1

Empirical Bayes and the James–Stein Estimator

Charles Stein shocked the statistical world in 1955 with his proof that maximum likelihood estimation methods for Gaussian models, in common use for more than a century, were inadmissible beyond simple one- or two-dimensional situations. These methods are still in use, for good reasons, but Stein-type estimators have pointed the way toward a radically different *empirical Bayes* approach to high-dimensional statistical inference. We will be using empirical Bayes ideas for estimation, testing, and prediction, beginning here with their path-breaking appearance in the James–Stein formulation.

Although the connection was not immediately recognized, Stein's work was half of an energetic post-war empirical Bayes initiative. The other half, explicitly named "empirical Bayes" by its principal developer Herbert Robbins, was less shocking but more general in scope, aiming to show how frequentists could achieve full Bayesian efficiency in large-scale parallel studies. Large-scale parallel studies were rare in the 1950s, however, and Robbins' theory did not have the applied impact of Stein's shrinkage estimators, which are useful in much smaller data sets.

All of this has changed in the 21st century. New scientific technologies, epitomized by the microarray, routinely produce studies of thousands of parallel cases — we will see several such studies in what follows — well-suited for the Robbins point of view. That view predominates in the succeeding chapters, though not explicitly invoking Robbins' methodology until the very last section of the book.

Stein's theory concerns estimation whereas the Robbins branch of empirical Bayes allows for hypothesis testing, that is, for situations where many or most of the true effects pile up at a specific point, usually called 0. Chapter 2 takes up large-scale hypothesis testing, where we will see, in Section 2.6, that the two branches are intertwined. Empirical Bayes theory blurs the distinction between estimation and testing as well as between fre-

quentist and Bayesian methods. This becomes clear in Chapter 2, where we will undertake frequentist estimation of Bayesian hypothesis testing rules.

1.1 Bayes Rule and Multivariate Normal Estimation

This section provides a brief review of Bayes theorem as it applies to multivariate normal estimation. Bayes rule is one of those simple but profound ideas that underlie statistical thinking. We can state it clearly in terms of densities, though it applies just as well to discrete situations. An unknown parameter vector $\boldsymbol{\mu}$ with prior density $g(\boldsymbol{\mu})$ gives rise to an observable data vector $\boldsymbol{z}$ according to density $f_{\boldsymbol{\mu}}(\boldsymbol{z})$,

$$\boldsymbol{\mu} \sim g(\cdot) \quad \text{and} \quad \boldsymbol{z}|\boldsymbol{\mu} \sim f_{\boldsymbol{\mu}}(\boldsymbol{z}). \tag{1.1}$$

Bayes rule is a formula for the conditional density of $\boldsymbol{\mu}$ having observed $\boldsymbol{z}$ (its *posterior distribution*),

$$g(\boldsymbol{\mu}|\boldsymbol{z}) = g(\boldsymbol{\mu})f_{\boldsymbol{\mu}}(\boldsymbol{z})/f(\boldsymbol{z}) \tag{1.2}$$

where $f(\boldsymbol{z})$ is the *marginal distribution* of $\boldsymbol{z}$,

$$f(\boldsymbol{z}) = \int g(\boldsymbol{\mu})f_{\boldsymbol{\mu}}(\boldsymbol{z})\, d\boldsymbol{\mu}, \tag{1.3}$$

the integral being over all values of $\boldsymbol{\mu}$.

The hardest part of (1.2), calculating $f(\boldsymbol{z})$, is usually the least necessary. Most often it is sufficient to note that the posterior density $g(\boldsymbol{\mu}|\boldsymbol{z})$ is proportional to $g(\boldsymbol{\mu})f_{\boldsymbol{\mu}}(\boldsymbol{z})$, the product of the prior density $g(\boldsymbol{\mu})$ and the *likelihood* $f_{\boldsymbol{\mu}}(\boldsymbol{z})$ of $\boldsymbol{\mu}$ given $\boldsymbol{z}$. For any two possible parameter values $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, (1.2) gives

$$\frac{g(\boldsymbol{\mu}_1|\boldsymbol{z})}{g(\boldsymbol{\mu}_2|\boldsymbol{z})} = \frac{g(\boldsymbol{\mu}_1)}{g(\boldsymbol{\mu}_2)}\,\frac{f_{\boldsymbol{\mu}_1}(\boldsymbol{z})}{f_{\boldsymbol{\mu}_2}(\boldsymbol{z})}, \tag{1.4}$$

that is, the posterior odds ratio is the prior odds ratio times the likelihood ratio. Formula (1.2) is no more than a statement of the rule of conditional probability but, as we will see, Bayes rule can have subtle and surprising consequences.

