \mathbb{R}.$$

Notice that by construction

$$\int_{-\infty}^{\infty} v(z) dz = 2\epsilon B + 2\epsilon \frac{B^2}{B} = 4\epsilon B, \tag{7.16}$$

and also

$$f_{\widehat{G}_n}(z) \leq \max_{j \in \mathcal{J}} \{f_j(z) + v(z)\} \quad \text{if } z \in [-B, B], \qquad \text{and} \qquad f_{\widehat{G}_n}(z) \leq (2\pi)^{-1/2} \quad \text{otherwise.}$$

For any f_G, write:

$$L_n(f_G, f_*) := \prod_{i=1}^n \frac{f_G(Z_i)}{f_*(Z_i)}.$$

Since $f_{\widehat{G}_n}$ is the NPMLE, it must hold that $L_n(f_{\widehat{G}_n}, f_*) \geq 1$. Next, on the event A_n:

$$\begin{aligned} L_n(f_{\widehat{G}_n}, f_*) &\leq \prod_{i:|Z_i| \leq B} \frac{f_{\hat{j}}(Z_i) + v(Z_i)}{f_*(Z_i)} \prod_{i:|Z_i| > B} \frac{(2\pi)^{-1/2}}{f_*(Z_i)} \\ &= \prod_{i=1}^n \frac{f_{\hat{j}}(Z_i) + v(Z_i)}{f_*(Z_i)} \prod_{i:|Z_i| > B} \frac{(2\pi)^{-1/2}}{f_{\hat{j}}(Z_i) + v(Z_i)} \\ &\leq \prod_{i=1}^n \frac{f_{\hat{j}}(Z_i) + v(Z_i)}{f_*(Z_i)} \prod_{i:|Z_i| > B} \frac{(2\pi)^{-1/2}}{v(Z_i)}. \end{aligned}$$

Observe that, $\mathbb{P}\left[\mathfrak{h}(f_{\widehat{G}_n}, f_*) \geq t\gamma_n\right]$ now equals

$$\begin{aligned} &\mathbb{P}\left[\mathfrak{h}(f_{\widehat{G}_n}, f_*) \geq t\gamma_n,\ L_n(f_{\widehat{G}_n}, f_*) \geq 1\right] \\ &\leq \mathbb{P}\left[\prod_{i=1}^n \frac{f_{\hat{j}}(Z_i) + v(Z_i)}{f_*(Z_i)} \cdot \left(\prod_{i:|Z_i| > B} \frac{(2\pi)^{-1/2}}{v(Z_i)}\right) \geq 1\right] \\ &\leq \mathbb{P}\left[\prod_{i=1}^n \frac{f_{\hat{j}}(Z_i) + v(Z_i)}{f_*(Z_i)} \geq e^{-2\gamma}\right] + \mathbb{P}\left[\prod_{i:|Z_i| > B} \frac{(2\pi)^{-1/2}}{v(Z_i)} \geq e^{2\gamma}\right], \end{aligned} \tag{7.17}$$

for some γ (to be chosen later). Let us first bound the first term in the above display. Observe that,

$$
\begin{aligned}
&\mathbb{P}\left[\prod_{i=1}^{n}\frac{f_{\hat{j}}(Z_i)+v(Z_i)}{f_*(Z_i)}\geq e^{-2\gamma}\right]\\
&\leq J\max_{j\in\mathcal{J}}\mathbb{P}\left[\prod_{i=1}^{n}\frac{f_j(Z_i)+v(Z_i)}{f_*(Z_i)}\geq e^{-2\gamma}\right]\\
&= J\sup_{j\in\mathcal{J}}\mathbb{P}\left[\prod_{i=1}^{n}\sqrt{\frac{f_j(Z_i)+v(Z_i)}{f_*(Z_i)}}\geq e^{-\gamma}\right]\\
&\leq J\sup_{j\in\mathcal{J}}\left\{e^{\gamma}\prod_{i=1}^{n}\mathbb{E}\left[\sqrt{\frac{f_j(Z_i)+v(Z_i)}{f_*(Z_i)}}\right]\right\} \qquad \text{(by Markov's inequality)} \qquad (7.18)\\
&\leq Je^{\gamma}\sup_{j\in\mathcal{J}}\left\{\exp\left(\sum_{i=1}^{n}\left\{\mathbb{E}\left[\sqrt{\frac{f_j(Z_i)+v(Z_i)}{f_*(Z_i)}}\right]-1\right\}\right)\right\} \qquad (\text{as } x\leq e^{x-1})\\
&\overset{(*)}{=} Je^{\gamma}\sup_{j\in\mathcal{J}}\left\{\exp\left(n\left\{\int_{-\infty}^{\infty}\sqrt{f_j(z)+v(z)}\sqrt{f_*(z)}dz-1\right\}\right)\right\}\\
&\overset{(**)}{\leq} Je^{\gamma}\exp\left(-n\mathfrak{h}^2(f_j,f_*)+2n\sqrt{\epsilon B}\right)\\
&\leq \exp\left\{-nt^2\gamma_n^2+2n\sqrt{\epsilon B}+\gamma+\log J\right\}.
\end{aligned}
$$

In $(*)$ we have used the following argument:

$$
\begin{aligned}
\sum_{i=1}^{n}\mathbb{E}\left[\sqrt{\frac{f_j(Z_i)+v(Z_i)}{f_*(Z_i)}}\right] &= \sum_{i=1}^{n}\int_{-\infty}^{\infty}\sqrt{\frac{f_j(z)+v(z)}{f_*(z)}}\,\phi(z-\theta_i)\,dz\\
&= n\int\sqrt{\frac{f_j(z)+v(z)}{f_*(z)}}\left(\frac{1}{n}\sum_{i=1}^{n}\phi(z-\theta_i)\right)dz\\
&= n\int_{0}^{\infty}\sqrt{f_j(z)+v(z)}\sqrt{f_*(z)}\,dz.
\end{aligned}
$$

In $(**)$ we used the following argument (and the fact that $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ and the Cauchy-Schwarz inequality and Eq. 7.16):

$$
\begin{aligned}
\int\sqrt{f_j(z)+v(z)}\sqrt{f_*(z)}dz-1 &\leq \int\left(\sqrt{f_j(z)f_*(z)}+\sqrt{v(z)f_*(z)}\right)dz-1\\
&\leq -\mathfrak{h}^2(f_j,f_*)+\left(\int v(z)dz\right)^{1/2}\left(\int f_*(z)dz\right)^{1/2}\\
&= -\mathfrak{h}^2(f_j,f_*)+2\sqrt{\epsilon B}.
\end{aligned}
$$

Let us now bound the second term in Eq. 7.17. By Markov's inequality, we have

$$\mathbb{P}\left[\prod_{i:|Z_i|>B}\frac{(2\pi)^{-1/2}}{v(Z_i)}\geq e^{2\gamma}\right]\leq\exp\left(-\frac{2\gamma}{C\log n}\right)\mathbb{E}\left[\left(\prod_{i:|Z_i|>B}\frac{(2\pi)^{-1/2}}{v(Z_i)}\right)^{\frac{1}{C\log n}}\right]. \tag{7.19}$$

For $z\notin[-B,B]$, as $1/v(z)=z^2/(\epsilon B^2)$, we can bound the expectation above, using Lemma 7.2 with $\lambda=\frac{2}{C\log n}$ and $a=(2\pi)^{-1/4}(\sqrt{\epsilon}B)^{-1}$, as

$$\begin{aligned}\mathbb{E}\left[\left(\prod_{i:|Z_i|>B}\frac{(2\pi)^{-1/2}}{v(Z_i)}\right)^{\frac{1}{C\log n}}\right] &\leq\mathbb{E}\left[\left(\prod_{i:|Z_i|>B}\frac{(2\pi)^{-1/4}|Z_i|}{\sqrt{\epsilon}B}\right)^{\frac{2}{C\log n}}\right]\\ \leq\ \exp\left\{\frac{4\left((2\pi)^{1/4}\sqrt{\epsilon}\right)^{-\frac{2}{C\log n}}B^{-1}}{\sqrt{2\pi}}\right\}.\end{aligned} \tag{7.20}$$

We are now ready to pick all parameters. We pick $\epsilon=1/n^2$ and $\gamma:=nt^2\gamma_n^2/2$.

Then, by Lemma Lemma 7.1, we get that:

$$\log J\leq C_R(\log n)^2.$$

Thus, the first term P_1 in Eq. 7.17 can be upper bounded, for all $t\geq 1$, by (using Eq. 7.18):

$$\begin{aligned}P_1 &\leq\ \exp\left\{-nt^2\gamma_n^2+2\sqrt{B}+nt^2\gamma_n^2/2+C_R(\log n)^2\right\}\\ &\leq\ \exp\left\{-Ct^2(\log n)^2/2+C'(\log n)^2t^2+C_R(\log n)^2t^2\right\}=e^{-t^2(\log n)^2}\end{aligned} \tag{7.21}$$

where we have used the facts: (i) $2\sqrt{B}\leq C'(\log n)^2t^2$, (ii) $C_R(\log n)^2\leq C_R(\log n)^2t^2$, and (iii) C is chosen such that $-C/2+C'+C_R=-1$.

Now, the second term P_2 in Eq. 7.17 can be upper bounded, for all $t\geq 1$, by (using Eq. 7.19 and Eq. 7.20) as

$$\begin{aligned}P_2 &\leq\ \exp\left(-t^2\log n\right)\exp\left\{\frac{4\left((2\pi)^{1/4}\sqrt{\epsilon}\right)^{-\frac{2}{C\log n}}B^{-1}}{\sqrt{2\pi}}\right\}\\ &\leq\ \exp\left(-t^2\log n+\frac{4\,(2\pi)^{\frac{1}{2C\log n}}\,e^{-2/C}}{\sqrt{2\pi}\sqrt{8\log n}}\right)\leq 2e^{-t^2\log n}\end{aligned} \tag{7.22}$$

for n large (such that $\frac{4\,e^{-2/C}\,(2\pi)^{\frac{1}{2C\log n}}}{\sqrt{2\pi}\sqrt{8\log n}}\leq\log 2$), where we have used the facts that: (i) $n^{\frac{2}{C\log n}}=e^{2/C}$, (ii) $B\geq\sqrt{8\log n}$. Then, combining Eq. 7.21 and Eq. 7.22, for n sufficiently large,

$$\mathbb{P}\left[A_n\right]\leq e^{-t^2(\log n)^2}+2e^{-t^2\log n}\leq 3e^{-t^2\log n}=3n^{-t^2}.$$

This proves the desired result.

□

Proof of Lemma 7.2. Observe that

$$\begin{aligned}\mathbb{E}\left[\left(\prod_{i=1}^{n}|aZ_i|^{\mathbf{1}\{|Z_i|>M\}}\right)^{\lambda}\right] &= \prod_{i=1}^{n}\mathbb{E}\left[|aZ_i|^{\lambda\mathbf{1}\{|Z_i|>M\}}\right] \\ &\leq \prod_{i=1}^{n}\left(1+|a|^{\lambda}\mathbb{E}\left[|Z_i|^{\lambda}\mathbf{1}\{|Z_i|>M\}\right]\right) \\ &\leq \exp\left\{|a|^{\lambda}n\int_{|z|>M}|z|^{\lambda}f_{G_n}(z)\,dz\right\},\end{aligned}$$

where in the last inequality we have used the facts: (i) $1+x \leq e^x$, (ii) for any $h(\cdot)$, $\sum_{i=1}^{n}\mathbb{E}\left[h(Z_i)\right] = n\int h(z)f_{G_n}(z)\,dz$. Let $Z \sim N(0,1)$ be independent of $\xi \sim G_n$. Since $Z+\xi \sim f_{G_n}$, and $\lambda \leq 1$,

$$\begin{aligned}\int_{|z|>M}|z|^{\lambda}f_{G_n}(z)\,dz &= \mathbb{E}\left[|Z+\xi|^{\lambda}\mathbf{1}\{|Z+\xi|>M\}\right] \\ &\leq \mathbb{E}\left[|2Z|^{\lambda}\mathbf{1}\{|Z|>\frac{M}{2}\}\right]+\mathbb{E}\left[|2\xi|^{\lambda}\mathbf{1}\{|\xi|>\frac{M}{2}\}\right] \\ &\leq 2M^{\lambda-1}\mathbb{E}\left[|Z|\mathbf{1}\{|Z|>\frac{M}{2}\}\right] \\ &\leq 4M^{\lambda-1}\int_{M/2}^{\infty}z\phi(z)\,dz = 4M^{\lambda-1}\frac{e^{-M^2/8}}{\sqrt{2\pi}} \leq \frac{4M^{\lambda-1}}{n\sqrt{2\pi}}\end{aligned}$$

where in the second inequality follows from the facts: (i) $|2Z|^{\lambda} \leq 2|Z|M^{\lambda-1}$ as $\lambda \leq 1$ and $|Z| > \frac{M}{2}$, and (ii) that $\xi \leq \frac{M}{2}$ a.s.~(as $|\theta_i| \leq R$, for all i). As $M \geq \sqrt{8\log n}$, we have $e^{-M^2/8} \leq \frac{1}{n}$.

