21

Empirical Bayes Estimation Strategies

Classic statistical inference was focused on the analysis of individual cases: a single estimate, a single hypothesis test. The interpretation of direct evidence bearing on the case of interest—the number of successes and failures of a new drug in a clinical trial as a familiar example—dominated statistical practice.

The story of modern statistics very much involves indirect evidence, "learning from the experience of others" in the language of Sections 7.4 and 15.3, carried out in both frequentist and Bayesian settings. The computer-intensive prediction algorithms described in Chapters 16–19 use regression theory, the frequentist's favored technique, to mine indirect evidence on a massive scale. False-discovery rate theory, Chapter 15, collects indirect evidence for hypothesis testing by means of Bayes' theorem as implemented through empirical Bayes estimation.

Empirical Bayes methodology has been less studied than Bayesian or frequentist theory. As with the James–Stein estimator (7.13), it can seem to be little more than plugging obvious frequentist estimates into Bayes estimation rules. This conceals a subtle and difficult task: learning the equivalent of a Bayesian prior distribution from ongoing statistical observations. Our final chapter concerns the empirical Bayes learning process, both as an exercise in applied deconvolution and as a relatively new form of statistical inference. This puts us back where we began in Chapter 1, examining the two faces of statistical analysis, the algorithmic and the inferential.

21.1 Bayes Deconvolution

A familiar formulation of empirical Bayes inference begins by assuming that an unknown prior density $g(\theta)$, our object of interest, has produced a random sample of real-valued variates $\Theta_1, \Theta_2, \dots, \Theta_N$,

$$\Theta_i \overset{\text{iid}}{\sim} g(\theta), \qquad i = 1, 2, \dots, N. \tag{21.1}$$

(The "density" $g(\cdot)$ may include discrete atoms of probability.) The Θ_i are unobservable, but each yields an observable random variable X_i according to a known family of density functions

$$X_i \overset{\text{ind}}{\sim} p_i(X_i|\Theta_i). \tag{21.2}$$

From the observed sample $X_1, X_2, \dots, X_N$ we wish to estimate the prior density $g(\theta)$.

A famous example has $p_i(X_i|\Theta_i)$ the Poisson family,

$$X_i \sim \text{Poi}(\Theta_i), \tag{21.3}$$

as in Robbins' formula, Section 6.1. Still more familiar is the normal model (3.28),

$$X_i \sim \mathcal{N}(\Theta_i, \sigma^2), \tag{21.4}$$

often with $\sigma^2 = 1$. A binomial model was used in the medical example of Section 6.3,

$$X_i \sim \text{Bi}(n_i, \Theta_i). \tag{21.5}$$

There the n_i differ from case to case, accounting for the need for the first subscript i in $p_i(X_i|\Theta_i)$ (21.2).

Let $f_i(X_i)$ denote the *marginal density* of X_i obtained from (21.1)–(21.2),

$$f_i(X_i) = \int_{\mathcal{T}} p_i(X_i|\theta_i) g(\theta_i)\, d\theta_i, \tag{21.6}$$

the integral being over the space $\mathcal{T}$ of possible Θ values. The statistician has only the marginal observations available,

$$X_i \overset{\text{ind}}{\sim} f_i(\cdot), \qquad i = 1, 2, \dots, N, \tag{21.7}$$

from which he or she wishes to estimate the density $g(\cdot)$ in (21.6).

In the normal model (21.4), f_i is the convolution of the unknown $g(\theta)$ with a known normal density, denoted

$$f = g * \mathcal{N}(0, \sigma^2) \tag{21.8}$$

(now f_i not depending on i). Estimating g using a sample $X_1, X_2, \dots, X_N$ from f is a problem in *deconvolution*. In general we might call the estimation of g in model (21.1)–(21.2) the "Bayes deconvolution problem."

An artificial example appears in Figure 21.1, where $g(\theta)$ is a mixture distribution: seven-eighths $\mathcal{N}(0, 0.5^2)$ and one-eighth uniform over the interval $[-3, 3]$. A normal sampling model $X_i \overset{\text{ind}}{\sim} \mathcal{N}(\Theta_i, 1)$ is assumed, yielding f by convolution as in (21.8). The convolution process makes f wider

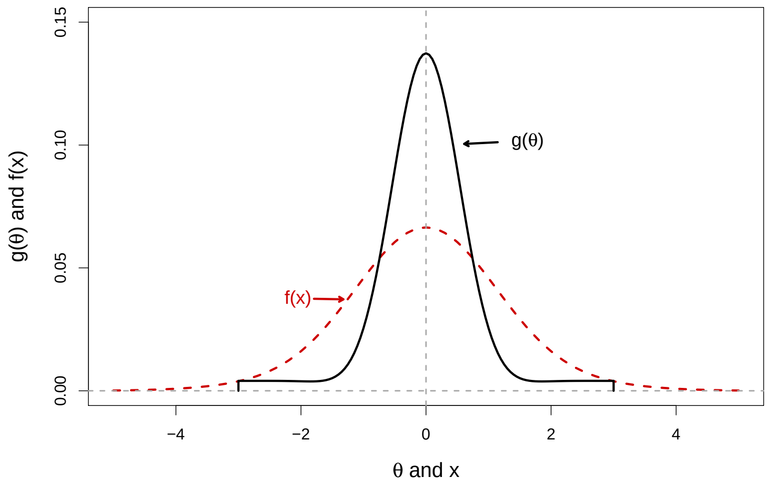


Figure 21.1 An artificial example of the Bayes deconvolution problem. The solid curve is $g(\theta)$, the prior density of Θ (21.1); the dashed curve is the density of an observation X from marginal distribution $f = g * \mathcal{N}(0,1)$ (21.8). We wish to estimate $g(\theta)$ on the basis of a random sample $X_1, X_2, \dots, X_N$ from $f(x)$.

and smoother than g, as illustrated in the figure. Having observed a random sample from f, we wish to estimate the deconvolute g, which begins to look difficult in the figure's example.

Deconvolution has a well-deserved reputation for difficulty. It is the classic ill-posed problem: because of the convolution process (21.6), large changes in $g(\theta)$ are smoothed out, often yielding only small changes in $f(x)$. Deconvolution operates in the other direction, with small changes in the estimation of f disturbingly magnified on the g scale. Nevertheless, modern computation, modern theory, and most of all modern sample sizes, together can make empirical deconvolution a practical reality.

Why would we want to estimate $g(\theta)$? In the **prostate** data example (3.28) (where Θ is called μ) we might wish to know $\Pr\{\Theta = 0\}$, the probability of a *null* gene, ones whose effect size is zero; or perhaps $\Pr\{|\Theta| \geq 2\}$, the proportion of genes that are substantially non-null. Or we might want to estimate Bayesian posterior expectations like $E\{\Theta|X = x\}$ in Figure 20.7, or posterior densities as in Figure 6.5.