Exercise 1.1 Suppose μ has a normal prior distribution with mean 0 and variance A, while z given μ is normal with mean μ and variance 1,

$$\mu \sim \mathcal{N}(0, A) \quad \text{and} \quad z|\mu \sim \mathcal{N}(\mu, 1). \tag{1.5}$$

Show that

$$\mu|z \sim \mathcal{N}(Bz, B) \qquad \text{where } B = A/(A+1). \tag{1.6}$$

Starting down the road to large-scale inference, suppose now we are dealing with many versions of (1.5),

$$\mu_i \sim \mathcal{N}(0, A) \quad \text{and} \quad z_i|\mu_i \sim \mathcal{N}(\mu_i, 1) \qquad [i = 1, 2, \ldots, N], \tag{1.7}$$

the (μ_i, z_i) pairs being independent of each other. Letting $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_N)'$ and $\boldsymbol{z} = (z_1, z_2, \ldots, z_N)'$, we can write this compactly using standard notation for the N-dimensional normal distribution,

$$\boldsymbol{\mu} \sim \mathcal{N}_N(\mathbf{0}, AI) \tag{1.8}$$

and

$$\boldsymbol{z}|\boldsymbol{\mu} \sim \mathcal{N}_N(\boldsymbol{\mu}, I), \tag{1.9}$$

I the $N \times N$ identity matrix. Then Bayes rule gives posterior distribution

$$\boldsymbol{\mu}|\boldsymbol{z} \sim \mathcal{N}_N(B\boldsymbol{z}, BI) \qquad [B = A/(A + 1)], \tag{1.10}$$

this being (1.6) applied component-wise.

Having observed $\boldsymbol{z}$ we wish to estimate $\boldsymbol{\mu}$ with some estimator $\hat{\boldsymbol{\mu}} = t(\boldsymbol{z})$,

$$\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_N)' . \tag{1.11}$$

We use total squared error loss to measure the error of estimating $\boldsymbol{\mu}$ by $\hat{\boldsymbol{\mu}}$,

$$L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}}) = \|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = \sum_{i=1}^{N} (\hat{\mu}_i - \mu_i)^2 \tag{1.12}$$

with the corresponding risk function being the expected value of $L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})$ for a given $\boldsymbol{\mu}$,

$$R(\boldsymbol{\mu}) = E_{\boldsymbol{\mu}} \{L(\boldsymbol{\mu}, \hat{\boldsymbol{\mu}})\} = E_{\boldsymbol{\mu}} \left\{\|t(\boldsymbol{z}) - \boldsymbol{\mu}\|^2\right\}, \tag{1.13}$$

$E_{\boldsymbol{\mu}}$ indicating expectation with respect to $\boldsymbol{z} \sim \mathcal{N}_N(\boldsymbol{\mu}, I)$, $\boldsymbol{\mu}$ fixed.

The obvious estimator of $\boldsymbol{\mu}$, the one used implicitly in every regression and ANOVA application, is $\boldsymbol{z}$ itself,

$$\hat{\boldsymbol{\mu}}^{(\mathrm{MLE})} = \boldsymbol{z}, \tag{1.14}$$

the maximum likelihood estimator (MLE) of $\boldsymbol{\mu}$ in model (1.9). This has risk

$$R^{(\mathrm{MLE})}(\boldsymbol{\mu}) = N \tag{1.15}$$

for every choice of $\boldsymbol{\mu}$; every point in the parameter space is treated equally by $\hat{\boldsymbol{\mu}}^{(\mathrm{MLE})}$, which seems reasonable for general estimation purposes.

Suppose though we have prior belief (1.8) which says that $\boldsymbol{\mu}$ lies more or less near the origin $\mathbf{0}$. According to (1.10), the Bayes estimator is

$$\hat{\boldsymbol{\mu}}^{(\text{Bayes})} = B\boldsymbol{z} = \left(1 - \frac{1}{A+1}\right)\boldsymbol{z}, \tag{1.16}$$

this being the choice that minimizes the expected squared error given $\boldsymbol{z}$. If $A = 1$, for instance, $\hat{\boldsymbol{\mu}}^{(\text{Bayes})}$ shrinks $\hat{\boldsymbol{\mu}}^{(\text{MLE})}$ halfway toward $\mathbf{0}$. It has risk

$$R^{(\text{Bayes})}(\boldsymbol{\mu}) = (1-B)^2\|\boldsymbol{\mu}\|^2 + NB^2, \tag{1.17}$$

(1.13), and overall Bayes risk

$$R_A^{(\text{Bayes})} = E_A\left\{R^{(\text{Bayes})}(\boldsymbol{\mu})\right\} = N\frac{A}{A+1}, \tag{1.18}$$

E_A indicating expectation with respect to $\boldsymbol{\mu} \sim \mathcal{N}_N(\mathbf{0}, AI)$.

Exercise 1.2 Verify (1.17) and (1.18).