□

7.4.2 Consistency of $\widehat{G}_n$

In this section we assume that Eq. 7.1 holds. In the following result we show that for a location mixture model with a symmetric kernel, the NPMLE $\widehat{G}_n$ converges weakly to the truth G a.s. The proof crucially uses the first-order characterization of the NPMLE and mimics the approach of Groeneboom and Wellner (1992) (Section 4.2) and Jewell (1982). Similar results can be derived for other nonparametric mixture models (e.g., for scale mixtures) by appropriately adapting some of the steps in the proof.

Theorem 7.5. *Suppose that $p(\cdot \mid \theta) \equiv p(\cdot - \theta)$ be a location family such that: (i) $p(\cdot)$ is symmetric about the origin (i.e., $p(z) = p(-z)$ for all $z \in \mathbb{R}$), (ii) $0 < p(z) \leq p_{\max} < \infty$ for*

all $z \in \mathbb{R}$ and $p(\cdot)$ is uniformly continuous on $\mathbb{R}$, (iii) $p(\cdot)$ is decreasing on $[0, \infty)$, and (iv) $f_G = f_{G'}$ implies that $G = G'$. Then, w.p. 1,

$$\widehat{G}_n \xrightarrow{d} G, \qquad \text{as } n \to \infty.$$

Proof. Let F_n denote the empirical distribution of the observed $Z_1, \dots, Z_n$, i.e., $F_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$. Let $F_\star$ be the c.d.f. of the marginal distribution of the Z_i's, i.e., $F_\star$ has density $f_G \equiv f_\star$ (see Eq. 7.2). We will show that there exists a set B with probability 1 such that for each $\omega \in B$, given any subsequence of $\{\widehat{G}_n(\cdot;\omega)\}_{n\geq 1}$ there exists a further subsequence which converges weakly to G; this will then complete the proof (see e.g., Theorem 3.2.9 in Durrett (2010)).

Recall that the first order optimality of $\widehat{G}_n$ implies that

$$\sum_{j=1}^m n_j \left\{ \frac{f_\star(y_j)}{f_{\widehat{G}_n}(y_j)} - 1 \right\} \leq 0 \quad \Leftrightarrow \quad \int \frac{f_\star(z)}{f_{\widehat{G}_n}(z)}\, dF_n(z) \leq 1. \tag{7.23}$$

For ϵ, define the compact set

$$A_\epsilon := \{z \in \mathbb{R} : f_\star(z) \geq \epsilon\} \cap [-\frac{1}{\epsilon}, \frac{1}{\epsilon}], \qquad \text{s.t.} \qquad F_\star(A_\epsilon) > 0. \tag{7.24}$$

By the Glivenko-Cantelli theorem, we have,

$$\mathbb{P}\left[B\right] = 1, \quad \text{where } B := \left\{\omega : \|F_n(\cdot;\omega) - F_\star\|_\infty := \sup_{z\in\mathbb{R}} |F_n(z;\omega) - F_\star(z)| \to 0, \text{ as } n \to \infty\right\}.$$

Fix $\epsilon := 1/j$ for some j. Let $\omega \in B$. Then,[17]

$$F_n(A_\epsilon) \to F_\star(A_\epsilon) > 0 \quad \text{as } n \to \infty. \tag{7.25}$$

By Helly's selection theorem (see Theorem 3.2.6 of Durrett (2010)), the sequence of distribution functions $\{\widehat{G}_{n_k}(\cdot;\omega)\}_{k\geq 1}$ has a subsequence $\{\widehat{G}_{n_{k_l}}(\cdot;\omega)\}_{l\geq 1}$ converging *vaguely*[18] to a sub-distribution function $\tilde{G}$ (say). We have to show that $\tilde{G} = G$.

Lemma 7.3. *We have, for $\omega \in B$,*

$$\lim_{l\to\infty} \int_{A_\epsilon} \frac{f_\star(z)}{f_{\widehat{G}_{n_{k_l}}}(z;\omega)}\, dF_{n_{k_l}}(\cdot;\omega) = \int_{A_\epsilon} \frac{f_\star(z)}{f_{\tilde{G}}(z)}\, dF_\star(z) \leq 1. \tag{7.26}$$

[17]This is because of the continuous mapping theorem (see e.g., Theorem 3.2.4 in Durrett (2010)): Note that $\psi(z) := \mathbf{1}_{A_\epsilon}(z)$, for $z \in \mathbb{R}$, is a measurable and bounded function such that $D_\psi := \{z \in \mathbb{R} : \psi \text{ is discontinuous at } z\}$ as Lebesgue measure 0. Thus, for $Z \sim F_\star$, $\mathbb{P}\left[Z \in D_\psi\right] = 0$, and then $F_n(A_\epsilon) = \int \psi\, dF_n \to \int \psi\, dF = F(A_\epsilon)$.

[18]See Chapter 4 in Chung (1974) for the formal definition and study of vague convergence.

Now, by monotone convergence theorem Eq. 7.26, where to take $\epsilon = 1/j \to 0$ as $j \to \infty$ (so that $A_{1/j}$ increases to $\mathbb{R}$) we have $\int \frac{f_\star(z)}{f_{\tilde{G}}(z)}\, dF_\star(z) \le 1$, i.e., $\int \frac{f_G^2(z)}{f_{\tilde{G}}(z)}\, dz \le 1$.

Lemma 7.4. $\int \frac{f_G^2(z)}{f_{\tilde{G}}(z)}\, dz \le 1$ *implies that* $f_{\tilde{G}} = f_G$.[19]

As $f_\star \equiv f_G$, under assumption (iv) of identifiability in the theorem, we immediately have $\tilde{G} = G$.

We can thus conclude from this that every subsequence of the sequence $\{\widehat{G}_n(\cdot;\omega)\}_{n\ge1}$ has a convergent subsequence, and that all these subsequences have the same weak limit G. This implies the consistency of the NPMLE $\widehat{G}_n$.

□

Proof of Lemma 7.3. For notational simplicity let us rename $\{\widehat{G}_{n_{k_l}}(\cdot)\}_{l\ge1}$ to $\{\widehat{G}_n(\cdot)\}_{n\ge1}$ and assume that $\widehat{G}_n \xrightarrow{v} \tilde{G}$ as $n \to \infty$, where $\xrightarrow{v}$ denotes vague convergence. Further, we fix $\omega \in B$ and hide the dependence on the random quantities on ω. Note that for every fixed $z \in \mathbb{R}$, the function $p(z - \cdot)$ is continuous and satisfies $\lim_{|y|\to\infty} p(z-y) = 0$. Thus, from the definition of vague convergence (see e.g., Theorem 4.4.1 in Chung (1974)), we have

$$\lim_{n\to\infty} f_{\widehat{G}_n}(z) = \lim_{n\to\infty} \int p(z-\theta)\, d\widehat{G}_n(\theta) \to \int p(z-\theta)\, d\tilde{G}(\theta) = f_{\tilde{G}}(z). \tag{7.27}$$

As $p(\cdot)$ is uniformly continuous on $\mathbb{R}$, so is $f_{\widehat{G}_n}$; in fact, the sequence $\{f_{\widehat{G}_n}\}_{n\ge1}$ is uniformly equicontinuous on $\mathbb{R}$.[20] Further, as $p(\cdot)$ is upper bounded, $\{f_{\widehat{G}_n}\}_{n\ge1}$ is pointwise bounded. Now, by the Arzela-Ascoli theorem (see e.g.,Theorem 4.44 in Folland (1999)), $\{f_{\widehat{G}_n}\}_{n\ge1}$ converges uniformly on A_ϵ to $f_{\tilde{G}}$.[21]

[19] This lemma immediately follows as $\int \frac{f_G^2(z)}{f_{\tilde{G}}(z)}\, dz - 1 = \int \left(\frac{f_G(z)}{f_{\tilde{G}}(z)} - 1\right)^2 f_{\tilde{G}}(z)\, dz \ge 0$ with equality iff $f_G = f_{\tilde{G}}$ a.s. As both f_G and $f_{\tilde{G}}$ are continuous, this implies we must have $f_G = f_{\tilde{G}}$.

[20] As $p(\cdot)$ is uniformly continuous on $\mathbb{R}$, given any $\eta > 0$, there exists $\delta > 0$ such that whenever $y_1, y_2 \in \mathbb{R}$ and $|y_1 - y_2| < \delta$, we have $|p(y_1) - p(y_2)| < \eta$. Then,

$$|f_{\widehat{G}_n}(y_1) - f_{\widehat{G}_n}(y_2)| \le \int |p(y_1-\theta) - p(y_2-\theta)|\, d\widehat{G}_n(\theta) < \eta,$$

as $|(y_2-\theta)-(y_1-\theta)| = |y_1 - y_2| < \delta$. Hence, $f_{\widehat{G}_n}(\cdot)$ is also uniformly continuous on $\mathbb{R}$. Further, as δ only depends on η, but not on y_1, y_2 or n, the sequence $\{f_{\widehat{G}_n}\}_{n\ge1}$ is uniformly equicontinuous on $\mathbb{R}$.

[21] Note that Arzela-Ascoli says that $\{f_{\widehat{G}_n}\}_{n\ge1}$ has a subsequence that converges uniformly on A_ϵ to a continuous limit function. However, as $\{f_{\widehat{G}_n}\}_{n\ge1}$ converges pointwise to $f_{\tilde{G}}$, the entire sequence $\{f_{\widehat{G}_n}\}_{n\ge1}$ converges uniformly to the limit $f_{\tilde{G}}$.

We now claim that there exists an interval $[-a, a] \subset \mathbb{R}$ such that $A_\epsilon \subset [-a, a]$ and $\widehat{G}_n([-a, a]) \geq \delta > 0$, for all large n, for some $\delta > 0$.[22]

Thus, for all large n, we have, for any $z \in A_\epsilon$, using the unimodality of $p(\cdot)$,

$$\begin{aligned} f_{\widehat{G}_n}(z) \geq \int_{[-a,a]} p(z-\theta)\, d\widehat{G}_n(\theta) &\geq \min\{p(z-a), p(z+a)\} \int_{[-a,a]} d\widehat{G}_n(\theta) \\ &\geq \delta \min\{p(z-a), p(z+a)\} \\ &\geq \delta\, p(2a) =: \delta_a > 0. \end{aligned}$$

Thus, $f_{\widehat{G}_n}$ is strictly positive on A_ϵ. As $f_{\widehat{G}_n}$ is uniformly continuous and nonzero on the compact set A_ϵ then $1/f_{\widehat{G}_n}$ is also uniformly continuous on A_ϵ.

Let $h_n(z) := \frac{f_\star(z)}{f_{\widehat{G}_n}(z)}$, for $z \in \mathbb{R}$. Then, for n large (say $n \geq n_0$),

$$h_n(z) := \frac{f_\star(z)}{f_{\widehat{G}_n}(z)} \leq \frac{p_{\max}}{\delta_a}, \qquad \text{for} \quad z \in A_\epsilon. \tag{7.28}$$

Note that by Eq. 7.27, for $z \in A_\epsilon$ (and as $f_{\tilde{G}}(z) = \lim_{n\to\infty} f_{\widehat{G}_n}(z) \geq \delta_a$),

$$h_n(z) \to h(z) := \frac{f_\star(z)}{f_{\tilde{G}}(z)} \quad \text{as} \quad n \to \infty.$$

Further, h_n is a continuous function on A_ϵ; moreover, the sequence $\{h_n\}_{n\geq n_0}$ is uniformly equicontinuous[23] on A_ϵ and pointwise bounded (by Eq. 7.28). Thus, again, by the Arzela-Ascoli theorem $\{h_n\}_{n\geq n_0}$ converges uniformly on the compact set A_ϵ to h.