Two main strategies have developed for carrying out empirical Bayes estimation: modeling on the θ scale, called *g-modeling* here, and modeling on the x scale, called *f-modeling*. We begin in the next section with g-modeling.

21.2 g-Modeling and Estimation

There has been a substantial amount of work on the asymptotic accuracy of estimates $\hat{g}(\theta)$ in the empirical Bayes model (21.1)–(21.2), most often in the normal sampling framework (21.4). The results are discouraging, with the rate of convergence of $\hat{g}(\theta)$ to $g(\theta)$ as slow as $(\log N)^{-1}$. In our terminology, much of this work has been carried out in a nonparametric g-modeling framework, allowing the unknown prior density $g(\theta)$ to be virtually anything at all. More optimistic results are possible if the g-modeling is pursued parametrically, that is, by restricting $g(\theta)$ to lie within some parametric family of possibilities.

We assume, for the sake of simpler exposition, that the space $\mathcal{T}$ of possible Θ values is finite and discrete, say

$$\mathcal{T} = \left\{\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(m)}\right\}. \tag{21.9}$$

The prior density $g(\theta)$ is now represented by a vector $\boldsymbol{g} = (g_1, g_2, \ldots, g_m)'$, with components

$$g_j = \Pr\left\{\Theta = \theta_{(j)}\right\} \qquad \text{for } j = 1, 2, \ldots, m. \tag{21.10}$$

A p-parameter exponential family (5.50) for $\boldsymbol{g}$ can be written as

$$\boldsymbol{g} = \boldsymbol{g}(\alpha) = e^{\boldsymbol{Q}\alpha - \psi(\alpha)}, \tag{21.11}$$

where the p-vector α is the natural parameter and $\boldsymbol{Q}$ is a known $m \times p$ *structure matrix*. Notation (21.11) means that the jth component of $\boldsymbol{g}(\alpha)$ is

$$g_j(\alpha) = e^{Q'_j\alpha - \psi(\alpha)}, \tag{21.12}$$

with Q'_j the jth row of $\boldsymbol{Q}$; the function $\psi(\alpha)$ is the normalizer that makes $\boldsymbol{g}(\alpha)$ sum to 1,

$$\psi(\alpha) = \log\left(\sum_{j=1}^{m} e^{Q'_j\alpha}\right). \tag{21.13}$$

In the **`nodes`** example of Figure 6.4, the set of possible Θ values was $\mathcal{T} = \{0.01, 0.02, \ldots, 0.99\}$, and $\boldsymbol{Q}$ was a fifth-degree polynomial matrix,

$$\boldsymbol{Q} = \texttt{poly}(\mathcal{T}, 5) \tag{21.14}$$

in R notation, indicating a five-parameter exponential family for $\boldsymbol{g}$, (6.38)–(6.39).

In the development that follows we will assume that the kernel $p_i(\cdot|\cdot)$ in (21.2) does not depend on i, i.e., that X_i has the same family of conditional distributions $p(X_i|\Theta_i)$ for all i, as in the Poisson and normal situations (21.3) and (21.4), but not the binomial case (21.5). And moreover we assume that the sample space $\mathcal{X}$ for the X_i observations is finite and discrete, say

$$\mathcal{X} = \left\{x_{(1)}, x_{(2)}, \dots, x_{(n)}\right\}. \tag{21.15}$$

None of this is necessary, but it simplifies the exposition.

Define

$$p_{kj} = \Pr\left\{X_i = x_{(k)}|\Theta_i = \theta_{(j)}\right\}, \tag{21.16}$$

for $k = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$, and the corresponding $n \times m$ matrix

$$\boldsymbol{P} = (p_{kj}), \tag{21.17}$$

having kth row $P_k = (p_{k1}, p_{k2}, \dots, p_{km})'$. The convolution-type formula (21.6) for the marginal density $f(x)$ now reduces to an inner product,

$$\begin{aligned} f_k(\alpha) &= \Pr_\alpha\left\{X_i = x_{(k)}\right\} = \textstyle\sum_{j=1}^m p_{kj} g_j(\alpha) \\ &= P_k' \boldsymbol{g}(\alpha). \end{aligned} \tag{21.18}$$

In fact we can write the entire marginal density $\boldsymbol{f}(\alpha) = (f_1(\alpha), f_2(\alpha), \dots, f_n(\alpha))'$ in terms of matrix multiplication,

$$\boldsymbol{f}(\alpha) = \boldsymbol{P}\boldsymbol{g}(\alpha). \tag{21.19}$$

The vector of counts $\boldsymbol{y} = (y_1, y_2, \dots, y_n)$, with

$$y_k = \#\left\{X_i = x_{(k)}\right\}, \tag{21.20}$$

is a sufficient statistic in the iid situation. It has a multinomial distribution (5.38),

$$\boldsymbol{y} \sim \text{Mult}_n(N, \boldsymbol{f}(\alpha)), \tag{21.21}$$

indicating N independent draws for a density $\boldsymbol{f}(\alpha)$ on n categories.

All of this provides a concise description of the g-modeling probability model:

$$\alpha \to \boldsymbol{g}(\alpha) = e^{\boldsymbol{Q}\alpha - \psi(\alpha)} \to \boldsymbol{f}(\alpha) = \boldsymbol{P}\boldsymbol{g}(\alpha) \to \boldsymbol{y} \sim \text{Mult}_n(N, \boldsymbol{f}(\alpha)). \tag{21.22}$$

The inferential task goes in the reverse direction,

$$\boldsymbol{y} \to \hat{\alpha} \to \boldsymbol{f}(\hat{\alpha}) \to \boldsymbol{g}(\hat{\alpha}) = e^{\boldsymbol{Q}\hat{\alpha} - \psi(\hat{\alpha})}. \tag{21.23}$$

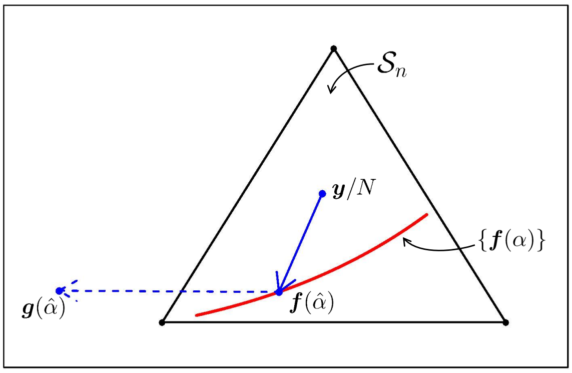

$\mathcal{S}_n$
$\boldsymbol{y}/N$
$\{\boldsymbol{f}(\alpha)\}$
$\boldsymbol{g}(\hat{\alpha})$
$\boldsymbol{f}(\hat{\alpha})$

Figure 21.2 A schematic diagram of empirical Bayes estimation, as explained in the text. $\mathcal{S}_n$ is the n-dimensional simplex, containing the p-parameter family $\mathcal{F}$ of allowable probability distributions $\boldsymbol{f}(\alpha)$. The vector of observed proportions $\boldsymbol{y}/N$ yields MLE $\boldsymbol{f}(\hat{\alpha})$, which is then deconvolved to obtain estimate $\boldsymbol{g}(\hat{\alpha})$.