The corresponding Bayes risk for $\hat{\boldsymbol{\mu}}^{(\text{MLE})}$ is

$$R_A^{(\text{MLE})} = N$$

according to (1.15). If prior (1.8) is correct then $\hat{\boldsymbol{\mu}}^{(\text{Bayes})}$ offers substantial savings,

$$R_A^{(\text{MLE})} - R_A^{(\text{Bayes})} = N/(A+1); \tag{1.19}$$

with $A = 1$, $\hat{\boldsymbol{\mu}}^{(\text{Bayes})}$ removes half the risk of $\hat{\boldsymbol{\mu}}^{(\text{MLE})}$.

1.2 Empirical Bayes Estimation

Suppose model (1.8) is correct but we don't know the value of A so we can't use $\hat{\boldsymbol{\mu}}^{(\text{Bayes})}$. This is where empirical Bayes ideas make their appearance. Assumptions (1.8), (1.9) imply that the marginal distribution of $\boldsymbol{z}$ (integrating $\boldsymbol{z} \sim \mathcal{N}_N(\mu, I)$ over $\boldsymbol{\mu} \sim \mathcal{N}_N(0, A \cdot I)$) is

$$\boldsymbol{z} \sim \mathcal{N}_N\left(\mathbf{0}, (A+1)I\right). \tag{1.20}$$

The sum of squares $S = \|\boldsymbol{z}\|^2$ has a scaled chi-square distribution with N degrees of freedom,

$$S \sim (A+1)\chi^2_N, \tag{1.21}$$

so that

$$E\left\{\frac{N-2}{S}\right\} = \frac{1}{A+1}. \tag{1.22}$$

Exercise 1.3 Verify (1.22).

The *James–Stein estimator* is defined to be

$$\hat{\boldsymbol{\mu}}^{(\mathrm{JS})} = \left(1 - \frac{N-2}{S}\right)\boldsymbol{z}. \tag{1.23}$$

This is just $\hat{\boldsymbol{\mu}}^{(\mathrm{Bayes})}$ with an unbiased estimator $(N-2)/S$ substituting for the unknown term $1/(A+1)$ in (1.16). The name "empirical Bayes" is satisfyingly apt for $\hat{\boldsymbol{\mu}}^{(\mathrm{JS})}$: the Bayes estimator (1.16) is itself being empirically estimated from the data. This is only possible because we have N similar problems, $z_i \sim \mathcal{N}(\mu_i, 1)$ for $i = 1, 2, \ldots, N$, under simultaneous consideration.

It is not difficult to show that the overall Bayes risk of the James–Stein estimator is

$$R_A^{(\mathrm{JS})} = N\frac{A}{A+1} + \frac{2}{A+1}. \tag{1.24}$$

Of course this is bigger than the true Bayes risk (1.18), but the penalty is surprisingly modest,

$$R_A^{(\mathrm{JS})} \Big/ R_A^{(\mathrm{Bayes})} = 1 + \frac{2}{N \cdot A}. \tag{1.25}$$

For $N = 10$ and $A = 1$, $R_A^{(\mathrm{JS})}$ is only 20% greater than the true Bayes risk.

The shock the James–Stein estimator provided the statistical world didn't come from (1.24) or (1.25). These are based on the zero-centric Bayesian model (1.8), where the maximum likelihood estimator $\hat{\boldsymbol{\mu}}^{(0)} = \boldsymbol{z}$, which doesn't favor values of $\boldsymbol{\mu}$ near $\mathbf{0}$, might be expected to be bested. The rude surprise came from the theorem proved by James and Stein in 1961[1]:

Theorem *For $N \geq 3$, the James–Stein estimator everywhere dominates the MLE $\hat{\boldsymbol{\mu}}^{(0)}$ in terms of expected total squared error; that is*

$$E_\mu\left\{\|\hat{\boldsymbol{\mu}}^{(JS)} - \boldsymbol{\mu}\|^2\right\} < E_\mu\left\{\|\hat{\boldsymbol{\mu}}^{(\mathrm{MLE})} - \boldsymbol{\mu}\|^2\right\} \tag{1.26}$$

for every *choice of $\boldsymbol{\mu}$.*

Result (1.26) is frequentist rather that Bayesian — it implies the superiority of $\hat{\boldsymbol{\mu}}^{(\mathrm{JS})}$ no matter what one's prior beliefs about $\boldsymbol{\mu}$ may be. Since versions of $\hat{\boldsymbol{\mu}}^{(\mathrm{MLE})}$ dominate popular statistical techniques such as linear regression, its apparent uniform inferiority was a cause for alarm. The fact that linear regression applications continue unabated reflects some virtues of $\hat{\boldsymbol{\mu}}^{(\mathrm{MLE})}$ discussed later.

[1] Stein demonstrated in 1956 that $\hat{\boldsymbol{\mu}}^{(0)}$ could be everywhere improved. The specific form (1.23) was developed with his student Willard James in 1961.