[22]To see this, note that for any interval $[-a, a] \supseteq A_\epsilon$, and $z \in A_\epsilon$,

$$\begin{aligned} f_{\widehat{G}_n}(z) = \int_{[-a,a]} p(z-\theta)\, d\widehat{G}_n(\theta) + \int_{[-a,a]^c} p(z-\theta)\, d\widehat{G}_n(\theta) \\ \leq p_{\max}\, \widehat{G}_n([-a, a]) + p(a-\epsilon^{-1}) =: \gamma_{a,n} \end{aligned}$$

where we have used the facts: (i) $p(\cdot)$ is upper bounded by $p_{\max}$, and (ii) for $z \in A_\epsilon \subseteq [-\epsilon^{-1}, \epsilon^{-1}]$ and $\theta \in [-a, a]^c$, $|z-\theta| \geq a - \epsilon^{-1}$ which implies that $p(z-\theta) \leq p(a-\epsilon^{-1})$ (as $p(\cdot)$ is decreasing on $[0, \infty)$). Therefore, using the facts that: (a) for $z \in A_\epsilon$, $f_\star(z) \geq \epsilon$, and (b) $F_n(A_\epsilon) \geq F_\star(A_\epsilon)/2$ for all n sufficiently large, we have, for all n sufficiently large

$$1 \geq \int_{A_\epsilon} \frac{f_\star(z)}{f_{\widehat{G}_n}(z)}\, dF_n(z) \geq \frac{\epsilon}{\gamma_{a,n}} F_n(A_\epsilon) \geq \frac{\epsilon}{2\gamma_{a,n}} F_\star(A_\epsilon),$$

which yields

$$\gamma_{a,n} \geq \frac{\epsilon}{2} F_\star(A_\epsilon) \quad \Leftrightarrow \quad \widehat{G}_n([-a, a]) \geq p_{\max}^{-1} \left\{ \frac{\epsilon}{2} - p(a-\epsilon^{-1}) \right\}.$$

But as \$a →∞\$, $p(a-\epsilon^{-1}) \to 0$, and thus, there exists $a \in \mathbb{R}$ such that $\widehat{G}_n([-a, a]) \geq \delta > 0$ for some $\delta > 0$.

[23]Show this.

$$
\begin{aligned}
&\left| \int_{A_\epsilon} h_n(z)\, dF_n(z) - \int_{A_\epsilon} h(z)\, dF_\star(z) \right| \\
&\leq \left| \int_{A_\epsilon} h_n(z)\, dF_n(z) - \int_{A_\epsilon} h(z)\, dF_n(z) \right| + \left| \int_{A_\epsilon} h(z)\, dF_n(z) - \int_{A_\epsilon} h(z)\, dF_\star(z) \right| \\
&\leq \sup_{z \in A_\epsilon} |h_n(z) - h(z)| + o(1) = o(1) \quad \text{as} \quad n \to \infty,
\end{aligned}
$$

where the first term in the above display converges to zero as $h_n(\cdot)$ converges uniformly to $h(\cdot)$ on A_ϵ and the second term converges to 0 by an application of the continuous mapping theorem (see Theorem 3.2.4 in Durrett (2010)) as in the proof of Eq. 7.25.

□

8 Multiple Testing and empirical Bayes

We now turn to study one of the most important and beautiful areas of statisics: multiple testing.

There are three views of multiple testing that we will describe here.

- Multiple testing as a **burden** (Section 8.2): the more hypotheses we are testing, the more false discoveries we are bound to make.
- Multiple testing as an **opportunity** to draw inferences that were not possible in classical statistics: an empirical Bayesian has good reasons to be excited about multiple testing!
- A middle-of-the-road approach: modern multiple testing is largely about recognizing both the possible burdens and the opportunities.

Before taking on the multiple testing problem, however, we provide a bird's overview of single hypothesis testing (and we refer the reader to, e.g., Erich L. Lehmann and Romano (2005) for a comprehensive textbook reference).

8.1 Single hypothesis testing recap

Consider a simple statistical decision problem with parameter $\theta \in \Theta$ where we observe:

$$Z \sim p(\cdot \mid \theta). \tag{8.1}$$

In hypothesis testing, we partition, Θ into a disjoint union,

$$\Theta = \Theta_0 \sqcup \Theta_1,$$

and then seek to decide between the two options:

$$H_0 : \theta \in \Theta_0 \text{ vs. } H_1 : \theta \in \Theta_1,$$

where we call H_0 the null hypothesis and H_1 the alternative. We consider decision rules $t(z) \in \mathcal{T} = \{0, 1\}$ where $t(z) = 1$ implies that we choose H_1; in statistical parlance we then reject the null hypothesis.

There are two possible errors that $t(z)$ can make. In hypothesis testing, one typically is more interested in preventing type-I errors, that is, to declare $t(z) = 1$ while H_0 is actually true.

Hence we typically only search for decision rules that satisfy the following guarantee for a pre-specified $\alpha \in (0,1)$

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta \left[t(Z) = 1 \right] \leq \alpha. \tag{8.2}$$

This guarantee is easily satisfied in the special case wherein $\Theta_0 = \{\theta_0\}$ is a singleton. In that case, we only require $\mathbb{P}_{\theta_0} \left[t(Z) = 1 \right] \leq \alpha$, i.e., we have a constraint with respect to a single and known probability measure. For simplicity in the remainder of this chapter we make the assumption that $\Theta_0 = \{\theta_0\}$ is indeed a singleton.

A large chunk of the multiple testing literature provides type-I error control guarantees when the null distribution of Z is known, but remains agnostic about its possible distribution under alternatives. This idea is typically captured through the notion of a p-value.

Definition 8.1 (p-value for a simple hypothesis)**.** Suppose we are testing a null hypothesis $H_0 : \theta = \theta_0$. A random variable $P \in [0,1]$ is called a **p-value** for the hypothesis $H_0 : \theta = \theta_0$ if it is uniformly distributed under θ_0, that is,

$$\mathbb{P}_{\theta_0} \left[P \leq t \right] = t \ \text{ for all } \ t \in [0,1].$$

It is also typically assumed that for $\theta \neq \theta_0$, the p-value will be stochastically smaller than uniform and that small realized values of the p-value provide evidence *against* H_0.

This is clearly not the most general definition of a p-value. This definition however suffices for our brief recap here and captures the following important notions:

- Typically P will be measurable function of Z in Eq. 8.1, i.e., $P = P(Z)$. As we know the distribution of Z under the null hypothesis by transformation we may set it to be uniform (at least when $Z \in \mathbb{R}$ and the distribution of Z is absolutely continuous).
- It provides an interpretable scale with which to base decisions. To get guarantee Eq. 8.2 at a fixed α, one can use the decision rule $t(Z) \equiv t(P) = \mathbf{1}(P \leq \alpha)$.

Definition 8.1 also helps with potential misunderstanding of what a p-value is. It makes it very clear that, e.g., under hypothetical replications of an experiment with no signal, one would expect to get p-values ≤ 0.05, 5% of the time.

Example 8.1. Suppose $Z \sim \mathcal{N}(\theta, 1)$ and we seek to test whether $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$. A p-value is then given by:

$$P = 2(1 - \Phi(|Z|)),$$

where Φ is the standard normal distribution function.

8.2 Multiple testing as a burden

We are ready to move on to the multiple testing setting. Here we are faced with n hypothesis tests

$$H_{i,0} : \theta_i = \theta_{i,0} \text{ vs } H_{i,1} : \theta_i \neq \theta_{i,0}$$

based on data $\mathbf{Z} = (Z_1, \dots, Z_n)$, with

$$Z_i \sim p_i(\cdot \mid \theta_i).$$

We write $\mathcal{H}_0 \subset \{1, \dots, n\}$ for the subset of hypotheses such that $H_{0,i}$ holds. We also write $n_0 := \#\mathcal{H}_0$ for the total number of null hypotheses.

A multiple testing procedure consists of a decision rule $\mathbf{t}(\mathbf{Z}) = (t_1(\mathbf{Z}), \dots, t_n(\mathbf{Z})) \in \{0,1\}^n$ with the interpretation that $t_i(\mathbf{Z}) = 1$ implies rejection of the null hypothesis $H_{0,i}$. This will be called a false discovery if $H_{0,i}$ is true. Hence the total number of discoveries is given by:

$$R = R(\mathbf{t}) = \sum_{i=1}^{n} \mathbf{1}(t_i(\mathbf{Z}) = 1), \tag{8.3}$$

and the total number of *false* discoveries is given by:

$$V = V(\mathbf{t}) = \sum_{i \in \mathcal{H}_0} \mathbf{1}(t_i(\mathbf{Z}) = 1). \tag{8.4}$$

Often we will first replace the Z_i by a p-value P_i, in which case we write $\mathbf{P} = (P_1, \dots, P_n)$. In that case we write $\mathbf{t}(\mathbf{P})$ multiple testing decision.

The fundamental challenge of multiple testing is captured by the following simple computation. If we proceed as in single hypothesis testing and take $t_i(\mathbf{P}) = \mathbf{1}(P_i \leq \alpha)$ for a fixed α, say $\alpha = 0.05$, then we will potentially incur a lot of false (spurious) discoveries. To see this, note that the above decision rule satisfies the following:

$$\mathbb{E}[V] = \sum_{i \in \mathcal{H}_0} \mathbb{E}[\mathbf{1}(P_i \leq \alpha)] = n_0 \alpha,$$

i.e., the expected number of false discoveries is equal to $n_0\alpha$. Furthermore, if the p-values $P_1, \dots, P_n$ are mutually independent, then we will make a false discovery with high probability:

$$\mathbb{P}[V \geq 1] = 1 - \mathbb{P}[V = 0] = 1 - \mathbb{P}[P_i > \alpha \text{ for all } i \in \mathcal{H}_0] = 1 - (1-\alpha)^{n_0},$$

and so for n_0 large enough, we are virtually guaranteed to make at least one false discovery.

A traditional goal in multiple testing is to avoid making any spurious discoveries with high probability, as formalized in the following definition.

Definition 8.2 (Family-Wise Error Rate)**.** The family-wise error rate (FWER) of a multiple testing procedure $\mathbf{t}$ is defined as:

$$\text{FWER} := \mathbb{P}\left[V(\mathbf{t}) \geq 1\right]$$

We say that a procedure controls the FWER at level α, if FWER $\leq \alpha$.

Perhaps the most classical procedure to achieve this goal is the famous and (very) conservative Bonferroni procedure:

Definition 8.3 (Bonferroni procedure)**.** Given n p-values $P_1, \dots, P_n$ and nominal level α, the Bonferroni procedure takes the following form:

$$\mathbf{t}(\mathbf{P}) = (\mathbf{1}(P_1 \leq \alpha/n), \dots, \mathbf{1}(P_n \leq \alpha/n)).$$

We can show that the Bonferroni procedure controls the FWER at level α.

$$\text{FWER} = \mathbb{P}\left[V \geq 1\right] \leq \mathbb{E}\left[V\right] = \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\mathbf{1}(P_i \leq \alpha/n)\right] = n_0 \frac{\alpha}{n} \leq \alpha.$$

Hence the Bonferroni procedure protects us from false discoveries, but this comes at a cost: we need to test each hypothesis at level α/n, and if n is large (say $n \approx 20,000$, as in genomics studies), then we may not be able to make any discovery at all. Hence multiple testing is often interpreted as a burden.

8.3 Multiple testing as an opportunity: empirical Bayes

We now switch gears and treat the multiple testing problem as an empirical Bayes problem.

We now model our setting in the following "typical" empirical Bayes fashion:

$$\theta_i \overset{\text{iid}}{\sim} G, \quad Z_i \mid \theta_i \sim p(\cdot \mid \theta_i), \tag{8.5}$$

and we assume for simplicity that all null hypotheses are given by $H_{0,i} : \theta_i = \theta_0$. We also introduce the (random) indicators $H_i = \mathbf{1}(\theta_i \neq \theta_0) \in \{0, 1\}$. Then we have that the set of null hypotheses $\mathcal{H}_0$ is in fact random an equal to $\{i : H_i = 0\}$ with (random) cardinality n_0. We write $\pi_0 := \mathbb{P}\left[H_i = 0\right] = \mathbb{P}_G\left[\theta_i = \theta_0\right]$.

8.3.1 The two-groups model

Eq. 8.5 and the definition of H_i imply the alternative equivalent data-generating mechanism.

$$\begin{aligned} H_i &\sim \text{Bernoulli}(1-\pi_0), \\ Z_i \mid H_i = 0 &\sim p(\cdot \mid \theta_0), \\ Z_i \mid H_i = 1 &\sim \frac{1}{1-\pi_0} \int_{\Theta \setminus \{\theta_0\}} p(\cdot \mid \theta) dG(\theta). \end{aligned} \tag{8.6}$$

The above connection is important to understand how a lot of the empirical Bayes multiple testing literature relates to empirical Bayes outside of multiple testing. Specifically, the starting point for the empirical Bayes multiple testing literature is often the following celebrated two-groups model (often Bradley Efron, Tibshirani, Storey, and Tusher (2001) for an early appearance of the two-groups model):

$$\begin{aligned} H_i &\sim \text{Bernoulli}(1-\pi_0), \\ Z_i \mid H_i = 0 &\sim F_{\text{null}}(\cdot), \\ Z_i \mid H_i = 1 &\sim F_{\text{alt}}(\cdot). \end{aligned} \tag{8.7}$$

Starting with Eq. 8.7 instead of Eq. 8.5 has the following interpretation: we only need to believe the distributional assumption of Eq. 8.5 at the null $\theta = \theta_0$, while for the alternatives we do not need to impose the form given in Eq. 8.6. Furthermore, we can forget about the latent θ_i (which may take values e.g., in $\mathbb{R}$), and instead we only need to focus on the binary latent variable $H_i \in \{0, 1\}$. Furthermore, once we turn the empirical Bayes crank, we typically assume that $F_{\text{null}}(\cdot)$ is known, and under Eq. 8.5 we would need to estimate the unknown G, while under Eq. 8.7 we would need to estimate F_{alt} and π_0.