A schematic diagram of the estimation process appears in Figure 21.2.

- The vector of observed proportions $\boldsymbol{y}/N$ is a point in $\mathcal{S}_n$, the simplex (5.39) of all possible probability vectors $\boldsymbol{f}$ on n categories; $\boldsymbol{y}/N$ is the usual nonparametric estimate of $\boldsymbol{f}$.
- The parametric family of allowable $\boldsymbol{f}$ vectors (21.19)

$$\mathcal{F} = \{\boldsymbol{f}(\alpha),\ \alpha \in A\}, \tag{21.24}$$

indicated by the red curve, is a curved p-dimensional surface in $\mathcal{S}_n$. Here A is the space of allowable vectors α in family (21.11).

- The nonparametric estimate $\boldsymbol{y}/N$ is "projected" down to the parametric estimate $\boldsymbol{f}(\hat{\alpha})$; if we are using MLE estimation, $\boldsymbol{f}(\hat{\alpha})$ will be the closest point in $\mathcal{F}$ to $\boldsymbol{y}/N$ measured according to a deviance metric, as in (8.35).
- Finally, $\boldsymbol{f}(\hat{\alpha})$ is mapped back to the estimate $\boldsymbol{g}(\hat{\alpha})$, by inverting mapping (21.19). (Inversion is not actually necessary with g-modeling since, having found $\hat{\alpha}$, $g(\hat{\alpha})$ is obtained directly from (21.11); the inversion step is more difficult for f-modeling, Section 21.6.)

The maximum likelihood estimation process for g-modeling is discussed in more detail in the next section, where formulas for its accuracy will be developed.

21.3 Likelihood, Regularization, and Accuracy[1]

Parametric g-modeling, as in (21.11), allows us to work in low-dimensional parametric families—just five parameters for the `nodes` example (21.14)—where classic maximum likelihood methods can be more confidently applied. Even here though, some regularization will be necessary for stable estimation, as discussed in what follows.

The g-model probability mechanism (21.22) yields a log likelihood for the multinomial vector $\boldsymbol{y}$ of counts as a function of α, say $l_{\boldsymbol{y}}(\alpha)$;

$$l_{\boldsymbol{y}}(\alpha) = \log\left(\prod_{k=1}^{n} f_k(\alpha)^{y_k}\right) = \sum_{k=1}^{n} y_k \log f_k(\alpha). \tag{21.25}$$

Its score function $\dot{l}_{\boldsymbol{y}}(\alpha)$, the vector of partial derivatives $\partial l_{\boldsymbol{y}}(\alpha)/\partial\alpha_h$ for $h = 1, 2, \ldots, p$, determines the MLE $\hat{\alpha}$ according to $\dot{l}_{\boldsymbol{y}}(\hat{\alpha}) = 0$. The $p \times p$ matrix of second derivatives $\ddot{l}_{\boldsymbol{y}}(\alpha) = (\partial^2 l_{\boldsymbol{y}}(\alpha)/\partial\alpha_h\partial\alpha_l)$ gives the *Fisher information matrix* (5.26)

$$\mathcal{I}(\alpha) = E\{-\ddot{l}_{\boldsymbol{y}}(\alpha)\}. \tag{21.26}$$

The exponential family model (21.11) yields simple expressions for $\dot{l}_{\boldsymbol{y}}(\alpha)$ and $\mathcal{I}(\alpha)$. Define

$$w_{kj} = g_j(\alpha)\left(\frac{p_{kj}}{f_k(\alpha)} - 1\right) \tag{21.27}$$

and the corresponding m-vector

$$W_k(\alpha) = (w_{k1}(\alpha), w_{k2}(\alpha), \ldots, w_{km}(\alpha))'. \tag{21.28}$$

Lemma 21.1 *The score function $\dot{l}_{\boldsymbol{y}}(\alpha)$ under model* (21.22) *is*

$$\dot{l}_{\boldsymbol{y}}(\alpha) = \boldsymbol{Q}W_+(\alpha), \qquad \textit{where } W_+(\alpha) = \sum_{k=1}^{n} W_k(\alpha)y_k \tag{21.29}$$

and $\boldsymbol{Q}$ is the $m \times p$ structure matrix in (21.11).

[1] The technical lemmas in this section are not essential to following the subsequent discussion.

Lemma 21.2 *The Fisher information matrix $\mathcal{I}(\alpha)$, evaluated at $\alpha = \hat{\alpha}$, is*

$$\mathcal{I}(\hat{\alpha}) = \boldsymbol{Q}' \left\{ \sum_{k=1}^{n} W_k(\hat{\alpha}) N f_k(\hat{\alpha}) W_k(\hat{\alpha})' \right\} \boldsymbol{Q}, \tag{21.30}$$

where $N = \sum_1^n y_k$ is the sample size in the empirical Bayes model (21.1)–(21.2).

†1 See the chapter endnotes[†] for a brief discussion of Lemmas 21.1 and 21.2. $\mathcal{I}(\hat{\alpha})^{-1}$ is the usual maximum likelihood estimate of the covariance matrix of $\hat{\alpha}$, but we will use a regularized version of the MLE that is less variable.

In the examples that follow, $\hat{\alpha}$ was found by numerical maximization.[2] Even though $\boldsymbol{g}(\alpha)$ is an exponential family, the marginal density $\boldsymbol{f}(\alpha)$ in (21.22) *is not*. As a result, some care is needed in avoiding local maxima of $l_{\boldsymbol{y}}(\alpha)$. These tend to occur at "corner" values of α, where one of its components goes to infinity. A small amount of regularization pulls $\hat{\alpha}$ away from the corners, decreasing its variance at the possible expense of increased bias.

Instead of maximizing $l_{\boldsymbol{y}}(\alpha)$ we maximize a *penalized likelihood*

$$m(\alpha) = l_{\boldsymbol{y}}(\alpha) - s(\alpha), \tag{21.31}$$

where $s(\alpha)$ is a positive penalty function. Our examples use

$$s(\alpha) = c_0 \|\alpha\| = c_0 \left(\sum_{h=1}^{p} \alpha_h^2 \right)^{1/2} \tag{21.32}$$

(with c_0 equal 1), which prevents the maximizer $\hat{\alpha}$ of $m(\alpha)$ from venturing too far into corners.

The following lemma is discussed in the chapter endnotes.

