CHAPTER 1

Introduction

This chapter constitutes an overview of our subject and an attempt to motivate our enthusiasm for it. The research program laid out by Robbins (1951, 1956) has captured our imagination over the last decade, and we have come to believe that it will continue to have profound and far reaching positive influence in our present era of data prosperity.

1.1 INCIDENTAL PARAMETERS

Maximum likelihood is a central precept of modern statistics, and yet in large parameter spaces it may behave badly, yielding unreliable inferences and even inconsistent estimators. Incidental parameters, in the terminology of Neyman and Scott (1948), whose dimension grows proportionally to sample size, can seriously disrupt the elegant optimality theory of the classical Fisherian paradigm. Many examples of the ill effects of incidental parameters are familiar in econometrics. Some are relatively benign, and some have proven to be more intransigent. Some may be simply swept away by some form of differencing, and others cannot. We will encounter many forms of the incidental parameter problem in what follows. The empirical Bayes approach of Robbins (1956) advocates that rather than treating incidental parameters as fixed constants to be estimated one by one, it is often preferable to estimate their *distribution*.

Motivated by the findings of Neyman and Scott (1948), Kiefer and Wolfowitz (1956) demonstrated that joint maximum likelihood estimation of structural – or what might be called coincidental – parameters and the distribution of the incidental parameters can succeed in delivering consistent estimates of both, thereby resolving the paradoxes posed by Neyman and Scott. Treating the incidental parameters as random variables rather than fixed constants leads naturally into Bayesian territory. We will argue that the traditional view of incidental parameters as "fixed effects" can benefit from this distributional perspective. Econometrics has been traditionally suspicious about such "random effects" methods, as exemplified by the influential work

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press

of Mundlak (1961, 1978), so skeptical readers should be on their guard for points at which our enthusiasm for empirical Bayes methods impinges upon prior dogma. No paradigm is a panacea, and empirical Bayes is certainly no exception. Nonetheless, empirical Bayes methods offer new insights and approaches to a wide variety of econometric problems and is therefore worthy of renewed consideration. Lancaster (2000) has ably charted the ebb and flow of econometric interest in incidental parameters over the 50 years following Neyman and Scott (1948), but it is an unfinished story.

1.2 THE UR COMPOUND DECISION PROBLEM

The origins of empirical Bayes methods may be traced to a 1951 Berkeley Symposium paper (Robbins, 1951), "Asymptotically subminimax solutions of compound statistical decision problems." In this paper Robbins mounts a direct challenge to the formulation of statistical decision theory advanced by Wald (1950) a few years earlier. We begin with a simple estimation problem, where we observe a realization, y, of a Gaussian random variable with mean $\theta \in \Theta = \{1, -1\}$, and our goal is to estimate the parameter θ. A frequentist would naturally use the maximum likelihood estimator, which is simply

$$\hat{\theta} = \underset{\theta \in \{1,-1\}}{\operatorname{argmax}} \log \phi(y - \theta) = \operatorname{sgn}(y).$$

A Bayesian, assuming a prior distribution for θ, $\mathbb{P}(\theta = 1) = p$, would update their belief about θ through the posterior and obtain

$$\mathbb{P}(\theta = 1 | Y = y) = \frac{p\phi(y-1)}{p\phi(y-1) + (1-p)\phi(y+1)}.$$

The Bayesian would then estimate θ to be 1 if $\mathbb{P}(\theta = 1|Y = y) \geq 1/2$, which results in the estimator

$$\check{\theta} = \operatorname{sgn}(y + 0.5 \log(p/(1-p))).$$

Unless $p = 1/2$, the two estimators are clearly different. For the Bayesian, if there is strong belief that $\mathbb{P}(\theta = 1) = p \neq 1/2$, rather than focusing on the sign of y, they will shift y by a log-odds term and then evaluate the sign. For example, when told that $p = 3/4$, their estimator is $\operatorname{sgn}(y + 0.549)$, making it more likely that they would estimate θ to be 1 than the frequentist. For the frequentist, however, the reliability of the imagined prior is always questionable, so the debate continues.

In Robbins (1951), Robbins made a truly remarkable suggestion. His argument is based on a twist of the original setup: Suppose that instead of having only one observation and one parameter to estimate, we observe n independent realizations, $y_1, \ldots, y_n$, of Gaussian random variables whose

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press

respective means, θ_i, are either $+1$ or -1, each with variance one. Our objective is to choose $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_n)$ to minimize the ℓ_1 loss,

$$L(\hat{\theta}, \theta) = (2n)^{-1} \sum_{i=1}^{n} |\hat{\theta}_i - \theta_i|.$$

We could, of course, ignore the fact that we now have n estimation problems and still use the maximum likelihood estimator by deciding $\hat{\theta}_i = \text{sgn}(y_i)$, which would result in an expected loss of $\Phi(-1) \approx 0.15866$. The maximum likelihood estimator is, in fact, the unique minimax decision rule for any $n \geq 1$. It corresponds to the least favorable setting of the problem, in which the probability of ± 1 is $1/2$.