It is important to note that under Eq. 8.7, the marginal distribution of Z_i is given by:

$$Z_i \sim F(\cdot) := \pi_0 F_{\text{null}}(\cdot) + (1-\pi_0) F_{\text{alt}}(\cdot). \tag{8.8}$$

The upshot of the two-groups model Eq. 8.7 is that it makes it easy to derive optimal multiple testing procedures. These typically take the form of rejecting hypothesis with a small value of the local false discovery rate, which is defined as:

$$\text{Lfdr}_i := \mathbb{P}\left[H_i = 0 \mid \mathbf{Z}\right]. \tag{8.9}$$

When we assume that the pairs (θ_i, Z_i) for $i \in \{1, \dots, n\}$ are mutually independent, then $\text{Lfdr}_i := \mathbb{P}\left[H_i = 0 \mid Z_i\right]$. Furthermore, if F and F_{null} have densities f, resp. f_{null}, then:

$$\text{Lfdr}_i = \frac{\pi_0 f_{\text{null}}(Z_i)}{f(Z_i)}. \tag{8.10}$$

The following optimality result is one of the cornerstones of the literature; it has appeared in various forms e.g., in Sun and Cai (2007) and Cai, Li, Maris, and Xie (2011).

Theorem 8.1 (Optimal weighted classification with local false discoveries)**.** *Let* $\mathbf{t}(\mathbf{Z})$ *be a multiple testing procedure. Suppose we evaluate it based on the following loss, where* $\mathbf{H} = (H_1, \dots, H_n)$ *and* $\lambda \in (0,1)$ *is fixed:*

$$\ell(\mathbf{t}(\mathbf{Z}), \mathbf{H}) = \sum_{i=1}^{n} \left[(1-\lambda)\mathbf{1}(t_i(\mathbf{Z}) = 1, H_i = 0) + \lambda \mathbf{1}(t_i(\mathbf{Z}) = 0, H_i = 1)\right].$$

Then the optimal decision takes the form:

$$t_i^*(\mathbf{Z}) = \mathbf{1}(\mathrm{Lfdr}_i \leq \lambda).$$

Proof. As we have argued before, it suffices to solve the following minimization problem:

$$t_i(\mathbf{Z}) \in \operatorname*{argmin}_{t_i \in \{0,1\}} \left\{(1-\lambda)\mathbb{E}\left[\lambda \mathbf{1}(t_i = 1, H_i = 0) + \lambda \mathbf{1}(t_i = 0, H_i = 1) \mid \mathbf{Z}\right]\right\}.$$

For $t_i = 1$, the above objective induces posterior risk $(1-\lambda)\mathbb{P}\left[H_i = 0 \mid \mathbf{Z}\right]$ and for $t_i = 0$, it induces posterior risk $\lambda(1 - \mathbb{P}\left[H_i = 0 \mid \mathbf{Z}\right])$. Hence we should choose $t_i = 1$, when:

$$(1-\lambda)\mathbb{P}\left[H_i = 0 \mid \mathbf{Z}\right] \leq \lambda(1 - \mathbb{P}\left[H_i = 0 \mid \mathbf{Z}\right]),$$

or equivalently:

$$\frac{\mathbb{P}\left[H_i = 0 \mid \mathbf{Z}\right]}{1 - \mathbb{P}\left[H_i = 0 \mid \mathbf{Z}\right]} \leq \frac{\lambda}{1-\lambda}.$$

Hence recalling that $\mathrm{Lfdr}_i = \mathbb{P}\left[H_i = 0 \mid \mathbf{Z}\right]$, we see that indeed the optimal decision is given by:

$$t_i(\mathbf{Z}) = \mathbf{1}(\mathbb{P}\left[H_i = 0 \mid \mathbf{Z}\right] \leq \lambda).$$

□

The result of Theorem 8.1 can be used also to prove that when searching for optimal multiple testing procedures, it often suffices to consider procedures of the form

$$\left\{\mathbf{t}(\mathbf{Z}) = (\mathbf{1}(\mathrm{Lfdr}_1 \leq \lambda), \dots, \mathbf{1}(\mathrm{Lfdr}_n \leq \lambda)) \,:\, \lambda \in [0,1]\right\}. \tag{8.11}$$

8.3.2 Empirical Bayes implementation of the local false discovery procedure

Theorem 8.1 established that optimal decision rules often reject for small values of the local false discovery rate, which however is not known. It can however be estimated through the empirical Bayes principle.

For example, suppose that we are willing to posit independence and existence of densities f_0, f so that Eq. 8.10 holds. Then one can estimate the local false discovery rate with the following two approaches.

1. **F-modeling:** Let $\hat{\pi}_0$ be an estimate of π_0 and let $\hat{f}$ be an estimate of the density of $Z_1, \dots, Z_n$ (e.g., a kernel density estimator). Then the local false discovery rate can be estimated as:
$$\widehat{\text{Lfdr}}_i = \frac{\hat{\pi}_0 f_{\text{null}}(Z_i)}{\hat{f}(Z_i)}.$$
This approach is pursued e.g., by Sun and Cai (2007), Bradley Efron (2007), and Strimmer (2008). We do not discuss estimation of π_0 here; we refer the reader to e.g., Storey (2002), Storey, Taylor, and Siegmund (2004), Langaas, Lindqvist, and Ferkingstad (2005),Meinshausen and Rice (2006), Jin (2008). In practice one often sets $\hat{\pi}_0 \equiv 1$ as a conservative choice (especially when it is believed that most null hypotheses are actually true).

2. **G-modeling:** If one is willing to also assume the full empirical Bayes model Eq. 8.5, then one could estimate $\widehat{G}$ (as in previous chapters), and then let:
$$\widehat{\text{Lfdr}}_i = \frac{\hat{\pi}_0 f_{\text{null}}(Z_i)}{f_{\widehat{G}}(Z_i)},$$
where we recall that $f_G(z) = \int p(z \mid \theta) dG(\theta)$. This approach is pursued e.g., by Scott et al. (2015),Gu and Shen (2018), Deb, Saha, Guntuboyina, and Sen (2022). A note of caution: we typically cannot estimate π_0 by $\widehat{G}(\{\theta_0\})$; the reason is that the typical weak convergence guarantees of $\widehat{G}$ to G do not lead to consistent estimation of point masses. We refer to Gu and Shen (2018) for further discussion of this point, and we suggest setting $\hat{\pi}_0 = 1$ as a reasonable default choice.

8.3.3 Empirical Bayes multiple testing decisions based on p-values

If we have already collapsed Z_i to p-values P_i, then the two-groups-model Eq. 8.7 takes the form:[1]
$$\begin{aligned} H_i &\sim \text{Bernoulli}(1-\pi_0), \\ P_i \mid H_i = 0 &\sim U[0,1], \\ P_i \mid H_i = 1 &\sim F_{\text{alt}}(\cdot). \end{aligned} \tag{8.12}$$
Analogously to Eq. 8.13, the marginal distribution of a p-value P_i is given by:
$$P_i \sim F(\cdot) := \pi_0 U[0,1] + (1-\pi_0) F_{\text{alt}}(\cdot). \tag{8.13}$$
Now suppose momentarily that following common practice, we seek multiple testing procedures that reject hypotheses with small p-values, i.e., we consider multiple testing procedures of the

[1] Below, F_{alt} refers generically to the distribution of the alternative p-values. If P_i is computed as a function of Z_i in Eq. 8.7, then the alternative distribution of Z_i will imply the alternative distribution of P_i as well. (Note the abuse of notation: F_{alt} in Eq. 8.7 and Eq. 8.12 in general do not refer to the same distribution unless Z_i is already a p-value.)

form:[2]

$$\left\{\mathbf{t}(\mathbf{P}) = \mathbf{t}_\gamma(\mathbf{P}) = (\mathbf{1}(P_1 \leq \gamma), \dots, \mathbf{1}(P_n \leq \gamma)) \,:\, \gamma \in [0,1]\right\}. \tag{8.14}$$

Does the empirical Bayes principle provide a basis for choosing among Eq. 8.14? One possibility, is the following. Let us define the marginal false discovery rate of the rule $\mathbf{t}_\gamma$ as:

$$\mathrm{mFDR}(\gamma) \equiv \mathrm{mFDR}(\mathbf{t}_\gamma) = \mathbb{P}\left[H_i = 0 \mid P_i \leq \gamma\right] = \frac{\pi_0 \gamma}{F(\gamma)}, \tag{8.15}$$

that is the posterior probability of being null conditionally on being rejected by t_γ. Suppose we seek to keep this probability below α, then we could proceed as follows:

$$\gamma^* = \sup\left\{\gamma \in [0,1] \,:\, \mathrm{mFDR}(\gamma) \leq \alpha\right\},$$

and then reject all hypotheses with $P_i \leq \gamma^*$.

The above procedure of course depends on the unknown marginal distribution $F(\cdot)$ and on π_0 in Eq. 8.15. Suppose we proceed in an empirical Bayes fashion, by first conservatively setting $\hat{\pi}_0 = 1$, and estimating $F(\cdot)$ by the empirical distribution function of $P_1, \dots, P_n$, i.e.,

$$\hat{F}(\cdot) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(P_i \leq \cdot).$$

Then we also have the following estimator of the marginal false discovery rate:

$$\widehat{\mathrm{mFDR}}(\gamma) := \frac{\gamma}{\max\left\{\hat{F}(\gamma), n^{-1}\right\}}. \tag{8.16}$$

This motivates the following multiple testing procedure:

Definition 8.4 (The Benjamini-Hochberg procedure (empirical process perspective))**.**

1. Let $\hat{\gamma}^* = \sup\left\{\gamma \in [0,1] \,:\, \widehat{\mathrm{mFDR}}(\gamma) \leq \alpha\right\}$, where $\widehat{\mathrm{mFDR}}(\gamma)$ is defined in Eq. 8.16.
2. Reject all hypotheses with $P_i \leq \hat{\gamma}^*$.

As already alluded to in the title, the above procedure is the famous multiple testing procedure of Benjamini and Hochberg (1995) that has been one of the most important innovations in statistical theory in the last three decades. We will study the procedure in more detail in the next section. Benjamini and Hochberg (1995) presented the procedure in a different, but equivalent form, and the formulation above is due to Storey, Taylor, and Siegmund (2004) (cf. Lemma 1 therein).

[2] It is worth noting that in general the following class of multiple testing rules is not equivalent to decision rules Eq. 8.11 when $\mathbf{P} \equiv \mathbf{Z}$, i.e., to rules of the form $t_i(\mathbf{P}) = \mathbf{1}(\mathrm{Lfdr}_i \leq \lambda)$ for some λ where $\mathrm{Lfdr}_i = \mathbb{P}\left[H_i = 0 \mid P_i\right]$. Hence the class Eq. 8.14 may not contain optimal rules. See Cao, Sun, and Kosorok (2013) for a discussion.

In view of Eq. 8.15, we have the following beautiful interpretation: $\widehat{\mathrm{mFDR}}(P_i)$ is the posterior probability that H_i is null given that the realized value of its p-value is below P_i. Perhaps interestingly, this is a common misconception of what a p-value is. The empirical Bayes perspective thus turned a misconception into an actionable and interpretable concept.

We now also point out another interpretation of the BH procedure. For $\gamma \in [0, 1]$, define:

$$\hat{q}(p) := \inf_{\gamma \geq p} \widehat{\mathrm{mFDR}}(\gamma). \tag{8.17}$$

In other words, the above definition implements a monotonization $\widehat{\mathrm{mFDR}}(\cdot)$: $p \mapsto \hat{q}(p)$ is always non-decreasing in p, while this won't be true for $\widehat{\mathrm{mFDR}}(\cdot)$.

We have the following further equivalent representation of the BH-procedure:

The Benjamini-Hochberg procedure (q-value perspective)

1. Compute $\hat{q}(P_1), \ldots, \hat{q}(P_n)$ with $\hat{q}(\cdot)$ defined in Eq. 8.17.
2. Reject H_i if $\hat{q}(P_i) \leq \alpha$.

$\hat{q}(P_i)$ are called q-values by Storey (2002) who introduced them. They are often also called adjusted p-values.

8.4 Middle-of-the-road: empirical Bayes powered multiple testing with frequentist guarantees

A major breakthrough in modern statistics was the definition of the false discovery rate by Benjamini and Hochberg (1995). Given a multiple testing decision procedure, let us recall that we used R to denote the total number of discoveries (Eq. 8.3) and V to denote the total number of false discoveries. The false discovery proportion FDP, is defined as follows:

$$\mathrm{FDP} := \frac{V}{\max\{R, 1\}}.$$

Note that when no discoveries are made, i.e., when $V = R = 0$, then $\mathrm{FDP} = 0$. On the other hand, when at least one discovery is made, i.e., $R \geq 1$, then the FDP is the proportion of false discoveries among all discoveries. Finally, the false discovery rate is defined as the expectation of the false discovery proportion, that is:

$$\mathrm{FDR} := \mathbb{E}\left[\mathrm{FDP}\right]. \tag{8.18}$$

Benjamini and Hochberg (1995) proposed to contruct multiple testing procedures, such that the FDR is controlled at a pre-specified level α.