†2 **Lemma 21.3** [†]*The maximizer $\hat{\alpha}$ of $m(\alpha)$ has approximate bias vector and covariance matrix*

$$\begin{aligned} \operatorname{Bias}(\hat{\alpha}) &= -\left(\mathcal{I}(\hat{\alpha}) + \ddot{s}(\hat{\alpha})\right)^{-1} \dot{s}(\hat{\alpha}) \\ \textit{and}\ \operatorname{Var}(\hat{\alpha}) &= \left(\mathcal{I}(\hat{\alpha}) + \ddot{s}(\hat{\alpha})\right)^{-1} \mathcal{I}(\hat{\alpha}) \left(\mathcal{I}(\hat{\alpha}) + \ddot{s}(\hat{\alpha})\right)^{-1}, \end{aligned} \tag{21.33}$$

where $\mathcal{I}(\hat{\alpha})$ is given in (21.30).

With $s(\alpha) \equiv 0$ (no regularization) the bias is zero and $\operatorname{Var}(\hat{\alpha}) = \mathcal{I}(\hat{\alpha})^{-1}$,

[2] Using the nonlinear maximizer `nlm` in R.

the usual MLE approximations: including $s(\alpha)$ reduces variance while introducing bias.

For $s(\alpha) = c_0\|\alpha\|$ we calculate

$$\dot{s}(\alpha) = c_0\alpha/\|\alpha\| \quad \text{and} \quad \ddot{s}(\alpha) = \frac{c_0}{\|\alpha\|}\left(\boldsymbol{I} - \frac{\alpha\alpha'}{\|\alpha\|^2}\right), \tag{21.34}$$

with $\boldsymbol{I}$ the $p \times p$ identity matrix. Adding the penalty $s(\alpha)$ in (21.31) pulls the MLE of α toward zero and the MLE of $\boldsymbol{g}(\alpha)$ toward a flat distribution over $\mathcal{T}$. Looking at $\mathrm{Var}(\hat{\alpha})$ in (21.33), a measure of the regularization effect is

$$\mathrm{tr}(\ddot{s}(\hat{\alpha}))/\,\mathrm{tr}(\mathcal{I}(\hat{\alpha})), \tag{21.35}$$

which was never more than a few percent in our examples.

Most often we will be more interested in the accuracy of $\hat{\boldsymbol{g}} = \boldsymbol{g}(\hat{\alpha})$ than in that of $\hat{\alpha}$ itself. Letting

$$\boldsymbol{D}(\hat{\alpha}) = \mathrm{diag}(\boldsymbol{g}(\hat{\alpha})) - \boldsymbol{g}(\hat{\alpha})\boldsymbol{g}(\hat{\alpha})', \tag{21.36}$$

the $m \times p$ derivative matrix $(\partial g_j/\partial\alpha_h)$ is

$$\partial\boldsymbol{g}/\partial\alpha = \boldsymbol{D}(\alpha)\boldsymbol{Q}, \tag{21.37}$$

with $\boldsymbol{Q}$ the structure matrix in (21.11). The usual first-order delta-method calculations then give the following theorem.

Theorem 21.4 *The penalized maximum likelihood estimate $\hat{\boldsymbol{g}} = \boldsymbol{g}(\hat{\alpha})$ has estimated bias vector and covariance matrix*

$$\begin{aligned} \mathrm{Bias}(\hat{\boldsymbol{g}}) &= \boldsymbol{D}(\hat{\alpha})\boldsymbol{Q}\mathit{Bias}(\hat{\alpha}) \\ \textit{and}\ \mathrm{Var}(\hat{\boldsymbol{g}}) &= \boldsymbol{D}(\hat{\alpha})\boldsymbol{Q}\mathit{Var}(\hat{\alpha})\boldsymbol{Q}'\boldsymbol{D}(\hat{\alpha}) \end{aligned} \tag{21.38}$$

with $\mathrm{Bias}(\hat{\alpha})$ *and* $\mathrm{Var}(\hat{\alpha})$ *as in* (21.33).[3]

The many approximations going into Theorem 21.4 can be short-circuited by means of the parametric bootstrap, Section 10.4. Starting from $\hat{\alpha}$ and $\boldsymbol{f}(\hat{\alpha}) = \boldsymbol{P}\boldsymbol{g}(\hat{\alpha})$, we resample the count vector

$$\boldsymbol{y}^* \sim \mathrm{Mult}_n(N, \boldsymbol{f}(\hat{\alpha})), \tag{21.39}$$

and calculate[4] the penalized MLE $\hat{\alpha}^*$ based on $\boldsymbol{y}^*$, yielding $\hat{\boldsymbol{g}}^* = \boldsymbol{g}(\hat{\alpha}^*)$.

[3] Note that the bias treats model (21.11) as the true prior, and arises as a result of the penalization.

[4] Convergence of the `nlm` search process is speeded up by starting from $\hat{\boldsymbol{\alpha}}$.

B replications $\hat{\boldsymbol{g}}^{*1}, \hat{\boldsymbol{g}}^{*2}, \dots, \hat{\boldsymbol{g}}^{*B}$ gives bias and covariance estimates

$$\begin{aligned} \widehat{\text{Bias}} &= \hat{\boldsymbol{g}}^{*\cdot} - \hat{\boldsymbol{g}} \\ \text{and } \widehat{\text{Var}} &= \sum_{b=1}^{B} (\hat{\boldsymbol{g}}^{*b} - \hat{\boldsymbol{g}}^{*\cdot})(\hat{\boldsymbol{g}}^{*b} - \hat{\boldsymbol{g}}^{*\cdot})' \big/ (B-1), \end{aligned} \tag{21.40}$$

and $\hat{\boldsymbol{g}}^{*\cdot} = \sum_1^B \hat{\boldsymbol{g}}^{*b}/B$.

Table 21.1 *Comparison of delta method* (21.38) *and bootstrap* (21.40) *standard errors and biases for the* `nodes` *study estimate of* $\boldsymbol{g}$ *in Figure 6.4. All columns except the first multiplied by 100.*

		Standard Error		Bias	
θ	$g(\theta)$	Delta	Boot	Delta	Boot
.01	12.048	.887	.967	−.518	−.592
.12	1.045	.131	.139	.056	.071
.23	.381	.058	.065	.025	.033
.34	.779	.096	.095	−.011	−.013
.45	1.119	.121	.117	−.040	−.049
.56	.534	.102	.100	.019	.027
.67	.264	.047	.051	.023	.027
.78	.224	.056	.053	.018	.020
.89	.321	.054	.048	.013	.009
.99	.576	.164	.169	−.008	.008

Table 21.1 compares the delta method of Theorem 20.4 with the parametric bootstrap ($B = 1000$ replications) for the surgical nodes example of Section 6.3. Both the standard errors—square roots of the diagonal elements of $\text{Var}(\hat{\boldsymbol{g}})$—and biases are well approximated by the delta method formulas (21.38). The delta method also performed reasonably well on the two examples of the next section.