A quick proof of the theorem begins with the identity

$$(\hat{\mu}_i - \mu_i)^2 = (z_i - \hat{\mu}_i)^2 - (z_i - \mu_i)^2 + 2\,(\hat{\mu}_i - \mu_i)\,(z_i - \mu_i). \tag{1.27}$$

Summing (1.27) over $i = 1, 2, \ldots, N$ and taking expectations gives

$$E_{\boldsymbol{\mu}}\left\{\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2\right\} = E_{\boldsymbol{\mu}}\left\{\|\boldsymbol{z} - \hat{\boldsymbol{\mu}}\|^2\right\} - N + 2\sum_{i=1}^{N} \mathrm{cov}_{\boldsymbol{\mu}}\,(\hat{\mu}_i, z_i)\,, \tag{1.28}$$

where $\mathrm{cov}_{\boldsymbol{\mu}}$ indicates covariance under $\boldsymbol{z} \sim \mathcal{N}_N(\boldsymbol{\mu}, I)$. Integration by parts involving the multivariate normal density function $f_{\boldsymbol{\mu}}(z) = (2\pi)^{-N/2}\exp\{-\frac{1}{2}\sum(z_i - \mu_i)^2\}$ shows that

$$\mathrm{cov}_{\boldsymbol{\mu}}\,(\hat{\mu}_i, z_i) = E_{\boldsymbol{\mu}}\left\{\frac{\partial \hat{\mu}_i}{\partial z_i}\right\} \tag{1.29}$$

as long as $\hat{\mu}_i$ is continuously differentiable in $\boldsymbol{z}$. This reduces (1.28) to

$$E_{\boldsymbol{\mu}}\,\|\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = E_{\boldsymbol{\mu}}\left\{\|\boldsymbol{z} - \hat{\boldsymbol{\mu}}\|^2\right\} - N + 2\sum_{i=1}^{N} E_{\boldsymbol{\mu}}\left\{\frac{\partial \hat{\mu}_i}{\partial z_i}\right\}. \tag{1.30}$$

Applying (1.30) to $\hat{\boldsymbol{\mu}}^{(\mathrm{JS})}$ (1.23) gives

$$E_{\boldsymbol{\mu}}\left\{\left\|\hat{\boldsymbol{\mu}}^{(\mathrm{JS})} - \boldsymbol{\mu}\right\|^2\right\} = N - E_{\boldsymbol{\mu}}\left\{\frac{(N-2)^2}{S}\right\} \tag{1.31}$$

with $S = \sum z_i^2$ as before. The last term in (1.31) is positive if N exceeds 2, proving the theorem.

Exercise 1.4 (a) Use (1.30) to verify (1.31). (b) Use (1.31) to verify (1.24).

The James–Stein estimator (1.23) shrinks each observed value z_i toward 0. We don't have to take 0 as the preferred shrinking point. A more general version of (1.8), (1.9) begins with

$$\mu_i \overset{\text{ind}}{\sim} \mathcal{N}(M, A) \quad \text{and} \quad z_i|\mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_0^2) \tag{1.32}$$

for $i = 1, 2, \ldots, N$, where M and A are the mean and variance of the prior distribution. Then (1.10) and (1.20) become

$$z_i \overset{\text{ind}}{\sim} \mathcal{N}\left(M, A + \sigma_0^2\right) \quad \text{and} \quad \mu_i|z_i \overset{\text{ind}}{\sim} \mathcal{N}\left(M + B(z_i - M), B\sigma_0^2\right) \tag{1.33}$$

for $i = 1, 2, \ldots, N$, where

$$B = \frac{A}{A + \sigma_0^2}. \tag{1.34}$$

Now Bayes rule $\hat{\mu}_i^{(\text{Bayes})} = M + B(z_i - M)$ has James–Stein empirical Bayes estimator

$$\hat{\mu}_i^{(\text{JS})} = \bar{z} + \left(1 - \frac{(N-3)\sigma_0^2}{S}\right)(z_i - \bar{z}), \tag{1.35}$$

with $\bar{z} = \sum z_i/N$ and $S = \sum(z_i - \bar{z})^2$. The theorem remains true as stated, except that we now require $N \geq 4$.

If the difference in (1.26) were tiny then $\hat{\boldsymbol{\mu}}^{(\text{JS})}$ would be no more than an interesting theoretical tidbit. In practice though, the gains from using $\hat{\boldsymbol{\mu}}^{(\text{JS})}$ can be substantial, and even, in favorable circumstances, enormous.

Table 1.1 illustrates one such circumstance. The batting averages z_i (number of successful hits divided by the number of tries) are shown for 18 major league baseball players early in the 1970 season. The true values μ_i are taken to be their averages over the remainder of the season, comprising about 370 more "at bats" each. We can imagine trying to predict the true values from the early results, using either $\hat{\mu}_i^{(\text{MLE})} = z_i$ or the James–Stein estimates (1.35) (with σ_0^2 equal the binomial estimate $\bar{z}(1-\bar{z})/45$, $\bar{z} = 0.265$ the grand average[2]). The ratio of prediction errors is

$$\sum_1^{18}\left(\hat{\mu}_i^{(\text{JS})} - \mu_i\right)^2 \Big/ \sum_1^{18}\left(\hat{\mu}_i^{(\text{MLE})} - \mu_i\right)^2 = 0.28, \tag{1.36}$$

indicating a tremendous advantage for the empirical Bayes estimates.