At least conceptually, the motivation for defining the FDR is similar to the definition of mFDR in Eq. 8.15; however the FDR may be defined as a purely frequentist notion, without reference to the e.g., the two-groups model Eq. 8.12. In fact, as we will see below, the Benjamini-Hochberg (BH) procedure that we motivated previously by applying the empirical Bayes principle to mFDR control in the two-groups model Eq. 8.12 controls the FDR under substantially more general assumptions.

Before turning to FDR control of the BH procedure, however, we explain how the FDR can be controlled under Eq. 8.12.

8.4.1 Controlling the FDR based on local false discoveries

Suppose that the two-group model Eq. 8.7 holds for $i = 1, \dots, n$ and that Lfdr_i is known. Recall that Theorem 8.1 established that optimal decision rules reject hypotheses with small values of the Lfdr_i. However, the exact cutoff value depends on the statistical goal. Sun and Cai (2007) proposed a procedure to set this cutoff in a data-driven way so that the resulting procedure controls the false discovery rate Eq. 8.18.

Theorem 8.2 (Oracle local false discovery procedure.)**.** *Let $\alpha \in (0, 1)$ be fixed. Suppose we are testing n hypotheses under the two-groups model, and the local false discovery rate of the i-th hypothesis is equal to Lfdr_i.*

1. *Let $\mathrm{Lfdr}_{(i)}$ be the i-th order statistic of $\mathrm{Lfdr}_1, \dots, \mathrm{Lfdr}_1$, sorted from smallest to largest. To be more expicit, $\mathrm{Lfdr}_{(i)}$ are such that:*

$$\mathrm{Lfdr}_{(1)} \leq \mathrm{Lfdr}_{(2)} \leq \dots \leq \mathrm{Lfdr}_{(n)}.$$

2. *Let*

$$k^* = \max \left\{ k \in \mathbb{N}_{\geq 1} \; : \; \frac{1}{k} \sum_{i=1}^{k} \mathrm{Lfdr}_{(i)} \leq \alpha \right\},$$

   with the understanding that $\max \emptyset = 0$.

3. *If $k^* \geq 1$, reject all hypotheses with $\mathrm{Lfdr}_i \leq \mathrm{Lfdr}_{(k^*)}$, else reject no hypothesis.*

Then the FDR is controlled at level α, i,e., $\mathrm{FDR} \leq \alpha$.

Proof. We will argue conditionally on $\mathbf{Z}$. Notice that $R = k^*$ by definition and also that k^* is measurable as a function of $\mathbf{Z}$.

It will also be helpful to note the following:

$$\sum_{i=1}^{k^*} \mathrm{Lfdr}_{(i)} = \sum_{i=1}^{n} \mathrm{Lfdr}_{(i)} \mathbf{1}(t_i(\mathbf{Z}) = 1),$$

the reason being that we reject the k^* hypotheses with the smallest $\text{Lfdr}_{(i)}$.

Next we will argue conditionally on $\mathbf{Z}$.

$$\begin{aligned}
\mathbb{E}\left[\text{FDP} \mid \mathbf{Z}\right] &= \mathbb{E}\left[\frac{1}{k^*}\sum_{i=1}^{k^*}\mathbf{1}(H_i = 0)\mathbf{1}(t_i(\mathbf{Z}) = 1) \mid \mathbf{Z}\right] \\
&= \frac{1}{k^*}\sum_{i=1}^{n}\text{Lfdr}_i\mathbf{1}(t_i(\mathbf{Z}) = 1) \\
&= \frac{1}{k^*}\sum_{i=1}^{k^*}\text{Lfdr}_{(i)} \leq \alpha.
\end{aligned}$$

The last inequality follows by construction. Hence, by iterated expectation, $\text{FDR} \leq \alpha$.

□

We note that the above argument holds in finite-sample only when Lfdr_i is known exactly. If we implemented the procedure with empirical Bayes estimates of the local false discovery rates, as in Section 8.3.2, then in general control will only be asymptotic.

8.4.2 Controlling the FDR with the Benjamini-Hochberg procedure

Theorem 8.3 (Benjamini and Hochberg (1995))**.** *Let $\mathcal{H}_0 \subset \{1, \dots, n\}$ be the (deterministic) index set of null hypotheses. Suppose that $\mathbf{P} = (P_1, \dots, P_n)$ are valid p-values that are uniformly distributed for $i \in \mathcal{H}_0$, and furthermore suppose that all p-values are jointly independent.*

Then the BH procedure (Definition 8.4) controls the FDR *at level α.*

This result is not the most general. The BH procedure also has some guarantees under dependence among the p-values, see Benjamini and Yekutieli (2001).

We will provide two proofs.

8.4.2.1 Optional stopping proof

Proof. In analogy to V in Eq. 8.4 and R in Eq. 8.3, we define the false discoveries, resp. total discoveries, when rejecting all hypotheses with p-value $\leq \gamma$, that is:

$$V(\gamma) := \sum_{i \in \mathcal{H}_0}\mathbf{1}(P_i \leq \gamma), \quad R(\gamma) := \sum_{i=1}^{n}\mathbf{1}(P_i \leq \gamma).$$

Note that under the above representation, we have that the false discoveries V of BH are equal to $V(\hat{\gamma}^*)$, and similarly the total discoveries R of BH are equal to $R(\hat{\gamma}^*)$.

By definition of the BH procedure in Definition 8.4, we may also verify that:[3]

$$\widehat{\mathrm{mFDR}}(\hat{\gamma}^*) = \alpha.$$

Then:

$$\begin{aligned}
\mathrm{FDR}_{\mathrm{BH}} &= \mathbb{E}\left[\frac{V(\hat{\gamma}^*)}{\max\{R(\hat{\gamma}^*), 1\}}\right] \\
&= \mathbb{E}\left[\frac{V(\hat{\gamma}^*)}{\hat{\gamma}^*}\frac{\hat{\gamma}^*}{\max\{R(\hat{\gamma}^*), 1\}}\right] \\
&= \mathbb{E}\left[\frac{V(\hat{\gamma}^*)}{n\hat{\gamma}^*}\widehat{\mathrm{mFDR}}((\hat{\gamma}^*))\right] \\
&= \frac{\alpha}{n}\mathbb{E}\left[\frac{V(\hat{\gamma}^*)}{\hat{\gamma}^*}\right].
\end{aligned}$$

Next we define the filtration with $s \in (0, 1]$

$$\mathcal{F}_s = \sigma\left(\mathbf{1}(P_i \leq \gamma),\, \gamma \in [s, 1],\ i = 1 \ldots, n\right),$$

where $\sigma(\cdot)$ denotes the σ-algebra generated by the random variable. This is a reverse filtration since $\mathcal{F}_s \subset \mathcal{F}_{s'}$ for $s' < s$.

Let us note the following. For $i \in \mathcal{H}_0$ and for $\gamma \in (0, s)$:

$$\mathbb{E}\left[\mathbf{1}(P_i \leq \gamma) \mid \mathbf{1}(P_i \leq s)\right] = \frac{\gamma}{s}\mathbf{1}(P_i \leq s).$$

Hence: $\mathbb{E}\left[\mathbf{1}(P_i \leq \gamma) \mid \mathcal{F}_s\right] = \frac{\gamma}{s}\mathbf{1}(P_i \leq s)$. By linearity of conditional expectation:

$$\mathbb{E}\left[\frac{V(\gamma)}{\gamma} \mid \mathcal{F}_s\right] = \frac{V(s)}{s},$$

hence $V(\gamma)/\gamma$ is a reverse-time martingale with respect to the filtration $(\mathcal{F})_s$. Furthermore, one can show that $\hat{\gamma}^*$ is a (reverse-time) stopping time with respect to that filtration. Hence:

$$\mathbb{E}\left[\frac{V(\hat{\gamma}^*)}{\hat{\gamma}^*}\right] = \mathbb{E}\left[\frac{V(1)}{1}\right] = n_0.$$

Hence:

$$\mathrm{FDR}_{\mathrm{BH}} = \frac{n_0}{n}\alpha \leq \alpha.$$

□

[3]Why is that? The reason is that $\widehat{\mathrm{mFDR}}(\cdot)$ is right-continuous, piecewise linear and increasing on each segment of linearity, and $\widehat{\mathrm{mFDR}}(0) = 0$, $\widehat{\mathrm{mFDR}}(1) = 1$.

We next present an alternative proof based on a famous leave-one-out argument that appears in Ferreira and Zwinderman (2006); also see A. Li and Barber (2019) for an application of the technique to multiple testing with side-information. Before stating the proof, we first write one more equivalent representation of the BH procedure, which is the one originally presented by Benjamini and Hochberg (1995).

Definition 8.5 (The Benjamini-Hochberg procedure (step-up interpretation))**.**

1. Let $P_{(i)}$ be the i-th order statistic of the p-values $P_1, \dots, P_n$, sorted from smallest to largest.
2. Let
$$k^*_{BH} := \max\left\{k \in \{1, \dots, n\} \,:\, P_{(k)} \leq \frac{\alpha k}{n}\right\},$$
with the convention $\max \emptyset = 0$.
3. Reject the hypotheses corresponding to the k^*_{BH} smallest p-values.

Exercise 8.1 (Proof of BH equivalence (Lemma 1 in Storey, Taylor, and Siegmund (2004)))**.** Prove that the procedure in Definition 8.5 is equivalent to the procedure in Definition 8.4 in the sense that they make the same rejections.

Hint: Suppose there are no ties, then $\widehat{\mathrm{mFDR}}(P_{(k)}) = \frac{nP_{(k)}}{k}$.

With the representation of BH in Definition 8.5 in-hand, we can proceed with the proof for the BH procedure.

Proof. Our strategy is the following. Write:

$$\mathrm{FDP} = \frac{V}{\max\{R, 1\}} = \sum_{i \in \mathcal{H}_0} \frac{\mathbf{1}(H_i \text{ rejected})}{\max\{R, 1\}}.$$

Below we will prove that for any $i \in \mathcal{H}_0$:

$$\mathbb{E}\left[\frac{\mathbf{1}(H_i \text{ rejected})}{\max\{R, 1\}}\right] = \frac{\alpha}{n}. \tag{8.19}$$

Hence:

$$\mathrm{FDR} = \mathbb{E}\left[\mathrm{FDP}\right] = \frac{n_0}{n}\alpha \leq \alpha.$$

Let us turn to Eq. 8.19. Fix $i \in \mathcal{H}_0$. It is first useful to make the following observation. $R = k^*_{BH}$, and:

$$H_i \text{ rejected} \iff P_i \leq \frac{\alpha k^*_{BH}}{n}.$$

Now here comes the key trick: a leave-one-out argument. Define:

$$\mathbf{P}_{i\to 0} = (P_1, \dots, P_{i-1}, 0, P_{i+1}, \dots, P_n),$$

that is, we replace P_i by 0 in the p-value vector $\mathbf{P}$. Then let $k_{BH}^{*,i} \geq 1$ be the number of rejection of BH applied to $\mathbf{P}_{i\to 0}$.

We make the following observation:

$$H_i \text{ rejected} \implies k_{BH}^* = k_{BH}^{*,i}. \tag{8.20}$$

Why? Well, by decreasing p-values, we can only increase the number of discoveries, so that $k_{BH}^* \leq k_{BH}^{*,i}$. On the other hand, since H_i is rejected, i must be among the k_{BH}^* smallest p-values. Hence by replacing P_i by 0, we are not changing the largest $n-k$ p-values $P_{(k_{BH}^*+1)}, \dots, P_{(n)}$. Since these were not rejected to begin with, it means that,

$$P_{(j)} > \frac{\alpha j}{n} \text{ for all } j = k_{BH}^* + 1, \dots, n,$$

and this remains true after replacing P_i by 0.