It did less well on the artificial example of Figure 21.1, where

$$g(\theta) = \frac{1}{8}\frac{I_{[-3,3](\theta)}}{6} + \frac{7}{8}\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{\theta^2}{\sigma^2}} \qquad (\sigma = 0.5) \tag{21.41}$$

(1/8 uniform on $[-3,3]$ and 7/8 $\mathcal{N}(0, 0.5^2)$). The vertical bars in Figure 21.3 indicate ± one standard error obtained from the parametric bootstrap, taking $\mathcal{T} = \{-3, -2.8, \dots, 3\}$ for the sample space of Θ, and assuming a natural spline model in (21.11) with five degrees of freedom,

$$\boldsymbol{g}(\alpha) = e^{\boldsymbol{Q}\alpha - \psi(\alpha)}, \qquad \boldsymbol{Q} = \texttt{ns}(\mathcal{T}\texttt{,df=5}). \tag{21.42}$$

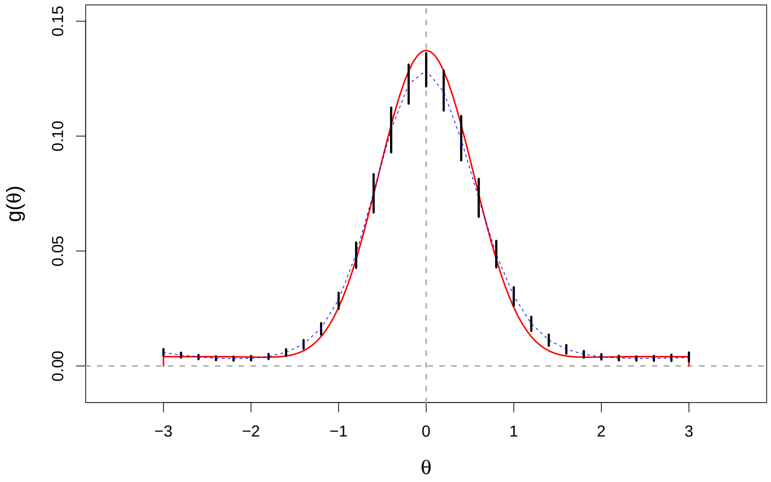


Figure 21.3 The red curve is $g(\theta)$ for the artificial example of Figure 21.1. Vertical bars are ± one standard error for g-model estimate $\boldsymbol{g}(\hat{\alpha})$; specifications (21.41)–(21.42), sample size $N = 1000$ observations $X_i \sim \mathcal{N}(\Theta_i, 1)$, using parametric bootstrap (21.40), $B = 500$. The light dashed line follows bootstrap means $\hat{g}_j^*$. Some definitional bias is apparent.

The sampling model was $X_i \sim \mathcal{N}(\Theta_i, 1)$ for $i = 1, 2, \ldots, N = 1000$. In this case the delta method standard errors were about 25% too small.

The light dashed curve in Figure 21.3 traces $\bar{g}(\theta)$, the average of the $B = 500$ bootstrap replications $\boldsymbol{g}^{*b}$. There is noticeable bias, compared with $g(\theta)$. The reason is simple: the exponential family (21.42) for $\boldsymbol{g}(\alpha)$ does not include $g(\theta)$ (21.41). In fact, $\bar{g}(\theta)$ is (nearly) the closest member of the exponential family to $g(\theta)$. This kind of *definitional bias* is a disadvantage of parametric g-modeling.

Our g-modeling examples, and those of the next section, bring together a variety of themes from modern statistical practice: classical maximum likelihood theory, exponential family modeling, regularization, bootstrap methods, large data sets of parallel structure, indirect evidence, and a combination of Bayesian and frequentist thinking, all of this enabled by massive computer power. Taken together they paint an attractive picture of the range of inferential methodology in the twenty-first century.

21.4 Two Examples

We now reconsider two previous data sets from a g-modeling point of view. the first is the artificial microarray-type example (20.24) comprising N =10,000 independent observations

$$z_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, 1), \qquad i = 1, 2, \dots, N = 10{,}000, \tag{21.43}$$

with

$$\mu_i \sim \begin{cases} 0 & \text{for } i = 1, 2, \dots, 9000 \\ \mathcal{N}(-3, 1) & \text{for } i = 9001, \dots, 10{,}000. \end{cases} \tag{21.44}$$

Figure 20.3 displays the points (z_i, μ_i) for $i = 9001, \dots, 10,000$, illustrating the Bayes posterior 95% conditional intervals (20.26),

$$\mu_i \in (z_i - 3)/2 \pm 1.96\Big/\sqrt{2}. \tag{21.45}$$

These required knowing the Bayes prior distribution $\mu_i \sim \mathcal{N}(-3, 1)$. We would like to recover intervals (21.45) using just the observed data z_i, $i = 1, 2, \dots, 10,000$, without knowledge of the prior.

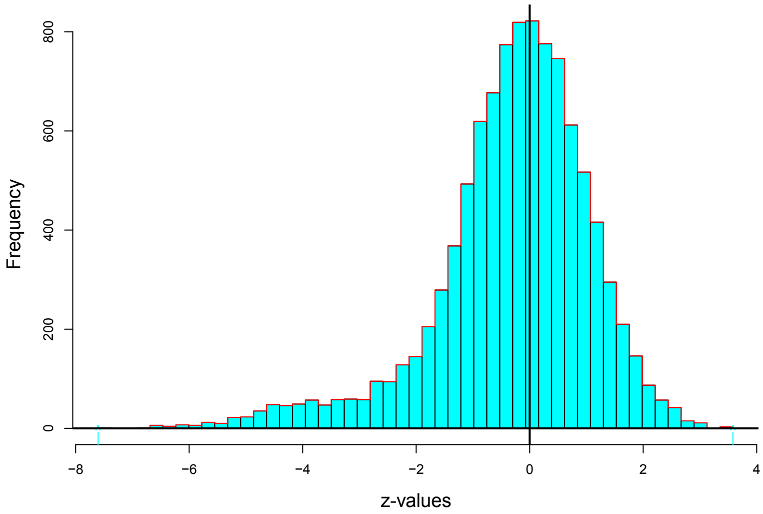


Figure 21.4 Histogram of observed sample of $N = 10{,}000$ values z_i from simulations (21.43)–(21.44).