The initial reaction to the Stein phenomena was a feeling of paradox: Clemente, at the top of the table, is performing independently of Munson, near the bottom. Why should Clemente's good performance increase our prediction for Munson? It does for $\hat{\boldsymbol{\mu}}^{(\text{JS})}$ (mainly by increasing $\bar{z}$ in (1.35)), but not for $\hat{\boldsymbol{\mu}}^{(\text{MLE})}$. There is an implication of indirect evidence lurking *among* the players, supplementing the direct evidence of each player's own average. Formal Bayesian theory supplies the extra evidence through a prior distribution. Things are more mysterious for empirical Bayes methods, where the prior may exist only as a motivational device.

1.3 Estimating the Individual Components

Why haven't James–Stein estimators displaced MLE's in common statistical practice? The simulation study of Table 1.2 offers one answer. Here $N = 10$, with the 10 μ_i values shown in the first column; $\mu_{10} = 4$ is much

[2] The z_i are binomial here, not normal, violating the conditions of the theorem, but the James–Stein effect is quite insensitive to the exact probabilistic model.

Table 1.1 *Batting averages $z_i = \hat{\mu}_i^{(\mathrm{MLE})}$ for 18 major league players early in the 1970 season; μ_i values are averages over the remainder of the season. The James–Stein estimates $\hat{\mu}_i^{(JS)}$ (1.35) based on the z_i values provide much more accurate overall predictions for the μ_i values. (By coincidence, $\hat{\mu}_i$ and μ_i both average 0.265; the average of $\hat{\mu}_i^{(JS)}$ must equal that of $\hat{\mu}_i^{(\mathrm{MLE})}$.)*

Name	hits/AB	$\hat{\mu}_i^{(\mathrm{MLE})}$	μ_i	$\hat{\mu}_i^{(\mathrm{JS})}$
Clemente	18/45	.400	**.346**	.294
F Robinson	17/45	.378	**.298**	.289
F Howard	16/45	.356	**.276**	.285
Johnstone	15/45	.333	**.222**	.280
Berry	14/45	.311	**.273**	.275
Spencer	14/45	.311	**.270**	.275
Kessinger	13/45	.289	**.263**	.270
L Alvarado	12/45	.267	**.210**	.266
Santo	11/45	.244	**.269**	.261
Swoboda	11/45	.244	**.230**	.261
Unser	10/45	.222	**.264**	.256
Williams	10/45	.222	**.256**	.256
Scott	10/45	.222	**.303**	.256
Petrocelli	10/45	.222	**.264**	.256
E Rodriguez	10/45	.222	**.226**	.256
Campaneris	9/45	.200	**.286**	.252
Munson	8/45	.178	**.316**	.247
Alvis	7/45	.156	**.200**	.242
Grand Average		.265	**.265**	.265

different than the others. One thousand simulations of $\boldsymbol{z} \sim \mathcal{N}_{10}(\boldsymbol{\mu}, I)$ each gave estimates $\hat{\boldsymbol{\mu}}^{(\mathrm{MLE})} = \boldsymbol{z}$ and $\hat{\boldsymbol{\mu}}^{(\mathrm{JS})}$ (1.23). Average squared errors for each μ_i are shown. For example $(\hat{\mu}_1^{(\mathrm{MLE})} - \mu_1)^2$ averaged 0.95 over the 1000 simulations, compared to 0.61 for $(\hat{\mu}_1^{(\mathrm{JS})} - \mu_1)^2$.

We see that $\hat{\mu}_i^{(\mathrm{JS})}$ gave better estimates than $\hat{\mu}_i^{(\mathrm{MLE})}$ for the first nine cases, but was much *worse* for estimating the outlying case μ_{10}. Overall, the total mean squared errors favored $\boldsymbol{\mu}^{(\mathrm{JS})}$, as they must.

Exercise 1.5 If we assume that the μ_i values in Table 1.2 were obtained from $\mu_i \overset{\mathrm{ind}}{\sim} \mathcal{N}(0, A)$, is the total error 8.13 about right?

The James–Stein theorem concentrates attention on the total squared error loss function $\sum(\hat{\mu}_i - \mu_i)^2$, without concern for the effects on individual cases. Most of those effects are good, as seen in Table 1.2, but genuinely

Table 1.2 *Simulation experiment:* $\boldsymbol{z} \sim \mathcal{N}_{10}(\boldsymbol{\mu}, I)$ *with* $(\mu_1, \mu_s, \ldots, \mu_{10})$ *as shown in first column.* $\mathrm{MSE}_i^{(\mathrm{MLE})}$ *is the average squared error* $(\hat{\mu}_i^{(\mathrm{MLE})} - \mu_i)^2$, *likewise* $\mathrm{MSE}_i^{(JS)}$. *Nine of the cases are better estimated by James–Stein, but for the outlying case 10,* $\hat{\mu}_{10}^{(JS)}$ *has nearly twice the error of* $\hat{\mu}_{10}^{(\mathrm{MLE})}$.