Eq. 8.20 now in fact implies the following:

$$\frac{\mathbf{1}(H_i \text{ rejected})}{\max\{R, 1\}} = \frac{\mathbf{1}\left(P_i \leq \frac{\alpha k_{BH}^*}{n}\right)}{\max\{k_{BH}^*, 1\}} = \frac{\mathbf{1}\left(P_i \leq \frac{\alpha k_{BH}^{*,i}}{n}\right)}{k_{BH}^{*,i}}$$

Notice that $k_{BH}^{*,i}$ is $\mathbf{P}_{i\to 0}$ measurable, and that P_i is independent of $\mathbf{P}_{i\to 0}$. Hence:

$$\mathbb{E}\left[\frac{\mathbf{1}(H_i \text{ rejected})}{\max\{R, 1\}}\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbf{1}\left(P_i \leq \frac{\alpha k_{BH}^{*,i}}{n}\right)}{k_{BH}^{*,i}}\middle|\mathbf{P}_{i\to 0}\right]\right] = \mathbb{E}\left[\frac{1}{k_{BH}^{*,i}}\frac{\alpha k_{BH}^{*,i}}{n}\right] = \frac{\alpha}{n}.$$

□

8.5 Multiple testing with side-information

We now turn to a further demonstration of how modern large-scale inference presents us with opportunities that were not available previously. We study the problem of multiple testing with side-information. That is, we assume that for the i-th hypothesis, we do not only observe a p-value P_i, but also a covariate $X_i \in \mathcal{X}$, for some space $\mathcal{X}$.[4]

[4]The setting we consider is analogous to the setting of empirical Bayes shrinkage with side-information that we studied in Section 4.3 in Chapter 4.

Throughout this section we make the following assumption, following e.g., Bourgon, Gentleman, and Huber (2010), Ignatiadis, Klaus, Zaugg, and Huber (2016), Boca and Leek (2018):

$$P_i \text{ is independent of } X_i \text{ for all } i \in \mathcal{H}_0, \tag{8.21}$$

where as before we denote by $\mathcal{H}_0 \subset \{1, \dots, n\}$ the subset of null hypotheses.

Assumption Eq. 8.21 guarantees that X_i does not influence the calibrated null distribution of the p-values. To have potential benefits in terms of power from the side-information X_i, it is important that X_i is associated with the power of the p-value under the alternative, or with the prior probability of the hypothesing being null. If this is the case, then multiple testing procedures may gain substantial power by ordering hypotheses not only based on the p-values, but also by accounting for the side-information.

We formalize this idea with a probabilistic model in the next section; later we will derive multiple testing procedures with type-I error control under Eq. 8.21.

8.5.1 The conditional two-groups model

Our goal in this section is to consider a generalization of the two-groups model in Eq. 8.12 that also accounts for side-information X_i. This model, which we call the conditional two-groups model has been considered by several authors, including Ferkingstad et al. (2008), Scott et al. (2015), Ignatiadis, Klaus, Zaugg, and Huber (2016), Xia, Zhang, Zou, and Tse (2017), Boca and Leek (2018), Lei and Fithian (2018), Cao, Chen, and Zhang (2022), Deb, Saha, Guntuboyina, and Sen (2022).

$$\begin{aligned}
&X_i \sim \mathbb{P}^X \\
&H_i \mid X_i \sim \text{Bernoulli}(1 - \pi_0(X_i)), \\
&P_i \mid X_i, H_i = 0 \ \sim \ U[0,1], \\
&P_i \mid X_i, H_i = 1 \ \sim \ F_{\text{alt}}(\cdot \mid X_i).
\end{aligned} \tag{8.22}$$

Here $\mathbb{P}^X$ denotes the distribution of covariates X_i which we do not further model. Compared to Eq. 8.12, the key new modeling assumptions are that:

1. The probability of the i-th hypothesis being null, $\pi_0(x) = \mathbb{P}\left[H_i = 0 \mid X_i = x\right]$, can be a function of x.
2. The alternative distribution $F_{\text{alt}}(\cdot \mid X_i = x)$ can also be a function of x.

In view of Eq. 8.21, the null distribution of P_i remains uniform.

Hence the conditional two-groups model in Eq. 8.22 captures our desiderata for modeling p-values in the presence of side-information X_i. If the model Eq. 8.22 is known (and assuming

independence across i for simplicity), then one can define the local false discovery rate Eq. 8.9 as:

$$\text{Lfdr}_i := \mathbb{P}\left[H_i = 0 \mid P_i, X_i\right]. \tag{8.23}$$

The local false discovery rates can then be used to (asymptotically) control the false discovery rate using the procedure of Section 8.4.1.

8.5.2 Multiple testing with (cross)-weighting: Independent Hypothesis Weighting

Our goal now is to develop a procedure that can account for the side-information in a data-driven way, but nevertheless comes with finite-sample type-I error guarantees. Our approach will be based on multiple testing weights.[5]

The general idea of multiple testing with weights is the following. Suppose that not all hypotheses are exchangeable a-priori. Then perhaps we may seek to prioritize some hypotheses more than others. Let us express such prioritization through deterministic weights w_i, i.e., fixed numbers $w_i > 0$ such that $\sum_{i=1}^n w_i = n$. Then, for many commonly used multiple testing procedures (see Genovese, Roeder, and Wasserman (2006)), one may apply the multiple testing procedure to P_i/w_i instead of P_i. Hence, if $w_i > 1$, then it is easier to reject the i-th hypothesis, and if $w_i < 1$, then it is more difficult.

As a concrete example, we consider the weighted Bonferroni procedure. Given n p-values $P_1, \dots, P_n$, weights $w_1, \dots, w_n$, and nominal level α, the weighted Bonferroni procedure takes the following form:

$$\text{Reject } H_i \iff \frac{P_i}{w_i} \leq \frac{\alpha}{n} \iff P_i \leq \frac{\alpha w_i}{n}.$$

It is immediate to show that this procedure controls the FWER at level α:

$$\text{FWER} = \mathbb{P}\left[V \geq 1\right] \leq \mathbb{E}\left[V\right] = \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\mathbf{1}(P_i \leq \alpha w_i/n)\right] \leq \sum_{i \in \mathcal{H}_0} \frac{\alpha w_i}{n} \leq \alpha,$$

where in the last step we used the fact that:

$$\sum_{i \in \mathcal{H}_0} w_i \leq \sum_{i=1}^n w_i = n.$$

The above argument treated the weights as deterministic numbers.[6] It turns out, however, that in the presence of side-information, it is possible to construct data-driven weights. That is, one can peek at the p-values to learn weights, yet control type-I error. This is possible through

[5] Lei and Fithian (2018) propose an alternative approach based on knockoff masking for control of the false discovery rate.

[6] Quoting Genovese, Roeder, and Wasserman (2006): "Whatever information one uses to construct p-value weights, the weight assignment remains a guess. This guess is to be made a priori, that is before seeing the p-values."

cross-weighting, introduced in Ignatiadis, Klaus, Zaugg, and Huber (2016) and Ignatiadis and Huber (2021):

1. Partition $\{1, \dots, m\}$ into K disjoint folds $I_1, \dots, I_K$.
2. For all $\ell \in \{1, \dots, K\}$, learn a potentially unnormalized weighting function $\widehat{W}^{-\ell}(\cdot) : \mathcal{X} \to \mathbb{R}_+$ as a function of all covariates and the p-values in the other folds, i.e,

$$\widehat{W}^{-\ell}(\cdot) = \widehat{W}^{-\ell}(\cdot\,; (X_i : i \in I_\ell), ((X_i, P_i) : i \in \{1, \dots, n\} \setminus I_\ell)).$$

3. For $\ell \in \{1, \dots, K\}$ and for all $i \in I_\ell$, assign data-driven weights as follows:

$$W_i = \#I_\ell \cdot \widehat{W}^{-\ell}(X_i) \Big/ \sum_{j \in I_\ell} \widehat{W}^{-\ell}(X_j).$$

4. Apply a weighted multiple testing procedure (e.g., Bonferroni or Benjamini-Hochberg) with p-values P_i and weights W_i.

The above algorithm is called **IHW** (Independent Hypothesis Weighting). The second step can be implemented e.g., by positing the conditional two-groups model Eq. 8.22, estimating $\pi_0(x)$ and $F_{\text{alt}}(\cdot \mid x)$ and learning the optimal weights under the estimated model (see Ignatiadis and Huber (2021) for details). The crucial part however is that Eq. 8.22 along with any additional modeling assumptions made during the estimation step are only treated as *working assumptions*, i.e., the type-I error control guarantees do not depend on these assumptions.

We prove the simplest example of a type-I error guarantee in the case of IHW-Bonferroni (that is, IHW, with weighted Bonferroni in the last step).

Theorem 8.4 (IHW-Bonferroni controls the FWER)**.** *Suppose all pairs (P_i, X_i) are jointly independent and that $P_i \sim U[0,1]$ for $i \in \mathcal{H}_0$. Furthermore, suppose that Eq. 8.21 holds and that the fold assignment $I_1, \dots, I_K$ is deterministic (or independent of $\{(P_i, X_i) : i \in \{1, \dots, n\}\}$). Then, the IHW-Bonferroni procedure controls the FWER.*

Proof. It will suffice to note the following two properties of the data-driven weights. First, by construction:

$$\sum_{i \in I_\ell} W_i = \#I_\ell,$$

for all $\ell \in \{1, \dots, K\}$. Hence, since the folds form a partition of $\{1, \dots, n\}$, it follows that:

$$\sum_{i=1}^{n} W_i = n. \tag{8.24}$$

Next, fix $i \in \mathcal{H}_0$. Under our independence assumption it holds that P_i is independent of

$$\mathcal{O}_i := \left\{(P_j : j \in \{1, \dots, n\} \setminus \{i\}), (X_i : i \in \{1, \dots, n\})\right\}.$$

By cross-weighting, W_i is a function of $\mathcal{O}_i$, and so, W_i is independent of P_i.

Hence:

$$\mathbb{P}\left[P_i \leq \frac{\alpha W_i}{n}\right] = \mathbb{E}\left[\mathbb{P}\left[P_i \leq \frac{\alpha W_i}{n} \mid W_i\right]\right] \leq \mathbb{E}\left[\alpha \frac{W_i}{n}\right].$$

We conclude by arguing as in the case with deterministic weights.

$$\text{FWER} \leq \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\mathbf{1}(P_i \leq \alpha W_i/n)\right] \leq \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\alpha \frac{W_i}{n}\right] \leq \frac{\alpha}{n} \mathbb{E}\left[\sum_{i=1}^{n} W_i\right] = \alpha.$$

In the last step we used the fact that the weights are almost surely normalized by Eq. 8.24.

□

8.6 Bibliographic remarks

Bradley Efron (2010) is a beautiful monograph on the connection between empirical Bayes and multiple testing and elaborates in a lot more detail on several points made in this chapter. Stephens (2016) makes a very convincing argument for the new opportunies presented by multiple testing.

9 Suggested papers for presentation

9.1 Methodological papers

1. Koenker and Mizera (2014): a long awaited paper that points out that nonparametric maximum likelihood is tractable
2. McAuliffe, Blei, and Jordan (2006), and Bradley Efron (2016) present modeling strategies for flexibly modeling the prior and provide an alternative e.g., to nonparametric maximum likelihood.
3. Banerjee, Liu, Mukherjee, and Sun (2021), Banerjee, Fu, James, and Sun (2021) use kernelized Stein discrepancies for optimal shrinkage estimation via score-based f-modeling.
4. Ignatiadis, Saha, Sun, and Muralidharan (2021) propose a method for near-optimal empirical Bayes shrinkage when both the likelihood and the prior are unknown.
5. Gu and Koenker (2023) study how empirical Bayes can be used for the ranking problem
6. Wang and Stephens (2021), and Zhong, Su, and Fan (2022) develop empirical Bayes methods for matrix factorizations.
7. Kim, Wang, Carbonetto, and Stephens (2022) considers connections between empirical Bayes and penalized regression.

9.2 Application papers

1. **Genomics**: Empirical Bayes has been highly influential in genomics. Several very commonly used methods address the problem of differential gene expression:
   a. Smyth (2004) for microarray data (limma)
   b. Love, Huber, and Anders (2014) for bulk RNAseq data (DESeq2)
   c. Liu et al. (2022) for single-cell RNAseq data
2. **Astronomy**: Bovy, Hogg, and Roweis (2011)

References

Anderson, Theodore Wilbur. 1969. "Confidence Limits for the Expected Value of an Arbitrary Bounded Random Variable with a Continuous Distribution Function." *Bulletin of the International Statistical Institute* 43: 249–51.

Armstrong, Timothy B., Michal Kolesár, and Mikkel Plagborg-Møller. 2022. "Robust Empirical Bayes Confidence Intervals." *Econometrica* 90 (6): 2567–2602.

Banerjee, Trambak, Luella J. Fu, Gareth M. James, and Wenguang Sun. 2021. "Nonparametric Empirical Bayes Estimation on Heterogeneous Data." arXiv. http://arxiv.org/abs/2002.12586.

Banerjee, Trambak, Qiang Liu, Gourab Mukherjee, and Wengunag Sun. 2021. "A General Framework for Empirical Bayes Estimation in Discrete Linear Exponential Family." *Journal of Machine Learning Research* 22 (67): 1–46.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300.

Benjamini, Yoav, and Daniel Yekutieli. 2001. "The Control of the False Discovery Rate in Multiple Testing Under Dependency." *The Annals of Statistics* 29 (4): 1165–88.