A histogram of the 10,000 z-values is shown in Figure 21.4; g-modeling (21.9)–(21.11) was applied to them (now with μ playing the role of "Θ"

and z being "x"), taking $\mathcal{T} = (-6, -5.75, \ldots, 3)$. $\boldsymbol{Q}$ was composed of a delta function at $\mu = 0$ and a fifth-degree polynomial basis for the nonzero μ, again a family of spike-and-slab priors. The penalized MLE $\hat{\boldsymbol{g}}$ (21.31), (21.32), $c_0 = 1$, estimated the probability of $\mu = 0$ as

$$\hat{g}(0) = 0.891 \pm 0.006 \tag{21.46}$$

(using (21.38), which also provided bias estimate 0.001).

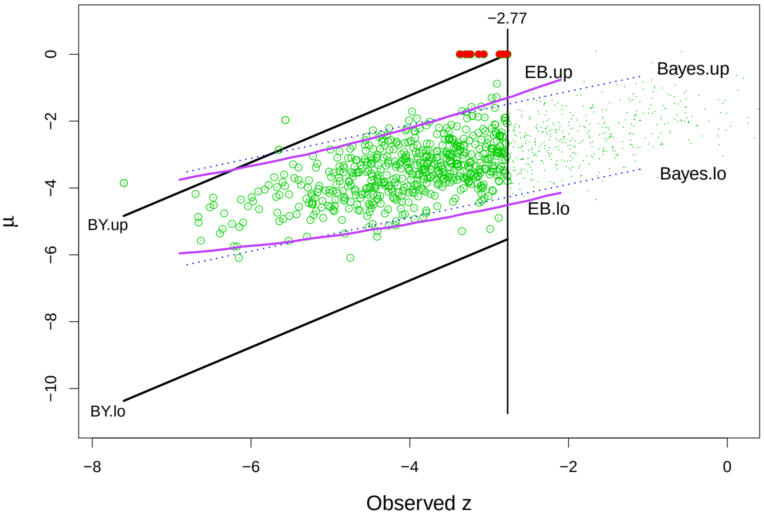


Figure 21.5 Purple curves show g-modeling estimates of conditional 95% credible intervals for μ given z in artificial microarray example (21.43)–(21.44). They are a close match to the actual Bayes intervals, dotted lines; cf. Figure 20.3.

The estimated posterior density of μ given z is

$$\hat{g}(\mu|z) = c_z \hat{g}(\mu)\phi(z - \mu), \tag{21.47}$$

$\phi(\cdot)$ the standard normal density and c_z the constant required for $\hat{g}(\mu|z)$ to integrate to 1. Let $q^{(\alpha)}(z)$ denote the αth quantile of $\hat{g}(\mu|z)$. The purple curves in Figure 21.5 trace the estimated 95% credible intervals

$$\left(q^{(.025)}(z), q^{(.975)}(z)\right). \tag{21.48}$$

They are a close match to the actual credible intervals (21.45).

The solid black curve in Figure 21.6 shows $\hat{g}(\mu)$ for $\mu \neq 0$ (the "slab" portion of the estimated prior). As an estimate of the actual slab density

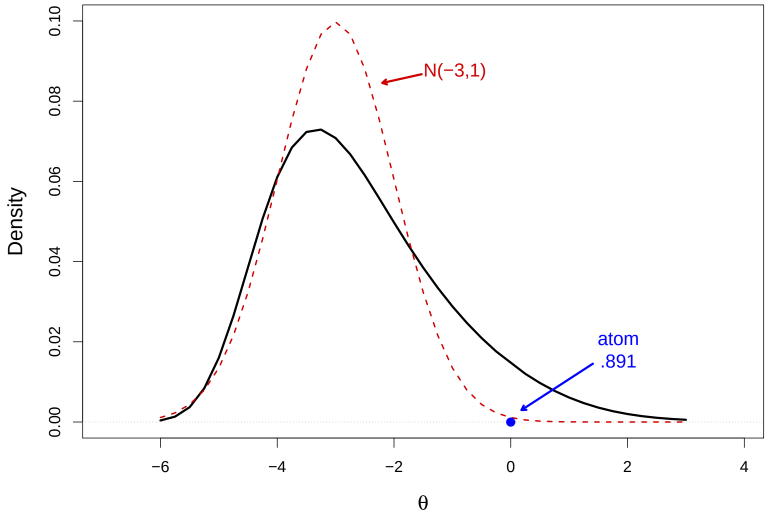


Figure 21.6 The heavy black curve is the g-modeling estimate of $g(\mu)$ for $\mu \neq 0$ in the artificial microarray example, suppressing the atom at zero, $\hat{g}(0) = 0.891$. It is only a rough estimate of the actual nonzero density $\mathcal{N}(-3, 1)$.

$\mu \sim \mathcal{N}(-3, 1)$ it is only roughly accurate, but apparently still accurate enough to yield the reasonably good posterior intervals seen in Figure 21.5. The fundamental impediment to deconvolution—that large changes in $g(\theta)$ produce only small changes in $f(x)$—can sometimes operate in the statistician's favor, when only a rough knowledge of g suffices for applied purposes.

Our second example concerns the `prostate` study data, last seen in Figure 15.1: $n = 102$ men, 52 cancer patients and 50 normal controls, each have had their genetic activities measured on a microarray of $N = 6033$ genes; gene$_i$ yields a test statistic z_i comparing patients with controls,

$$z_i \sim \mathcal{N}(\mu_i, \sigma_0^2), \tag{21.49}$$

with μ_i the gene's effect size. (Here we will take the variance σ_0^2 as a parameter to be estimated, rather than assuming $\sigma_0^2 = 1$.) What is the prior density $g(\mu)$ for the effects?

The local false-discovery rate program `locfdr`, Section 15.5, was applied to the 6033 z_i values, as shown in Figure 21.7. `Locfdr` is an "f-modeling" method, where probability models are proposed directly for

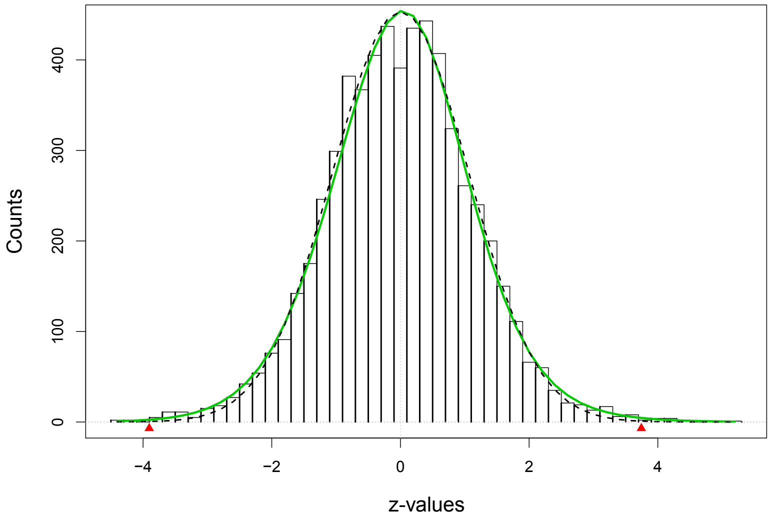


Figure 21.7 The green curve is a six-parameter Poisson regression estimate fit to counts of the observed z_i values for the **prostate** data. The dashed curve is the empirical null (15.48), $z_i \sim \mathcal{N}(0.00, 1.06^2)$. The f-modeling program **locfdr** estimated null probability $\Pr\{\mu = 0\} = 0.984$. Genes with z-values lying beyond the red triangles have estimated fdr values less than 0.20.