	μ_i	$\mathrm{MSE}_i^{(\mathrm{MLE})}$	$\mathrm{MSE}_i^{(\mathrm{JS})}$
1	−.81	.95	.61
2	−.39	1.04	.62
3	−.39	1.03	.62
4	−.08	.99	.58
5	.69	1.06	.67
6	.71	.98	.63
7	1.28	.95	.71
8	1.32	1.04	.77
9	1.89	1.00	.88
10	4.00	1.08	2.04!!
Total Sqerr		10.12	8.13

unusual cases, like μ_{10}, can suffer. Baseball fans know that Clemente was in fact an extraordinarily good hitter, and shouldn't have been shrunk so drastically toward the mean of his less-talented cohort. Current statistical practice is quite conservative in protecting individual inferences from the tyranny of the majority, accounting for the continued popularity of stand-alone methods like $\hat{\boldsymbol{\mu}}^{(\mathrm{MLE})}$. On the other hand, large-scale simultaneous inference, our general theme here, focuses on favorable group inferences.

Compromise methods are available, that capture most of the group savings while protecting unusual individual cases. In the baseball example, for instance, we might decide to follow the James–Stein estimate (1.35) subject to the restriction of not deviating more than $D\,\sigma_0$ units away from $\hat{\mu}_i^{(\mathrm{MLE})} = z_i$ (the so-called "limited translation estimator" $\hat{\mu}_i^{(D)}$):

$$\hat{\mu}_i^{(D)} = \begin{cases} \max\left(\hat{\mu}_i^{(\mathrm{JS})}, \hat{\mu}_i^{(\mathrm{MLE})} - D\sigma_0\right) & \text{for } z_i > \bar{z} \\ \min\left(\hat{\mu}_i^{(\mathrm{JS})}, \hat{\mu}_i^{(\mathrm{MLE})} + D\sigma_0\right) & \text{for } z_i \le \bar{z}. \end{cases} \tag{1.37}$$

Exercise 1.6 Graph $\hat{\mu}_i^{(D)}$ as a function of z_i for the baseball data.

Taking $D = 1$ says that $\hat{\mu}_i^{(D)}$ will never deviate more than $\sigma_0 = 0.066$ from z_i, so Clemente's prediction would be $\hat{\mu}_1^{(D)} = 0.334$ rather than $\hat{\mu}_1^{(\mathrm{JS})} =$

0.294. This sacrifices some of the $\hat{\boldsymbol{\mu}}^{(\mathrm{JS})}$ savings relative to $\hat{\boldsymbol{\mu}}^{(\mathrm{MLE})}$, but not a great deal: it can be shown to lose only about 10% of the overall James–Stein advantage in the baseball example.

1.4 Learning from the Experience of Others

Bayes and empirical Bayes techniques involve learning from the experience of others, e.g., each baseball player learning from the other 17. This always raises the question, "Which others?" Chapter 10 returns to this question in the context of hypothesis testing. There we will have thousands of other cases, rather than 17, vastly increasing the amount of "other" experience.

Figure 1.1 diagrams James–Stein estimation, with case 1 learning from the $N-1$ others. We imagine that the others have been observed first, giving estimates $(\hat{M}, \hat{A})$ for the unknown Bayes parameters in (1.32) (taking $\sigma_0^2 = 1$). The estimated prior distribution $\mathcal{N}(\hat{M}, \hat{A})$ is then used to supplement the direct evidence $z_1 \sim \mathcal{N}(\mu_1, 1)$ for the estimation of μ_1. (Actually $\hat{\mu}_i^{(\mathrm{JS})}$ includes z_i as well as the others in estimating $(\hat{M}, \hat{A})$ for use on μ_1: it can be shown that this improves the accuracy of $\hat{\mu}_1^{(\mathrm{JS})}$.) Versions of this same diagram apply to the more intricate empirical Bayes procedures that follow.

Learning from the experience of others is not the sole property of the Bayes world. Figure 1.2 illustrates a common statistical situation. $N = 157$ healthy volunteers have had their kidney function evaluated by a somewhat arduous medical procedure. The scores are plotted versus age, higher scores indicating better function, and it is obvious that function tends to decrease with age. (At one time, kidney donation was forbidden for donors exceeding 60, though increasing demand has relaxed this rule.) The heavy line indicates the least squares fit of function to age.