Bichsel, Fritz. 1964. "Erfahrungs-Tarifierung in Der Motorfahrzeughaftpflicht-Versicherung." *Mitteilungen Der Vereinigung Schweizerischer Versicherungsmathematiker / Bulletin of the Association of Swiss Actuaries* 64: 119–30.

Billingsley, P. 1995. *Probability and Measure*. Third. Wiley Series in Probability and Statistics. New York: Wiley.

Blyth, Colin R. 1951. "On Minimax Statistical Decision Procedures and Their Admissibility." *Ann. Math. Statistics* 22: 22–42.

Boca, Simina M., and Jeffrey T. Leek. 2018. "A Direct Approach to Estimating False Discovery Rates Conditional on Covariates." *PeerJ* 6: e6035.

Böhning, D. 1999. *Computer-Assisted Analysis of Mixtures and Applications: Meta-Analysis, Disease Mapping and Others*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC.

Bourgon, Richard, Robert Gentleman, and Wolfgang Huber. 2010. "Independent Filtering Increases Detection Power for High-Throughput Experiments." *Proceedings of the National Academy of Sciences* 107 (21): 9546–51.

Bovy, Jo, David W. Hogg, and Sam T. Roweis. 2011. "Extreme Deconvolution: Inferring Complete Distribution Functions from Noisy, Heterogeneous and Incomplete Observations." *The Annals of Applied Statistics* 5 (2B).

Brown, Lawrence D, and Eitan Greenshtein. 2009. "Nonparametric Empirical Bayes and Compound Decision Approaches to Estimation of a High-Dimensional Vector of Normal Means." *The Annals of Statistics*, 1685–1704.

Brown, Lawrence D, and Linda H Zhao. 2009. "Estimators for Gaussian Models Having a Block-Wise Structure." *Statistica Sinica*, 885–903.

Bühlmann, Hans. 1964. "Optimale Prämienstufensysteme." *Mitteilungen Der Vereinigung Schweizerischer Versicherungsmathematiker / Bulletin of the Association of Swiss Actuaries* 64.

———. 1976. "Minimax Credibility." *Scandinavian Actuarial Journal* 1976 (2): 65–78.

———. 2005. *A Course in Credibility Theory and Its Applications*. Universitext. Berlin/Heidelberg: Springer-Verlag.

Cai, T. Tony, Hongzhe Li, John Maris, and Jichun Xie. 2011. "Optimal False Discovery Rate Control for Dependent Data." *Statistics and Its Interface* 4 (4): 417–30.

Cao, Hongyuan, Jun Chen, and Xianyang Zhang. 2022. "Optimal False Discovery Rate Control for Large Scale Multiple Testing with Auxiliary Information." *The Annals of Statistics* 50 (2): 807–57.

Cao, Hongyuan, Wenguang Sun, and Michael R. Kosorok. 2013. "The Optimal Power Puzzle: Scrutiny of the Monotone Likelihood Ratio Assumption in Multiple Testing." *Biometrika* 100 (2): 495–502.

Carlin, Bradley P., and Alan E. Gelfand. 1990. "Approaches for Empirical Bayes Confidence Intervals." *Journal of the American Statistical Association* 85 (409): 105–14.

———. 1991. "A Sample Reuse Method for Accurate Parametric Empirical Bayes Confidence Intervals." *Journal of the Royal Statistical Society: Series B (Methodological)* 53 (1): 189–200.

Casella, George, and J. T. Gene Hwang. 2012. "Shrinkage Confidence Procedures." *Statistical Science* 27 (1): 51–60.

Casella, George, and Jiunn Tzon Hwang. 1982. "Limit Expressions for the Risk of James-Stein Estimators." *Canadian Journal of Statistics* 10 (4): 305–9.

Chung, Kai Lai. 1974. *A Course in Probability Theory*. Second. Probability and Mathematical Statistics, Vol. 21. New York-London: Academic Press [Harcourt Brace Jovanovich, Publishers].

Cox, D. R. 1975. "Prediction Intervals and Empirical Bayes Confidence Intervals." *Journal of Applied Probability* 12 (S1): 47–55.

Deb, Nabarun, Sujayam Saha, Adityanand Guntuboyina, and Bodhisattva Sen. 2022. "Two-Component Mixture Model in the Presence of Covariates." *Journal of the American Statistical Association* 117 (540): 1820–34.

Deely, J. J., and R. L. Kruse. 1968. "Construction of Sequences Estimating the Mixing Distribution." *The Annals of Mathematical Statistics* 39 (1): 286–88.

Donoho, David L., and Iain M. Johnstone. 1995. "Adapting to Unknown Smoothness via Wavelet Shrinkage." *J. Amer. Statist. Assoc.* 90 (432): 1200–1224.

Durrett, Rick. 2010. *Probability: Theory and Examples*. Fourth. Cambridge University Press. https://doi.org/10.1017/CBO9780511779398.

Dyson, Frank. 1926. "A Method for Correcting Series of Parallax Observations." *Monthly*

Notices of the Royal Astronomical Society 86: 686.

Efron, B. 1987. "Comment on "Empirical Bayes Confidence Intervals Based on Bootstrap Samples," by N. M. Laird and T. A. Louis." *Journal of the American Statistical Association* 82 (399): 754–54.

Efron, Bradley. 2006. "Minimum Volume Confidence Regions for a Multivariate Normal Mean Vector." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68 (4): 655–70.

———. 2007. "Size, Power and False Discovery Rates." *The Annals of Statistics* 35 (4).

———. 2010. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge: Cambridge University Press.

———. 2011. "Tweedie's Formula and Selection Bias." *Journal of the American Statistical Association* 106 (496): 1602–14.

———. 2014. "Two Modeling Strategies for Empirical Bayes Estimation." *Statistical Science* 29 (2).

———. 2016. "Empirical Bayes Deconvolution Estimates." *Biometrika* 103 (1): 1–20.

———. 2019. "Bayes, Oracle Bayes and Empirical Bayes." *Statistical Science* 34 (2).

Efron, Bradley, and Trevor Hastie. 2016. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Institute of Mathematical Statistics Monographs. Cambridge: Cambridge University Press.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. "Least Angle Regression." *Ann. Statist.* 32 (2): 407–99.

Efron, Bradley, and Carl Morris. 1975. "Data Analysis Using Stein's Estimator and Its Generalizations." *Journal of the American Statistical Association* 70 (350): 311–19.

Efron, Bradley, Robert Tibshirani, John D Storey, and Virginia Tusher. 2001. "Empirical Bayes Analysis of a Microarray Experiment." *Journal of the American Statistical Association* 96 (456): 1151–60.

Everson, Phil. 2007. "A Statistician Reads the Sports Pages: Stein's Paradox Revisited." *CHANCE* 20 (3): 49–56.

Fan, Jianqing. 1991. "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems." *The Annals of Statistics* 19 (3): 1257–72.

Fay III, Robert E, and Roger A Herriot. 1979. "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data." *Journal of the American Statistical Association* 74 (366a): 269–77.

Ferkingstad, Egil, Arnoldo Frigessi, Håvard Rue, Gudmar Thorleifsson, and Augustine Kong. 2008. "Unsupervised Empirical Bayesian Multiple Testing with External Covariates." *The Annals of Applied Statistics* 2 (2): 714–35.

Ferreira, J. A., and A. H. Zwinderman. 2006. "On the Benjamini–Hochberg Method." *The Annals of Statistics* 34 (4): 1827–49.

Fisher, R. A., A. Steven Corbet, and C. B. Williams. 1943. "The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population." *The Journal of Animal Ecology* 12 (1): 42.

Folland, G. B. 1999. *Real Analysis: Modern Techniques and Their Applications*. 2nd ed. Pure

and Applied Mathematics. New York: Wiley.

Genovese, Christopher R., Kathryn Roeder, and Larry Wasserman. 2006. "False Discovery Control with p-Value Weighting." *Biometrika* 93 (3): 509–24.

Gholami, Sepideh, Lucas Janson, David J. Worhunsky, Thuy B. Tran, Malcolm Hart Squires, Linda X. Jin, Gaya Spolverato, et al. 2015. "Number of Lymph Nodes Removed and Survival After Gastric Cancer Resection: An Analysis from the US Gastric Cancer Collaborative." *Journal of the American College of Surgeons* 221 (2): 291–99.

Girshick, M. A., and L. J. Savage. 1951. "Bayes and Minimax Estimates for Quadratic Loss Functions." In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, 53–73. University of California Press, Berkeley-Los Angeles, Calif.

Good, I. J. 1992. "Introduction to Robbins (1955) An Empirical Bayes Approach to Statistics." In *Breakthroughs in Statistics*, edited by Samuel Kotz and Norman L. Johnson, 379–87. New York, NY: Springer New York.

Good, I. J., and G. H. Toulmin. 1956. "The Number of New Species, and the Increase in Population Coverage, When a Sample Is Increased." *Biometrika* 43 (1-2): 45–63.

Green, Edwin J., and William E. Strawderman. 1991. "A James-Stein Type Estimator for Combining Unbiased and Possibly Biased Estimators." *Journal of the American Statistical Association* 86 (416): 1001–6.

Greenshtein, Eitan, and Theodor Itskov. 2018. "Application of Non-Parametric Empirical Bayes to Treatment of Non-Response." *Statistica Sinica* 28 (4): 2189–2208.

Groeneboom, Piet, and Jon A. Wellner. 1992. *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Basel: Birkhäuser Basel. https://doi.org/10.1007/978-3-0348-8621-5.

Gu, Jiaying, and Roger Koenker. 2017. "Unobserved Heterogeneity in Income Dynamics: An Empirical Bayes Perspective." *Journal of Business & Economic Statistics* 35 (1): 1–16.

———. 2023. "Invidious Comparisons: Ranking and Selection as Compound Decisions." *Econometrica* 91 (1): 1–41.

Gu, Jiaying, and Shu Shen. 2018. "Oracle and Adaptive False Discovery Rate Controlling Methods for One-Sided Testing: Theory and Application in Treatment Effect Evaluation." *The Econometrics Journal* 21 (1): 11–35.

Harper, F. Maxwell, and Joseph A. Konstan. 2016. "The MovieLens Datasets: History and Context." *ACM Transactions on Interactive Intelligent Systems* 5 (4): 1–19.

Hodges, J. L., Jr., and E. L. Lehmann. 1951. "Some Applications of the Cramér-Rao Inequality." In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 1950*, 13–22. University of California Press, Berkeley-Los Angeles, Calif.

Hoff, Peter. 2022. "Coverage Properties of Empirical Bayes Intervals: A Discussion of 'Confidence Intervals for Nonparametric Empirical Bayes Analysis' by Ignatiadis and Wager." *Journal of the American Statistical Association* 117 (539): 1175–78.

Hwang, J. T. Gene, and George Casella. 1984. "Improved Set Estimators for a Multivariate Normal Mean." *Statistics & Decisions*, no. Suppl. 1: 3–16.

Ignatiadis, Nikolaos, and Wolfgang Huber. 2021. "Covariate Powered Cross-Weighted Multiple Testing." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 83 (4): 720–51.

Ignatiadis, Nikolaos, Bernd Klaus, Judith B Zaugg, and Wolfgang Huber. 2016. "Data-Driven Hypothesis Weighting Increases Detection Power in Genome-Scale Multiple Testing." *Nature Methods* 13 (7): 577–80.

Ignatiadis, Nikolaos, Sujayam Saha, Dennis L. Sun, and Omkar Muralidharan. 2021. "Empirical Bayes Mean Estimation with Nonparametric Errors via Order Statistic Regression on Replicated Data." *Journal of the American Statistical Association*, (forthcoming).

Ignatiadis, Nikolaos, and Stefan Wager. 2019. "Covariate-Powered Empirical Bayes Estimation." In *Advances in Neural Information Processing Systems*. Vol. 32.

———. 2022. "Confidence Intervals for Nonparametric Empirical Bayes Analysis." *Journal of the American Statistical Association* 117 (539): 1149–66.

James, W., and Charles Stein. 1961. "Estimation with Quadratic Loss." In *Proc. 4th Berkeley Sympos. Math. Statist. And Prob., Vol. I*, 361–79. Univ. California Press, Berkeley, Calif.

Jewell, Nicholas P. 1982. "Mixtures of Exponential Distributions." *The Annals of Statistics* 10 (2).

Jiang, Wenhua. 2019. "Comment: Empirical Bayes Interval Estimation." *Statistical Science* 34 (2).

Jiang, Wenhua, and Cun-Hui Zhang. 2009. "General Maximum Likelihood Empirical Bayes Estimation of Normal Means." *The Annals of Statistics* 37 (4): 1647–84.

———. 2010. "Empirical Bayes in-Season Prediction of Baseball Batting Averages." In *Institute of Mathematical Statistics Collections*, 263–73. Institute of Mathematical Statistics.

Jin, Jiashun. 2008. "Proportion of Non-Zero Normal Means: Universal Oracle Equivalences and Uniformly Consistent Estimators." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 70 (3): 461–93.