We are entitled to ask, however, "is the minimax rule reasonable?" Suppose we observe that three quarters of our sample observations are positive. Should we really stick with our pessimistic premise that ± 1 are equally likely? The reliability of a prior is always questionable, so maybe it would be preferable or rely on evidence provided by the observations at hand. Robbins suggests a method of moments approach to estimate p by $\hat{p} = (\bar{y} + 1)/2$, and simply plugging it into the Bayesian estimator:

$$\hat{\theta}_i = \text{sgn}(y + 0.5 \log(\hat{p}/(1 - \hat{p}))).$$

This is the essential insight of Robbins's compound decision method: Given many similar decision problems, one can "borrow strength" from all of them to improve decision making for each.

Alternatively, one might adopt a more conventional Bayesian approach to obtain a $\hat{p}$. The log likelihood for p given the entire sample, y, is given by

$$\ell_n(p|y) = \sum_{i=1}^{n} \log(p\phi(y_i - 1) + (1 - p)\phi(y_i + 1)).$$

It is convenient to adopt a symmetric beta prior, $B(a,a)$, that becomes more concentrated around $1/2$ as the parameter a grows. The posterior mean for p can then be easily computed and used as an alternative plug-in estimator. This is similar to the idea behind hierarchical Bayes.

Figure 1.1 illustrates the advantages of such methods over the minimax rule for two sample sizes and several values of a. Of course, there is still a small risk penalty to be paid near $p = 1/2$, but this vanishes as the sample size grows, vindicating the phrase "asymptotically subminimax" in Robbins's title.

This compound statistical decision problem vividly illustrates the potential advantages of "borrowing strength" from the full sample to improve performance across multiple decision problems. Minimax procedures – and as we shall see, many likelihood based procedures – fail to exploit such opportunities. Robbins's crucial message is: When losses are aggregated over many seemingly related, or even seemingly unrelated, problems, we should be wary of decision rules that treat the problems in isolation and look for ways to treat them as an ensemble.

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press

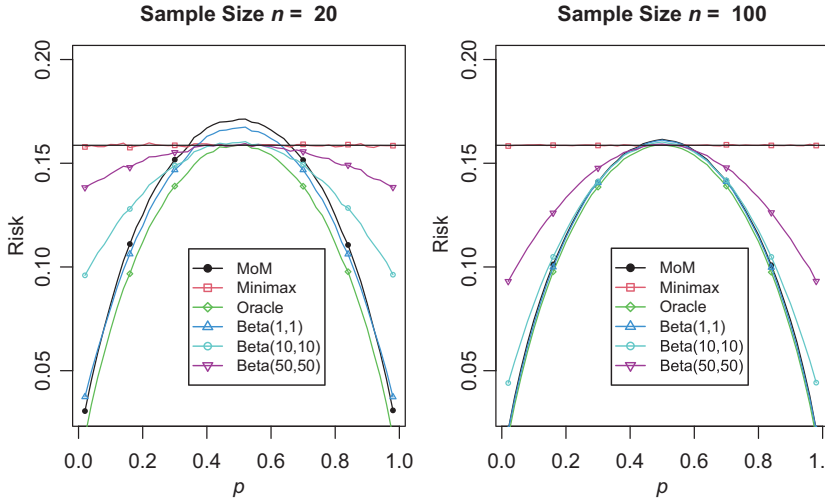


Figure 1.1. Empirical risk of several plug-in decision rules for the Ur Robbins problem. Mean loss is computed over 20,000 replications.

1.3 COMPOUND DECISIONS AND CLASSICAL DECISION THEORY

The final section of Robbins's seminal 1951 paper is devoted to embedding compound decision theory into the existing framework of Wald's decision theory. Formally, this seems straightforward; however its practical implications are quite revolutionary.

As in Wald (1950) we have a canonical problem with observations, y, on random variables, Y, from a sample space $\mathcal{Y}$, with probability distribution, P_θ, indexed by $\theta \in \Theta$. There are decision functions, $d \in \mathcal{D}$ that map actions, $a = d(y)$ from $\mathcal{Y} \to \mathcal{A}$ with loss function $L(d(y),\theta)$ and the expected loss denoted as

$$R(d,\theta) = \mathbb{E}_\theta L(d(Y),\theta) = \int L(d(y),\theta) dP_\theta(y).$$

Our objective is to choose d to minimize risk, either viewed from a Bayesian perspective as Bayes risk,

$$B(d) = \int R(d,\theta) dG(\theta),$$

where $G(\theta)$ is viewed as our prior distribution for θ, or from a minimax perspective simply as

$$\max_{\theta\in\Theta} R(d,\theta).$$

Compound decision problems expand upon this framework by considering ensembles of such problems with losses evaluated by aggregating in some