the marginal density $f(\cdot)$ rather than for the prior density $g(\cdot)$; see Section (21.6). Here we can compare **locfdr**'s results with those from g-modeling. The former gave[5]

$$\left(\hat{\delta}_0, \hat{\sigma}_0, \hat{\pi}_0\right) = (0.00, 1.06, 0.984) \tag{21.50}$$

in the notation of (15.50); that is, it estimated the null distribution as $\mu \sim \mathcal{N}(0, 1.06^2)$, with probability $\hat{\pi}_0 = 0.984$ of a gene being null ($\mu = 0$).

Only 22 genes were estimated to have local fdr values less than 0.20, the 9 with $z_i \leq -3.71$ and the 12 with $z_i \geq 3.81$. (These are more pessimistic results than in Figure 15.5, where we used the theoretical null $\mathcal{N}(0, 1)$ rather than the empirical null $\mathcal{N}(0, 1.06^2)$.)

The g-modeling approach (21.11) was applied to the **prostate** study data, assuming $z_i \sim \mathcal{N}(\mu_i, \sigma_0^2)$, $\sigma_0 = 1.06$ as suggested by (21.50). The

[5] Using a six-parameter Poisson regression fit to the z_i values, of the type employed in Section 10.4.

structure matrix $\boldsymbol{Q}$ in (21.11) had a delta function at $\mu = 0$ and a five-parameter natural spline basis for $\mu \neq 0$; $\mathcal{T} = (-3.6, -3.4, \ldots, 3.6)$ for the discretized Θ space (21.9). This gave a penalized MLE $\hat{\boldsymbol{g}}$ having null probability

$$\hat{g}(0) = 0.946 \pm 0.011. \tag{21.51}$$

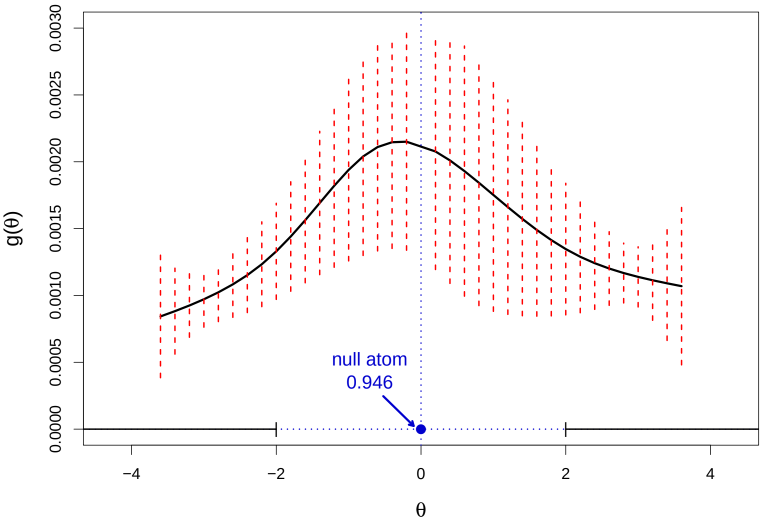


Figure 21.8 The g-modeling estimate for the non-null density $\hat{g}(\mu)$, $\mu \neq 0$, for the `prostate` study data, also indicating the null atom $\hat{g}(0) = 0.946$. About 2% of the genes are estimated to have effect sizes $|\mu_i| \geq 2$. The red bars show ± one standard error as computed from Theorem 21.4 (page 429).

The non-null distribution, $\hat{g}(\mu)$ for $\mu \neq 0$, appears in Figure 21.8, where it is seen to be modestly unimodal around $\mu = 0$. Dashed red bars indicate ± one standard error for the $\hat{g}(\theta_{(j)})$ estimates obtained from Theorem 21.4 (page 429). The accuracy is not very good. It is better for larger regions of the Θ space, for example

$$\widehat{\Pr}\{|\theta| \geq 2\} = 0.020 \pm 0.0014. \tag{21.52}$$

Here g-modeling estimated less prior null probability, 0.946 compared with 0.984 from f-modeling, but then attributed much of the non-null probability to small values of $|\mu_i|$.

Taking (21.52) literally suggests 121 ($= 0.020 \cdot 6033$) genes with true

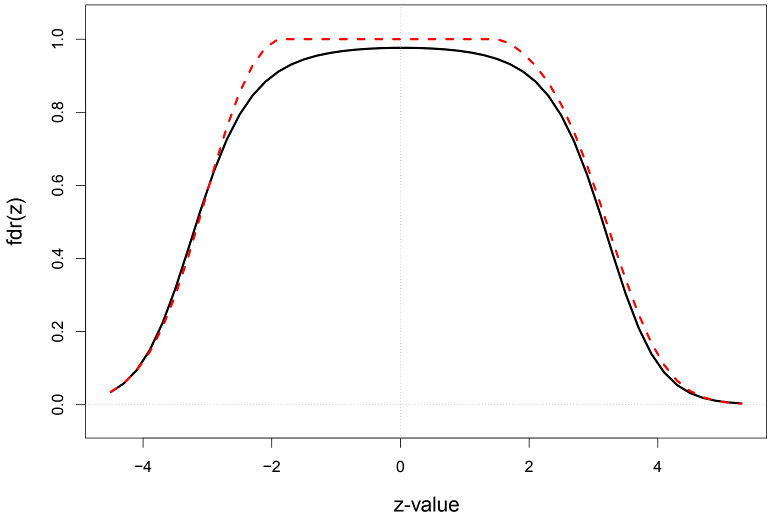


Figure 21.9 The black curve is the empirical Bayes estimated false-discovery rate $\widehat{\Pr}\{\mu = 0|z\}$ from g-modeling. For large values of $|z|$ it nearly matches the **`locfdr`** f-modeling estimate $\mathrm{fdr}(z)$, red curve.

effect sizes $|\mu_i| \geq 2$. That doesn't mean we can say with certainty *which* 121. Figure 21.9 compares the g-modeling empirical Bayes false-discovery rate

$$\widehat{\Pr}\{\mu = 0|z\} = c_z \hat{g}(0)\phi\left(\frac{z-\mu}{\hat{\sigma}_0}\right), \tag{21.53}$$

as in (21.47), with the f-modeling estimate $\widehat{\mathrm{fdr}}(z)$ produced by **`locfdr`**. Where it counts, in the tails, they are nearly the same.