Johns, M. V. 1974. "Discussion of the Paper "Minimax Credibility" by Hans Bühlmann." In *Proceedings of the Berkeley Actuarial Research Conference on Credibility*. University of California, Berkeley.

Keener, Robert W. 2010a. *Theoretical Statistics*. Springer Texts in Statistics. Springer, New York.

———. 2010b. *Theoretical Statistics*. Springer Texts in Statistics. New York, NY: Springer New York.

Kiefer, Jack, and Jacob Wolfowitz. 1956. "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters." *The Annals of Mathematical Statistics* 27 (4): 887–906.

Kim, Youngseok, Peter Carbonetto, Matthew Stephens, and Mihai Anitescu. 2020. "A Fast Algorithm for Maximum Likelihood Estimation of Mixture Proportions Using Sequential Quadratic Programming." *Journal of Computational and Graphical Statistics* 29 (2): 261–73.

Kim, Youngseok, Wei Wang, Peter Carbonetto, and Matthew Stephens. 2022. "A Flexible Empirical Bayes Approach to Multiple Linear Regression and Connections with Penalized Regression." arXiv. http://arxiv.org/abs/2208.10910.

Koenker, Roger. 2020. "Empirical Bayes Confidence Intervals: An R Vinaigrette." http://www.econ.uiuc.edu/~roger/research/ebayes/cieb.pdf.

Koenker, Roger, and Jiaying Gu. 2017. "REBayes : An R Package for Empirical Bayes Mixture Methods." *Journal of Statistical Software* 82 (8).

Koenker, Roger, and Ivan Mizera. 2014. "Convex Optimization, Shape Constraints, Compound Decisions, and Empirical Bayes Rules." *Journal of the American Statistical Association* 109 (506): 674–85.

Kou, S. C., and Justin J. Yang. 2017. "Optimal Shrinkage Estimation in Heteroscedastic Hierarchical Linear Models." In *Big and Complex Data Analysis*, edited by S. Ejaz Ahmed, 249–84. Cham: Springer International Publishing.

Laird, Nan. 1978. "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution." *Journal of the American Statistical Association* 73 (364): 805–11.

Laird, Nan M, and Thomas A Louis. 1987. "Empirical Bayes Confidence Intervals Based on Bootstrap Samples." *Journal of the American Statistical Association* 82 (399): 739–50.

Langaas, Mette, Bo Henry Lindqvist, and Egil Ferkingstad. 2005. "Estimating the Proportion of True Null Hypotheses, with Application to DNA Microarray Data." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (4): 555–72.

Lehmann, Erich Leo, and George Casella. 1998. *Theory of Point Estimation*. Second. Springer Texts in Statistics. New York: Springer-Verlag.

Lehmann, Erich L., and Joseph P Romano. 2005. *Testing Statistical Hypotheses*. Springer Texts in Statistics. New York, NY: Springer New York. https://doi.org/10.1007/0-387-27605-X.

Lei, Lihua, and William Fithian. 2018. "AdaPT: An Interactive Procedure for Multiple Testing with Side Information." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80 (4): 649–79.

Li, Ang, and Rina Foygel Barber. 2019. "Multiple Testing with the Structure-Adaptive Benjamini–Hochberg Algorithm." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 81 (1): 45–74.

Li, Ker-Chau. 1985. "From Stein's Unbiased Risk Estimates to the Method of Generalized Cross Validation." *Ann. Statist.* 13 (4): 1352–77.

———. 1986. "Asymptotic Optimality of C_L and Generalized Cross-Validation in Ridge Regression with Application to Spline Smoothing." *Ann. Statist.* 14 (3): 1101–12.

———. 1987. "Asymptotic Optimality for C_p, C_L, Cross-Validation and Generalized Cross-Validation: Discrete Index Set." *Ann. Statist.* 15 (3): 958–75.

Lindsay, B. G. 1995. *Mixture Models: Theory, Geometry, and Applications*. Conference Board of the Mathematical Sciences: NSF-CBMS Regional Conference Series in Probability and Statistics. Institute of Mathematical Statistics.

Lindsay, Bruce G. 1983. "The Geometry of Mixture Likelihoods: A General Theory." *The Annals of Statistics* 11 (1): 86–94.

Lindsay, Bruce G., and Kathryn Roeder. 1993. "Uniqueness of Estimation and Identifiability in Mixture Models." *Canadian Journal of Statistics* 21 (2): 139–47.

Liu, Yusha, Peter Carbonetto, Michihiro Takahama, Adam Gruenbaum, Dongyue Xie, Nicolas Chevrier, and Matthew Stephens. 2022. "A Flexible Model for Correlated Count Data, with Application to Analysis of Gene Expression Differences in Multi-Condition Experiments." arXiv. http://arxiv.org/abs/2210.00697.

Lord, Frederic M, and Noel Cressie. 1975. "An Empirical Bayes Procedure for Finding an Interval Estimate." *Sankhyā: The Indian Journal of Statistics, Series B*, 1–9.

Lord, Frederic M, and Martha L Stocking. 1976. "An Interval Estimate for Making Statistical Inferences about True Scores." *Psychometrika* 41 (1): 79–87.

Love, Michael I, Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2." *Genome Biology* 15 (12): 550.

Massart, Pascal. 1990. "The Tight Constant in the Dvoretzky–Kiefer–Wolfowitz Inequality." *The Annals of Probability*, 1269–83.

McAuliffe, Jon D., David M. Blei, and Michael I. Jordan. 2006. "Nonparametric Empirical Bayes for the Dirichlet Process Mixture Model." *Statistics and Computing* 16 (1): 5–14.

Meinshausen, Nicolai, and John Rice. 2006. "Estimating the Proportion of False Null Hypotheses Among a Large Number of Independently Tested Hypotheses." *The Annals of Statistics* 34 (1): 373–93.

Meyer, Mary, and Michael Woodroofe. 2000. "On the Degrees of Freedom in Shape-Restricted Regression." *Ann. Statist.* 28 (4): 1083–1104.

Morris, Carl N. 1983. "Parametric Empirical Bayes Confidence Intervals." In *Scientific Inference, Data Analysis, and Robustness*, edited by G. E. P. Box, Tom Leonard, and Chien-Fu Wu, 25–50. Academic Press.

Narasimhan, Balasubramanian, and Bradley Efron. 2020. "deconvolveR: A G-modeling Program for Deconvolution and Empirical Bayes Estimation." *Journal of Statistical Software* 94 (11).

Neyman, J. 1962. "Two Breakthroughs in the Theory of Statistical Decision Making." *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* 30 (1): 11.

Polyanskiy, Yury, and Yihong Wu. 2020. "Self-Regularizing Property of Nonparametric Maximum Likelihood Estimator in Mixture Models." arXiv. http://arxiv.org/abs/2008.08244.

Pratt, John W. 1961. "Length of Confidence Intervals." *Journal of the American Statistical Association* 56 (295): 549–67.

———. 1963. "Shorter Confidence Intervals for the Mean of a Normal Distribution with Known Variance." *The Annals of Mathematical Statistics* 34 (2): 574–86.

Robbins, Herbert. 1950. "A Generalization of the Method of Maximum Likelihood: Estimating a Mixing Distribution (Abstract)." *The Annals of Mathematical Statistics* 21: 314–15.

———. 1951. "Asymptotically Subminimax Solutions of Compound Statistical Decision Problems." In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, 2:131–49. University of California Press.

———. 1956. "An Empirical Bayes Approach to Statistics." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

———. 1964. "The Empirical Bayes Approach to Statistical Decision Problems." *The Annals of Mathematical Statistics* 35 (1): 1–20.

———. 1982. "Estimating Many Variances." In *Statistical Decision Theory and Related Topics*

III, 251–61. Elsevier.

Romano, Joseph P., and Michael Wolf. 2000. "Finite Sample Nonparametric Inference and Large Sample Efficiency." *The Annals of Statistics* 28 (3): 756–78.

Saha, Sujayam, and Adityanand Guntuboyina. 2020. "On the Nonparametric Maximum Likelihood Estimator for Gaussian Location Mixture Densities with Application to Gaussian Denoising." *The Annals of Statistics* 48 (2): 738–62.

Samworth, Richard. 2005. "Small Confidence Sets for the Mean of a Spherically Symmetric Distribution." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (3): 343–61.

Scott, James G., Ryan C. Kelly, Matthew A. Smith, Pengcheng Zhou, and Robert E. Kass. 2015. "False Discovery Rate Regression: An Application to Neural Synchrony Detection in Primary Visual Cortex." *Journal of the American Statistical Association* 110 (510): 459–71.

Shen, Yandi, and Yihong Wu. 2022. "Empirical Bayes Estimation: When Does g-Modeling Beat f-Modeling in Theory (and in Practice)?" arXiv.

Smyth, Gordon K. 2004. "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments." *Statistical Applications in Genetics and Molecular Biology* 3 (1): 1–25.

Soloff, Jake A., Adityanand Guntuboyina, and Bodhisattva Sen. 2021. "Multivariate, Heteroscedastic Empirical Bayes via Nonparametric Maximum Likelihood." arXiv. http://arxiv.org/abs/2109.03466.

Stein, Charles. 1956. "Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, Vol. I*, 197–206. University of California Press, Berkeley-Los Angeles, Calif.

Stein, Charles M. 1962. "Confidence Sets for the Mean of a Multivariate Normal Distribution." *Journal of the Royal Statistical Society: Series B (Methodological)* 24 (2): 265–85.

———. 1981. "Estimation of the Mean of a Multivariate Normal Distribution." *Ann. Statist.* 9 (6): 1135–51.

Stephens, Matthew. 2016. "False Discovery Rates: A New Deal." *Biostatistics* 18 (2): 275–94.

Stigler, Stephen M. 1990. "The 1988 Neyman Memorial Lecture: A Galtonian Perspective on Shrinkage Estimators." *Statistical Science* 5 (1).

Storey, John D. 2002. "A Direct Approach to False Discovery Rates." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64 (3): 479–98.

Storey, John D., Jonathan E. Taylor, and David Siegmund. 2004. "Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66 (1): 187–205.

Strimmer, Korbinian. 2008. "A Unified Approach to False Discovery Rate Estimation." *BMC Bioinformatics* 9 (1): 303.

Sun, Wenguang, and T. Tony Cai. 2007. "Oracle and Adaptive Compound Decision Rules for False Discovery Rate Control." *Journal of the American Statistical Association* 102 (479):

901–12.

Thyrion, P. 1960. "Contribution a l'etude Du Bonus Pour Non Sinistre En Assurance Automobile." *ASTIN Bulletin* 1 (3): 142–62.

Tibshirani, Ryan J., and Saharon Rosset. 2019. "Excess Optimism: How Biased Is the Apparent Error of an Estimator Tuned by SURE?" *Journal of the American Statistical Association* 114 (526): 697–712.

Tibshirani, Ryan J., and Jonathan Taylor. 2012. "Degrees of Freedom in Lasso Problems." *Ann. Statist.* 40 (2): 1198–1232.

van der Vaart, Aad W., and Jon A. Wellner. 1996. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. New York, NY: Springer New York.

Wang, Wei, and Matthew Stephens. 2021. "Empirical Bayes Matrix Factorization." *Journal of Machine Learning Research* 22 (120): 1–40.

Xia, Fei, Martin J Zhang, James Y Zou, and David Tse. 2017. "NeuralFDR: Learning Discovery Thresholds from Hypothesis Features." In *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30.

Xie, Xianchao, SC Kou, and Lawrence D Brown. 2012. "SURE Estimates for a Heteroscedastic Hierarchical Model." *Journal of the American Statistical Association* 107 (500): 1465–79.

Yoshimori, Masayo, and Partha Lahiri. 2014. "A Second-Order Efficient Empirical Bayes Confidence Interval." *The Annals of Statistics* 42 (4): 1233–61.

Yu, C, and P D Hoff. 2018. "Adaptive Multigroup Confidence Intervals with Constant Coverage." *Biometrika* 105 (2): 319–35.

Zhang, Cun-Hui. 1990. "Fourier Methods for Estimating Mixing Densities and Distributions." *The Annals of Statistics* 18 (2): 806–31.

———. 2009. "Generalized Maximum Likelihood Estimation of Normal Mixture Densities." *Statistica Sinica*, 1297–1318.

Zhang, Yangjing, Ying Cui, Bodhisattva Sen, and Kim-Chuan Toh. 2022. "On Efficient and Scalable Computation of the Nonparametric Maximum Likelihood Estimator in Mixture Models." arXiv. http://arxiv.org/abs/2208.07514.

Zhong, Xinyi, Chang Su, and Zhou Fan. 2022. "Empirical Bayes PCA in High Dimensions." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 84 (3): 853–78.

Zwet, Erik, Simon Schwab, and Stephen Senn. 2021. "The Statistical Properties of RCTs and a Proposal for Shrinkage." *Statistics in Medicine* 40 (27): 6107–17.