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press

fashion over the entire ensemble. The informal discussion here will be elaborated in the following chapter. For concreteness, we will consider independent instances of our canonical problem. They need not be closely related, but they should share the same problem structure, so each y_i is a realization of a Y_i that has a probability distribution P_{θ_i}. Loss is evaluated as

$$L(d,\theta) = n^{-1}\sum_{i=1}^{n} L(d_i,\theta_i).$$

If we consider decision rules of the general form $d_i(y_1,\ldots,y_n)$ and write the compound risk as

$$R_n(d,\theta) = n^{-1}\sum_{i=1}^{n}\int\ldots\int L(d_i(y_1,\ldots,y_n),\theta_i)dP_{\theta_1}(y_1)\ldots dP_{\theta_n}(y_n),$$

we can again try to minimize risk over this bigger class of decision rules. As a starting point, it is natural to assume that there is nothing special about the ordering of the problems so we can restrict attention to permutation-invariant rules. Following Robbins (1951), we will further restrict attention to what Robbins calls simple symmetric rules of the form $d_i(y_1,\ldots,y_n) = d(y_i)$. Subsequent authors, for example, Jiang and Zhang (2009), Weinstein et al. (2018) and Greenshtein and Ritov (2019), use the term "separable rules" for this class, and we will adopt this terminology.

When we restrict our decision rules d_i to depend only on their own y_i, risk is additively separable so it looks like we are led back to treating the compound problem as if it were just n independent problems. One's initial reaction to this observation is probably, "sure, why not?" The observations are assumed to be independent, so what could be gained by trying to use the y_i to inform our decision about θ_j? But we have seen that such borrowing of information across problem instances can be beneficial in the compound decision framework of the previous section. Is there a more general principle at work?

Assuming that the distributions P_θ have densities $\varphi(y|\theta)$, we can further write

$$\begin{aligned} R_n(d,\theta) &= n^{-1}\int\sum_{i=1}^{n} L(d(y),\theta_i)\varphi(y|\theta_i)dy \\ &= \int\left(\int L(d(y),\theta)\varphi(y|\theta)dG_n(\theta)\right)dy, \end{aligned}$$

where G_n denotes the empirical distribution of the θ_i's. Of course this G_n is unknown, so we would need to find a way to estimate it. Viewing the n problems as an ensemble allows us to improve upon the decision rules that would emerge from Wald's minimax approach that viewed each problem in isolation. In Robbins's Ur problem there were only two known support points for G_n, so estimation required only the relative probabilities of the two values ± 1, and this could be easily accomplished with Robbins's proposed method of moments estimator. What about the more general case?

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press

At this point, Robbins makes a truly brilliant, yet remarkably simple, suggestion. Our sample consists of independent realizations of n random variables, each with a common density function,

$$f_G(y) = \int \varphi(y|\theta) dG(\theta),$$

and therefore the likelihood of the observed sample is

$$\ell(G) = \prod_{i=1}^{n} f_G(y_i),$$

which can then be maximized with respect to G. Solutions to problems of this type, albeit infinite-dimensional, can be efficiently computed by modern convex optimization methods that we will describe in detail in Chapter 12. We will refer to these solutions as nonparametric maximum likelihood estimators (NPMLE). We will encounter many varieties of these estimators for various mixing distributions in the remainder of the book.

Whether we choose to interpret G at this stage as the empirical distribution of "fixed" θ_i's or as a subjective prior distribution on them is entirely irrelevant. To support this proposal Robbins published an abstract, Robbins (1950), claiming the consistency of such a procedure. Unfortunately, no further details seem to have been provided. It would be another five years before Kiefer and Wolfowitz (1956) established firmer foundations for this claim. A practical implementation of this method of nonparametric maximum likelihood for mixture models required an even longer wait until Laird (1978) revealed how this could be done with the nascent EM (expectation maximization) algorithm. Modern computational tools relying on advances in convex optimization have made Robbins's suggestion much more practical, and this will prove to be an essential tool in our exposition of empirical Bayes methods.

1.4 EMPIRICAL BAYESBALL: SPRING TRAINING

As a preview of this approach we may consider the problem of evaluating the batting performance of US Major League baseball players, a mainstay of the empirical Bayes literature from its earliest days of Efron and Morris (1977).[1] Following Gu and Koenker (2017a), suppose we have data on the batting performance of 898 players, excluding pitchers, from 2002 to 2011. A complicating factor with this unbalanced panel is that players have different numbers of "at bats," so measures of past performance have differing precision. Let H_{it} be the number of hits of player i in period t, and N_{it} be the number of their "at bats." Rather than relying on binomial response directly, we

[1] It is a truth universally acknowledged that a statistician in possession of a good shrinkage estimator must be desirous of predicting baseball batting averages.

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press

will consider, as in Brown (2008), the approximately Gaussian transformed response,

$$y_{it} = \arcsin\left(\sqrt{\frac{H_{it}+0.25}{N_{it}+0.5}}\right),$$

with means $\mu_i = \arcsin(\sqrt{p_i})$, where p_i is the individual specific (latent) batting success probability, and variances $\theta_i v_{it}^2 = \theta_i/(4N_{it})$. The additional individual specific scale parameter θ_i allows us to consider deviations, over or under dispersion, from the variances dictated by the binomial transformation. Following standard practice, we divide each season into two half-seasons, and for each player compute

$$\hat{\mu}_i = \left(\sum_{t=1}^{m_i} v_{it}^{-2}\right)^{-1} \sum_{t=1}^{m_i} y_{it}/v_{it}^2$$

and

$$S_i = \frac{1}{m_i - 1}\sum_{t=1}^{m_i}(y_{it} - \hat{\mu}_i)^2/v_{it}^2.$$

These sufficient statistics have Gaussian and Gamma distributions, respectively, $\hat{\mu}_i|\mu_i,\theta_i \sim \mathcal{N}(\mu_i,\theta_i v_i^2)$ and $S_i|\theta_i \sim \gamma(r_i,\theta_i/r_i)$, where we set $v_i^2 = \left(4\sum_{t=1}^{m_i} N_{it}\right)^{-1}$ and $r_i = (m_i - 1)/2$.