A potential new donor, age 55, has appeared, but it is not practical to evaluate his kidney function by the arduous medical procedure. Figure 1.2 shows two possible predictions: the starred point is the function score (-0.01) for the only 55-year-old person among the 157 volunteers, while the squared point reads off the value of the least square line (-1.46) at age = 55. Most statisticians, frequentist or Bayesian, would prefer the least squares prediction.

Tukey's evocative term "borrowing strength" neatly captures the regression idea. This is certainly "learning from the experience of others", but in a more rigid framework than Figure 1.1. Here there is a simple covari-

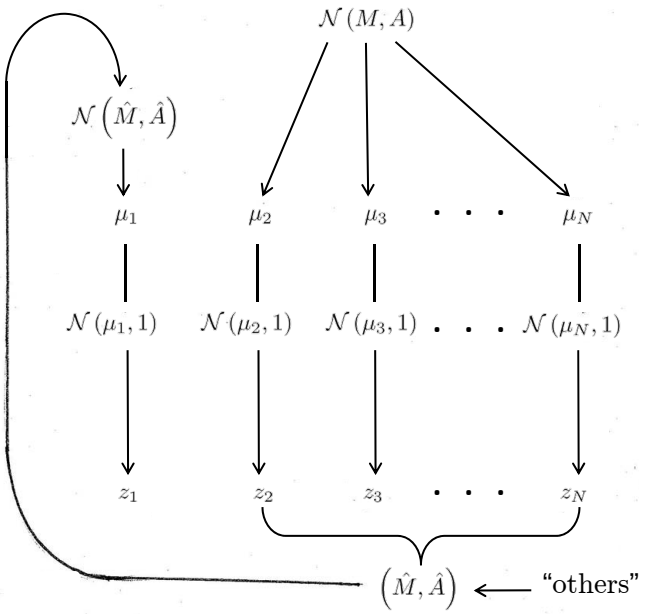

Figure 1.1 Schematic diagram of James–Stein estimation, showing case 1 learning from the experience of the other $N - 1$ cases.

ate, age, convincingly linking the volunteers with the potential donor. The linkage is more subtle in the baseball example.

Often the two methods can be combined. We might extend model (1.32) to

$$\mu_i \overset{\text{ind}}{\sim} \mathcal{N}(M_0 + M_1 \cdot \text{age}_i, A) \quad \text{and} \quad z_i \sim \mathcal{N}\left(\mu_i, \sigma_0^2\right). \tag{1.38}$$

The James–Stein estimate (1.35) takes the form

$$\hat{\mu}_i^{(\text{JS})} = \hat{\mu}_i^{(\text{reg})} + \left(1 - \frac{(N-4)\sigma_0^2}{S}\right)\left(z_i - \hat{\mu}_i^{(\text{reg})}\right), \tag{1.39}$$

where $\hat{\mu}_i^{(\text{reg})}$ is the linear regression estimate $(\hat{M}_0 + \hat{M}_1 \cdot \text{age}_i)$ and $S = \sum(z_i - \hat{\mu}_i^{(\text{reg})})^2$. Now $\hat{\mu}_i^{(\text{JS})}$ is shrunk toward the linear regression line instead of toward $\bar{z}$.

Exercise 1.7 $S = 503$ for the kidney data. Assuming $\sigma_0^2 = 1$, what is the James–Stein estimate for the starred point in Figure 1.2 (i.e., for the healthy volunteer, age 55)?

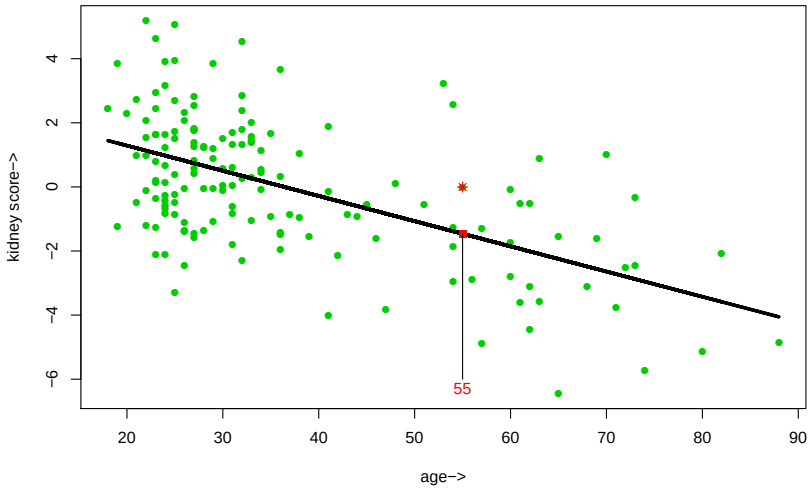

Figure 1.2 Kidney scores plotted versus age for 157 healthy volunteers. The least squares line shows the decrease of function with age. How should we predict the score of a potential donor, age 55?