The likelihood of the joint density, g, of the latent parameters (μ,θ) can be written as

$$\ell(g) = K\prod_{i=1}^{n}\int\int \phi((\hat{\mu}_i - \mu)/\sqrt{\theta}v_i)/(\sqrt{\theta}v_i)\frac{e^{-R_i}R_i^{r_i}}{S_i\Gamma(r_i)}g(\mu,\theta)d\mu d\theta,$$

where $R_i = r_i S_i/\theta_i$. It is convenient at this point to make the further assumption that the mixing density $g(\mu,\theta)$ factors into $g = h_\mu h_\theta$. In this case, since the likelihood contribution of the S_i is independent of the μ_i's, we can solve the resulting (Kiefer–Wolfowitz) nonparametric maximum likelihood problem by first estimating the mixing density, h_θ, and then solving for h_μ. For the independent prior case this is especially convenient since the likelihood can be decomposed into Gamma and Student components. This independent prior case can be estimated with the function `WTLVmix` from the R package **REBayes** (Koenker and Gu, 2015–2024). Chapter 12 describes this computational strategy along with several other methods for computing the NPMLE. The independence assumption can also be alleviated, as we will discuss in Chapter 6.

In Figure 1.2 we show the resulting maximum likelihood estimates of h_μ and h_θ. Both are discrete distributions with only a few atoms. The h_θ is particularly simple with one large mass point somewhat above one indicating over-dispersion for most players, and a smaller mass point below one

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press

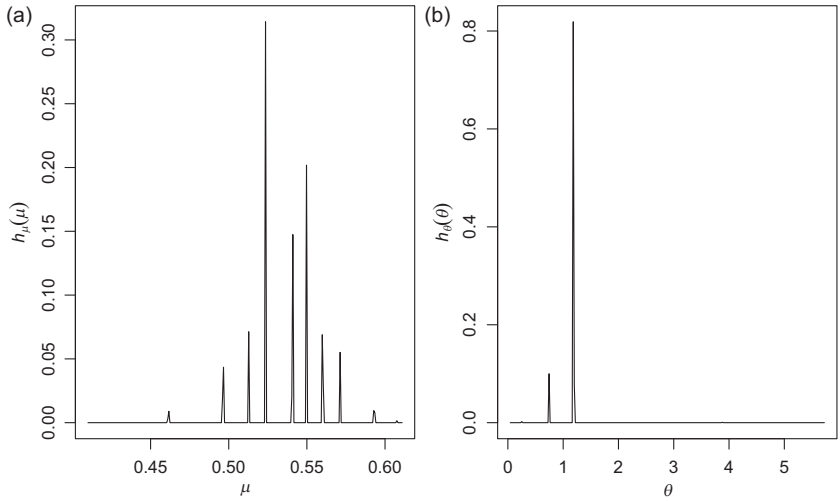


Figure 1.2. Nonparametric maximum likelihood estimates of the latent parameter distribution of a longitudinal model of baseball batting performance. Panel (a) depicts the estimated distribution of the mean of the transformed response, μ. Panel (b) shows the estimated distribution of the dispersion parameter of the transformed response, θ. The two components of the prior distribution, h_μ and h_θ, are assumed to be independent, so the likelihood can be factored into a Student component and a Gamma component, which can be estimated sequentially.

indicating under-dispersion for a smaller group of players. Under-dispersion in this setting reflects relatively stable performance over the sample period, while over-dispersion reflects more erratic batting. From these estimates it is easy to compute posterior mean estimates for μ, and after their transformation back to the original batting average scale for the parameters p_i. See Exercise 1.

It is plausible that better batting induces managers to give players more "at bats," casting doubt on the independent prior assumption. In Chapter 6 we will describe a model for longitudinal income dynamics that incorporates joint estimation of a bivariate prior for latent location and scale parameters. It is also possible to incorporate covariate effects employing profile likelihood methods: Gu and Koenker (2017a) introduce player's age as a quadratic effect; we find that batting performance is maximized at age 27.

Muting the dispersion factor θ_i and taking the model that $\hat{\mu}_i|\mu_i \sim \mathcal{N}(\mu_i, v_i^2)$, if we further assume that the latent μ_i follows a Gaussian distribution with mean μ_0 and variance σ_0^2, then we have

$$\hat{\mu}_i \sim \mathcal{N}(\mu_0, v_i^2 + \sigma_0^2).$$

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press

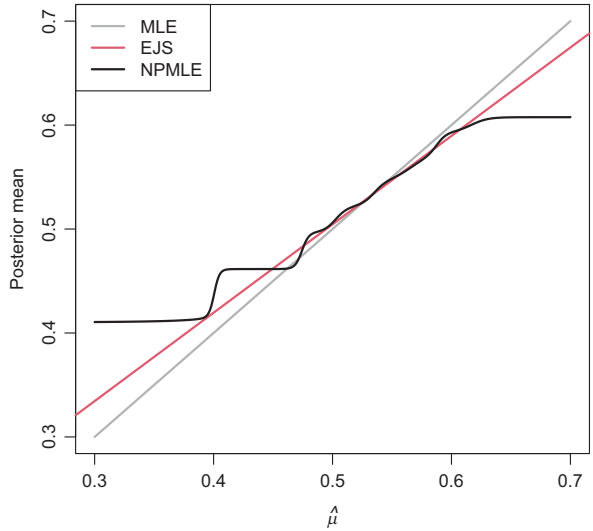


Figure 1.3. Comparison of posterior mean shrinkage rules for the Bayesball data. The linear Stein-type shrinkage rule (labeled EJS) is contrasted with the posterior mean rule derived from the NPMLE estimate shown in Figure 1.2. The gray 45° line (labeled MLE) represents the unshrunken maximum likelihood estimates.

We can estimate the unknown parameters (μ_0, σ_0^2) from the two equations

$$\bar{\mu} = n^{-1} \sum_i \hat{\mu}_i,$$
$$\bar{\sigma}^2 = n^{-1} S - n^{-1} \sum_i v_i^2,$$

with $S = \sum_i (\hat{\mu}_i - \bar{\mu})^2$. The empirical Bayes rule for estimating the latent μ_i's given the $\hat{\mu}_i$ and v_i^2 can be expressed as

$$\tilde{\mu}_i = \bar{\mu} + \Big(\frac{\bar{\sigma}^2}{\bar{\sigma}^2 + v_i^2} \Big) (\hat{\mu}_i - \bar{\mu}).$$

In Figure 1.3 we compare the shrinkage effects of this linear rule with the nonlinear shrinkage rule derived from the nonparametric maximum likelihood estimator, $\hat{h}_\mu$, both fixing $v_i = 0.01$. The two shrinkage procedures agree quite well for players in the middle of the distribution. Note that in the original "batting average" scale, players with $\hat{\mu} \in [0.5, 0.6]$ are those with hitting success rate between 0.23 and 0.319. The NPMLE posterior mean rule shrinks very aggressively in both tails, while in the shoulders of the distribution it is somewhat less aggressive than the linear rule. Efron (2010, p 9) refers to the tendency of linear Stein rules to over-shrink observations in the tails as the

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press

"tyranny of the majority." NPMLE rules, although more flexible, are obviously not fully exempt from this criticism.

1.5 FROM MOMENTS TO EXCHANGEABILITY TO MIXTURES

Moments play an important role in our subsequent developments, especially in the context of discrete response models. Method of moments estimators like Robbins's $\hat{p}$ for his 1951 problem will also appear with some regularity. A deeper connection between moment problems and the empirical Bayes paradigm is developed by Diaconis (1987) linking the Hausdorff moment problem to de Finetti's theorem.

Hausdorff (1921) considered moment sequences,

$$m_k = \int_0^1 x^k d\mu(x), \quad k = 0, 1, \ldots,$$

for probability measures, μ, on [0,1] and showed that such sequences must satisfy a strong monotonicity property. In particular, $\Delta m_k = m_{k+1} - m_k \leq 0$ and $\Delta^2 m_k = m_{k+2} - 2m_{k+1} + m_k \geq 0$. More generally, it follows that

$$(-1)^r \Delta^r m_k = \int_0^1 x^k (1-x)^r d\mu(x) \geq 0.$$

Hausdorff proved that such completely monotone sequences were the only ones that could be valid moment sequences.

Theorem 1.1 (Hausdorff). *A sequence,* $\{1, m_1, m_2, \ldots\}$ *can be a valid moment sequence for a probability measure* μ *on* $[0,1]$ *if and only if* $(-1)^r \Delta^r m_k \geq 0$ *for all* k *and* r.

De Finetti considered probability measures for infinite sequences of Bernoulli random variables and showed that such measures, P, were exchangeable, that is, were invariant to permutations,

$$P(e_1, \ldots, e_k) = P(e_{\pi(1)}, \ldots, e_{\pi(k)}),$$

for every sequence $e_1, \ldots, e_k$, all permutations, π and all k under strikingly similar conditions.

Theorem 1.2 (de Finetti). *P is exchangeable if and only if for all n,*

$$P(e_1, \ldots, e_n) = \int_0^1 x^r (1-x)^{n-r} d\mu(x),$$

where $r = e_1 + \cdots + e_n$, *and* μ *is a unique probability measure on* $[0,1]$ *not depending on n or the* e_i*'s.*

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press

Thus, P is exchangeable if and only if it can be expressed as a mixture of Bernoulli random variables with differing success probabilities. This brings us to a fundamental Bayesian decomposition that expresses the probability of elementary exchangeable events in terms of a likelihood, $x^r(1-x)^{n-r}$, and a prior μ over the parameter x. We have already seen that this is a canonical empirical Bayes setting in which likelihood is expressed in terms of some given parametric likelihood with mixing distribution represented by the prior.