1.5 Empirical Bayes Confidence Intervals

Returning to the situation in Section 1.1, suppose we have $N + 1$ independent normal observations z_i, with

$$\mu_i \overset{\text{ind}}{\sim} \mathcal{N}(0, A) \quad \text{and} \quad z_i|\mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1) \tag{1.40}$$

for $i = 0, 1, 2, \ldots, N$, and we want to assign a "confidence interval" to the parameter μ_0. The quotes are necessary here because we wish to take advantage of empirical Bayes information as in Figure 1.1, now with the "others" being $\boldsymbol{z} = (z_1, z_2, \ldots, z_N)$ and with (μ_0, z_0) playing the role of (μ_1, z_1) — taking us beyond classical confidence interval methodology.

If A were known we could calculate the Bayes posterior distribution for μ_0 according to (1.10),

$$\mu_0|z_0 \sim \mathcal{N}(Bz_0, B) \qquad [B = A/(A + 1)]\,, \tag{1.41}$$

yielding

$$\mu_0 \in Bz_0 \pm 1.96\sqrt{B} \tag{1.42}$$

as the obvious 95% posterior interval. A reasonable first try in the empirical

Bayes situation of Section 1.2 is to substitute the unbiased estimate

$$\hat{B} = 1 - \frac{N-2}{S} \qquad \left[S = \|\boldsymbol{z}\|^2\right] \tag{1.43}$$

into (1.41), giving the approximation

$$\mu_0|z_0, \boldsymbol{z} \doteq \mathcal{N}\left(\hat{B}z_0, \hat{B}\right) \tag{1.44}$$

and similarly $\hat{B}z_0 \pm 1.96\sqrt{\hat{B}}$ for (1.42). In doing so, however, we have ignored the variability of $\hat{B}$ as an estimate of B, which can be substantial when N is small.

Here is a more accurate version of (1.44):

$$\mu_0|z_0, \boldsymbol{z} \doteq \mathcal{N}\left(\hat{B}z_0,\ \hat{B} + \frac{2}{N-2}\left[z_0\left(1-\hat{B}\right)\right]^2\right) \tag{1.45}$$

and its corresponding posterior interval

$$\mu_0 \in \hat{B}z_0 \pm 1.96\left\{\hat{B} + \frac{2}{N-2}\left[z_0\left(1-\hat{B}\right)\right]^2\right\}^{\frac{1}{2}}. \tag{1.46}$$

Exercise 1.8 (a) Show that the relative length of (1.46) compared to the interval based on (1.44) is

$$\left\{1 + \frac{2}{N-2}\frac{z_0^2\left(1-\hat{B}\right)^2}{\hat{B}}\right\}^{\frac{1}{2}}. \tag{1.47}$$

(b) For $N = 17$ and $\hat{B} = 0.21$ (appropriate values for the baseball example), graph (1.47) for z_0 between 0 and 3.

Formula (1.45) can be justified by carefully following through a simplified version of Figure 1.1 in which $M = 0$, using familiar maximum likelihood calculations to assess the variability of $\hat{A}$ and its effect on the empirical Bayes estimation of μ_0 (called μ_1 in the figure).

Hierarchical Bayes methods offer another justification. Here the model (1.40) would be preceded by some Bayesian prior assumption on the *hyperparameter* A, perhaps A uniformly distributed over $(0, 10^6)$, chosen not to add much information beyond that in $\boldsymbol{z}$ to A's estimation. The term *objective Bayes* is used to describe such arguments, which are often insightful and useful. Defining $V = A + 1$ in model (1.40) and formally applying Bayes rule to the (impossible) prior that takes V to be uniformly distributed over $(0, \infty)$ yields exactly the posterior mean and variance in (1.45).

Notes

Herbert Robbins, paralleling early work by R.A. Fisher, I.J. Good, and Alan Turing (of Turing machine fame) developed a powerful theory of empirical Bayes statistical inference, some references being Robbins (1956) and Efron (2003). Robbins reserved the name "empirical Bayes" for situations where a genuine prior distribution like (1.8) was being estimated, using "compound Bayes" for more general parallel estimation and testing situations, but Efron and Morris (1973) hijacked "empirical Bayes" for James–Stein-type estimators.

Stein (1956) and James and Stein (1961) were written entirely from a frequentist point of view, which has much to do with their bombshell effect on the overwhelmingly frequentist statistical literature of that time. Stein (1981) gives the neat identity (1.28) and the concise proof of the theorem.

Limited translation estimates (1.37) were developed in Efron and Morris (1972), amid a more general theory of *relevance functions*, modifications of the James–Stein estimator that allowed individual cases to partially opt out of the overall shrinkage depending on how relevant the other cases appeared to be. Relevance functions for hypothesis testing will be taken up here in Chapter 10. Efron (1996) gives a more general version of Figure 1.1.

The kidney data in Figure 1.2 is from the Stanford nephrology lab of Dr. B. Myers; see Lemley et al. (2008). Morris (1983) gives a careful derivation of empirical Bayes confidence intervals such as (1.46), along with an informative discussion of what one should expect from such intervals.