Diaconis notes that Hausdorff's and de Finetti's theorems are equivalent. To show this, he defines, $m_k = P(1,\ldots,1)$, the probability of a sequence of k ones, and using $P(1) = P(1,0) + P(1,1)$, notes that

$$P(1,0) = P(1) - P(1,1) = m_1 - m_2 \geq 0,$$

and likewise,

$$\begin{aligned} P(1,0,0) &= P(1,0) - P(1,0,1) \\ &= P(1) - P(1,1) - P(1,1) + P(1,1,1) = \Delta^2 m_1 \geq 0. \end{aligned}$$

Continuing by induction, Diaconis shows that if the m_k have all such differences positive, they satisfy consistency requirements to be a valid probability, and from Hausdorff's theorem we obtain de Finetti's; conversely, if Hausdorff's condition is satisfied then de Finetti's theorem implies that $m_k = \int x^k d\mu(x)$. This is all very classical, appearing in Feller (1971), but has been greatly expanded to much larger classes of exchangeable processes that exhibit similar mixture structure. It will suffice for our purpose to note that infinite exchangeable sequences, $Y_1, Y_2, \ldots$, that is, those with joint distributions invariant to permutations, may be viewed as mixtures of conditionally iid[2] random variables. In our context this will generally be expressed by elements of the sequence, Y having common density,

$$f_G(y) = \int \varphi(y|\theta) dG(\theta)$$

with respect to some prior measure, G, on the parameter θ of some base density φ. This formalism is adequate for our present needs, but we should stress that the modern theory of exchangeable sequences and partitions thereof extends far beyond this domain. Interested readers are encouraged to consult Zabell (1992) for a more comprehensive view. A very readable introduction to modern developments is available in Diaconis and Skyrms (2018). Exchangeability and its consequential mixture representations serve as a foundation for all of Bayesian statistics, no less so than for empirical Bayes methods.

For multivariate arrays exchangeability generalizes nicely, but there are some unexpected subtleties when permutation invariance is restricted to marginal distributions. Consider P being a measure on matrices with 0–1 entries, Y_{ij}, such that

$$P(Y_{ij} : 1 \leq i, j \leq n) = P(Y_{\sigma(i)\tau(j)} : 1 \leq i, j \leq n)$$

[2] Independent and identically distributed.

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press

for all permutations, σ and τ, and all n. In the univariate setting we can imagine constructing exchangeable Bernoulli sequences in the following way: First, a θ is drawn from the distribution μ on $[0,1]$, then a coin is flipped which lands heads with probability θ; if heads is observed our entry is 1, otherwise 0. In the bivariate setting we consider an arbitrary Borel function $\psi : [0,1]^2 \to [0,1]$. We generate a pair of independent uniforms (U_i, V_j) on $[0,1]$, and now flip a $\psi(U_i, V_j)$ coin, entering a 1 if it lands heads, 0 if tails. It transpires that any row/column-exchangeable Y is a mixture of such ψ-processes. See Diaconis (2009).

1.6 COMPOUND DECISIONS AND MULTIPLE TESTING

Another theme of our exposition will be the close connection between compound decision theory and multiple testing. Robbins's Ur problem for $\theta_1, \ldots, \theta_n \in \{-1, 1\}^n$ is essentially a multiple testing problem in which we weight Type I and Type II errors equally. Treating $\theta = -1$ as the null hypothesis and $\theta = 1$ as the alternative, we can construct a p-value procedure with individual p-value, $T_i = 1 - \Phi(Y_i + 1)$, cutoff $\Phi(-1)$, and decision rule,

$$\delta(T_i) = \operatorname{sgn}(\Phi(-1) - T_i),$$

that is equivalent to our earlier minimax rule, $\delta(y) = \operatorname{sgn}(y)$.

A p-value testing procedure that is equivalent to the empirical Bayes compound decision procedure described earlier can also be constructed. Under the null hypothesis, $Y_i \sim \mathcal{N}(-1, 1)$, we know that $T_i = 1 - \Phi(Y_i + 1)$ is uniformly distributed, $U[0,1]$, while if $Y_i \sim \mathcal{N}(1,1)$,

$$\mathbb{P}(T_i < u) = \mathbb{P}(Y_i + 1 > \Phi^{-1}(1-u)) = 1 - \Phi(\Phi^{-1}(1-u) - 2).$$

Thus, under the null, the density of T is $f_0(t) = 1$ for $t \in [0,1]$, and under the alternative,

$$f_1(t) = \phi(\Phi^{-1}(1-t) - 2)/\phi(\Phi^{-1}(1-t)),$$

and the posterior probability of $\theta_i = 1$ given t_i and assuming for the moment that the unconditional probability, $p = \mathbb{P}(\theta_i = 1)$ is known, is given by

$$\mathbb{P}(\theta = 1 | t, p) = \frac{p f_1(t)}{p f_1(t) + (1-p) f_0(t)}.$$

Under symmetric loss we were led to the decision $\hat{\theta}_i = 1$ if $\mathbb{P}(\theta_i = 1 | T_i, p) > 1/2$, which is equivalent to the p-value testing rule,

$$\delta(T_i) = \operatorname{sgn}(\Phi(-1 + 0.5 \log(p/(1-p))) - T_i).$$

Again, we are led back to the problem of estimating p. In these two-point mixture problems, ℓ_1 loss is equivalent to $0-1$ loss since the median and the mode are identical. We will see in Chapter 7 that when we consider more general settings beyond the point null and point alternative then we will need

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press

to depart from the realm of p-value multiple testing. Nevertheless, the link between estimation and multiple testing will provide a fruitful perspective throughout our exposition.

1.7 PARAMETRIC VS. NONPARAMETRIC EMPIRICAL BAYES

There is an extensive literature on hierarchical Bayes modeling with influential contributions by Lindley and Smith (1972) and Chamberlain and Leamer (1976) among many others. Early on such efforts were constrained to conjugate priors for latent effects, but the rapid development of Markov chain Monte Carlo methods have opened the way toward much more flexible formulations. Still, the modeling language available in Stan (Carpenter et al., 2017) and its predecessors is predominantly parametric in form. Because these methods have authoritative expositions elsewhere, we can afford to be quite brief in our treatment of them here. Any attempt to distinguish empirical from principled Bayes methods of this sort is fraught with pointless controversy. Indeed the distinction between likelihood and prior has become somewhat blurred as models become more complex and what might have been viewed as prior parameters are now routinely estimated and even, heaven forfend, plugged in. We make some comparisons with parametric hierarchical models in later chapters, especially with Dirichlet prior methods, which are more closely related to the mixture models that predominate our exposition.

We will focus much of our attention on nonparametric empirical Bayes methods that rely on the nonparametric maximum likelihood methods illustrated in the preceding baseball application. Efron (2014) has identified two basic modeling strategies for nonparametric empirical Bayes methods. The first he calls g-modeling. It proceeds by formulating a mixture model for some latent parameters like players' batting ability, and then derives decision rules that follow from estimating the mixing distribution, G, or its density, g, by some suitable method, typically maximum likelihood. The second approach, called f-modeling, exploits the fact that in many circumstances optimal Bayes decision rules can be formulated purely in terms of the marginal density of the observed response variable. In these cases, estimation of the mixing distribution, G, can be circumvented, and decision rules can be formulated entirely in terms of an estimated marginal density, f.

A leading example of this dichotomy is the classical Gaussian mixture model. We observe

$$Y_i = \theta_i + u_i, \quad i = 1, \ldots, n,$$

with independent $u_i \sim \mathcal{N}(0, 1)$ and $\theta_i \sim G$, so the Y_i's have marginal (mixture) density,

$$f_G(y) = \int \phi(y - \theta) dG(\theta).$$

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press

The posterior mean of θ given $Y = y$, which is the optimal Bayes rule under squared error loss, is given by the Tweedie formula,

$$\mathbb{E}[\theta|Y = y] = y + f'_G(y)/f_G(y),$$

see Theorem 4.1. Thus, the analyst wanting to predict next season's batting average, y_{T+1}, from the average in the past season, y_T, might be inclined to estimate the marginal density, f_G, and to use it to predict

$$\hat{y}_{T+1} = y_T + \hat{f}'_G(y_T)/\hat{f}_G(y_T).$$

This would be an f-modeling strategy. Alternatively, we might instead favor estimating G with the NPMLE, constructing $f_{\hat{G}}$ and predicting with

$$\hat{y}_{T+1} = y_T + f'_{\hat{G}}(y_T)/f_{\hat{G}}(y_T).$$

This would be a g-modeling strategy. Both approaches have advantages and disadvantages that are application dependent. Like Efron, we will generally favor the g-modeling approach when it is feasible, but there will be several places where f-modeling makes an appealing appearance.

1.8 REGULARIZATION AND ITS TUNING

Modern statistical practice is replete with various forms of regularization intended to control smoothness, model selection and other irregularities. All such activity requires the choice of tuning parameters. From a Bayesian perspective this corresponds to a decision about how much you want to believe in your prior: How smooth should an estimated density be, or how sparse is sparse enough for an estimated regression model. Tuning parameter selection has spawned a small industry, perhaps not as scientific as piano tuning but probably employing more people. Since most of these choices are eventually resolved by some form of cross-validation or information criterion optimization, they naturally fall under the empirical Bayes rubric. And yet, we don't have anything original to add to the extensive existing literature that deals with these choices. Thus, while we will encounter various forms of tuning parameter selection in what follows, readers should not expect to encounter any magic elixirs.

Indeed, the whole idea of committing to a parametric prior up to but not including a scalar tuning parameter may seem a bit odd when it is first encountered. Why, we might ask, should we consider a prior, like that of Silverman (1982) for a univariate density, g, to look like

$$\left(\lambda \int ((\log g(x))''')^2 dx\right),$$

except that when one adjusts the λ knob one gets a nice transition from very wiggly $\hat{g}$ to very smooth, with a Gaussian limit estimate as $\lambda \to \infty$. Targeting a particular shape constraint like this is an appealing way to incorporate undogmatic forms of prior information. Parametric priors with estimated tuning parameters fall naturally into the empirical Bayes framework and will be discussed in Chapter 3.

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press

1.9 ILLICIT PRIOR ELICITATION

In the mythical world of Savage (1954), subjective priors appear fully formed as from the head of Zeus, but as anyone who has tried to elicit prior distributions even in dimension one can attest, it's not easy, as acknowledged by no less than Lindley et al. (1979):

> . . . the measurement of subjective probability is considerably more problematic and less satisfactory than the measurement of physical attributes such as length or mass and psychological attributes such as loudness or brightness. Indeed, the assessment of subjective probability is beset with severe problems of both theoretical and practical nature.

The review by Kadane and Wolfson (1998) provides a bracing description of various approaches, notably the earlier work by Kadane (1980) and Kadane et al. (1980) to extract Gaussian priors for regression settings. Thus, there has been a recent resurgence of interest in meta-analysis as a source for prior distributions. This is clearly in the spirit of Bayesian updating, and in settings with many similar problems it shares features with the mixture models we will consider. In general there seems to be a growing willingness to explore a variety of approaches to sensitivity of estimation and inference procedures to prior specifications. This may involve diagnostics, simulations or some form of external data validation.

The rigid prohibitions that one once saw regarding potential sources of prior beliefs have faded somewhat and left the field more open to empirical Bayes methods that quite explicitly confess to illicit elicitation of prior beliefs. In view of these developments, "objective Bayes" seems increasingly difficult to defend in the algorithmic climate of modern applied statistics. Empirical Bayes encompasses a wide variety of schemes to extract and use prior information from the data.

1.10 CONNECTIONS TO EMPIRICAL LIKELIHOOD

Empirical likelihood (Owen, 1990, 2001) especially in its Bayesian attire, Lazar (2003), shares many similarities with nonparametric empirical Bayes methods. However, rather than starting from a simple, scientifically plausible parametric mixture model and trying to construct a data-driven prior for the mixing distribution, empirical likelihood reverses this process and begins with a scientifically plausible prior in the form of moment conditions and seeks to estimate a data-driven likelihood. In this respect, empirical likelihood shares some features with approximate Bayesian computation, see for example, Marin et al. (2012). In both cases one arrives at nonparametric maximum likelihood procedures that promise a more flexible representation of the data than those offered by conventional parametric models. In both cases there are challenging profile likelihood problems and intriguing inferential questions. From a modeling perspective in which prior and likelihood are increasingly difficult to distinguish in many applications, it would be profitable

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press

to explore the interconnections of empirical Bayes and empirical likelihood more thoroughly. Unfortunately, we don't have anything novel to contribute to this intriguing topic.

1.11 THE WAY FORWARD

This brief introductory chapter is followed by a more detailed description in Chapter 2 of the Robbins compound decision paradigm and how it expands upon the classical framework of Wald. Parametric empirical Bayes methods exemplified by Stein estimators and hierarchical modeling are considered in Chapter 3 as a prelude to an extensive discussion of nonparametric empirical Bayes methods in Chapter 4, which constitutes the central focus of our exposition. Identification in nonparametric mixture models and profile likelihood methods are considered in Chapter 5. Heteroscedastic Gaussian mixture models are treated in Chapter 6. We return to compound decision making in Chapter 7 to consider a variety of ranking and selection applications. Empirical Bayes methods for survival and duration data are discussed in Chapter 8. Mixture models for categorical data are considered in Chapter 9. Random coefficient models for binary response and ecological inference are discussed in Chapter 10. Empirical Bayes methods have traditionally focused on prediction and point estimation, so the inference and confidence regions discussed in Chapter 11 are ripe topics for future research. Computational methods are discussed in the penultimate chapter including software implementations available in our R package **REBayes**.

Exercises

(1) Given the estimates in Figure 1.2, which can be reproduced in R with the commands: `library(REBayes); demo(Bayesball)`, transform the estimated distribution h_μ back to the original scale of p_i, the conventional batting average.
(2) Compute posterior mean estimates of μ_i as a function of $\hat{\mu}$ based on the Bayesball NPMLE using the `predict.GLmix` function of **REBayes**, and interpret the resulting "shrinkage" effects. Hint: This requires a somewhat arbitrary choice of the `newsigma` argument of the `predict` function. Try `newsigma = 0.01`, initially, and then suggest a more principled choice.
(3) Compare the posterior mean function of the previous problem with a version of the James-Stein estimator of the posterior mean.
(4) In early work on computer vision it was conjectured that 0-1 arrays generated so that (i) the chance of a one, appearing as a black pixel, in any given position was 1/2 and (ii) the chance that two ones appearing in any given pair of positions was 1/4 would be sufficient to ensure an appearance of randomness. Suggest a counterexample.

https://doi.org/10.1017/9781009275354.002 Published online by Cambridge University Press