21.5 Generalized Linear Mixed Models

The g-modeling theory can be extended to the situation where each observation X_i is accompanied by an observed vector of covariates c_i, say of dimension d. We return to the generalized linear model setup of Section 8.2, where each X_i has a one-parameter exponential family density indexed by its own natural parameter λ_i,

$$f_{\lambda_i}(X_i) = \exp\{\lambda_i X_i - \gamma(\lambda_i)\} f_0(X_i) \tag{21.54}$$

in notation (8.20).

Our key assumption is that each λ_i is the sum of a deterministic component, depending on the covariates c_i, and a random term Θ_i,

$$\lambda_i = \Theta_i + c_i'\beta. \tag{21.55}$$

Here Θ_i is an unobserved realization from $\boldsymbol{g}(\alpha) = \exp\{\boldsymbol{Q}\alpha - \psi(\alpha)\}$ (21.11) and β is an unknown d-dimensional parameter. If $\beta = 0$ then (21.55) is a g-model as before,[6] while if all the $\Theta_i = 0$ then it is a standard GLM (8.20)–(8.22). Taken together, (21.55) represents a *generalized linear mixed model* (GLMM). The likelihood and accuracy calculations of Section 21.3 extend to GLMMs, as referenced in the endnotes, but here we will only discuss a GLMM analysis of the `nodes` study of Section 6.3.

In addition to n_i the number of `nodes` removed and X_i the number found `positive` (6.33), a vector of four covariates

$$c_i = (\texttt{age}_i,\ \texttt{sex}_i,\ \texttt{smoke}_i,\ \texttt{prog}_i) \tag{21.56}$$

was observed for each patient: a standardized version of `age` in years; `sex` being 0 for female or 1 for male; `smoke` being 0 for no or 1 for yes to long-term smoking; and `prog` being a post-operative prognosis score with large values more favorable.

GLMM model (21.55) was applied to the `nodes` data. Now λ_i was the logit $\log[\pi_i/(1-\pi_i)]$, where

$$X_i \sim \text{Bi}(n_i, \pi_i) \tag{21.57}$$

as in Table 8.4, i.e., π_i is the probability that any one node from patient i is positive. To make the correspondence with the analysis in Section 6.3 exact, we used a variant of (21.55)

$$\lambda_i = \text{logit}(\Theta_i) + c_i'\beta. \tag{21.58}$$

Now with $\beta = 0$, Θ_i is exactly the binomial probability π_i for the ith case. Maximum likelihood estimates were calculated for α in (21.11)—with $\mathcal{T} = (0.01, 0.02, \ldots, 0.99)$ and $\boldsymbol{Q} = \texttt{poly}(\mathcal{T}\texttt{,5})$ (21.14)—and β in (21.58). The MLE prior $\boldsymbol{g}(\hat{\alpha})$ was almost the same as that estimated without covariates in Figure 6.4.

Table 21.2 shows the MLE values $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$, their standard errors (from a parametric bootstrap simulation), and the z-values $\hat{\beta}_k/\widehat{\text{se}}_k$. `Sex` looks like it has a significant effect, with males tending toward larger values of π_i, that is, a greater number of positive nodes. The big effect though is `prog`, larger values of `prog` indicating smaller values of π_i.

[6] Here the setup is more specific; f is exponential family, and Θ_i is on the natural-parameter scale.

Table 21.2 *Maximum likelihood estimates* $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4)$ *for GLMM analysis of the* `nodes` *data, and standard errors from a parametric bootstrap simulation; large values of* $\texttt{prog}_i$ *predict low values of* π_i.

	`age`	`sex`	`smoke`	`prog`
MLE	−.078	.192	.089	−.698
Boot st err	.066	.070	.063	.077
***z*-value**	**−1.18**	**2.74**	**1.41**	**9.07**

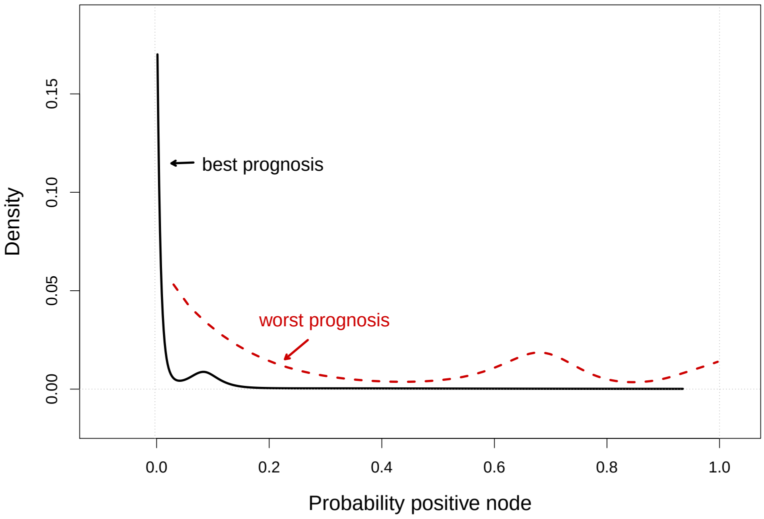


Figure 21.10 Distribution of π_i, individual probabilities of a positive node, for best and worst levels of factor `prog`; from GLMM analysis of `nodes` data.

Figure 21.10 displays the distribution of $\pi_i = 1/[1+\exp(-\lambda_i)]$ implied by the GLMM model for the best and worst values of prog (setting `age`, `sex`, and `smoke` to their average values and letting Θ have distribution $g(\hat{\alpha})$). The implied distribution is concentrated near $\pi = 0$ for the best-level `prog`, while it is roughly uniform over [0, 1] for the worst level.

The random effects we have called Θ_i are sometimes called *frailties*: a composite of unmeasured individual factors lumped together as an index of disease susceptibility. Taken together, Figures 6.4 and 21.10 show substantial frailty and covariate effects both at work in the `nodes` data. In the language of Section 6.1, we have amassed "indirect evidence" for each patient, using both Bayesian and frequentist methods.

21.6 Deconvolution and f-Modeling

Empirical Bayes applications have traditionally been dominated by f-modeling—not the g-modeling approach of the previous sections—where probability models for the marginal density $f(x)$, usually exponential families, are fit directly to the observed sample $X_1, X_2, \dots, X_N$. We have seen several examples: Robbins' estimator in Table 6.1 (particularly the bottom line), `locfdr`'s Poisson regression estimates in Figures 15.6 and 21.7, and Tweedie's estimate in Figure 20.7.

Both the advantages and the disadvantages of f-modeling can be seen in the inferential diagram of Figure 21.2. For f-modeling the red curve now can represent an exponential family $\{\boldsymbol{f}(\alpha)\}$, whose concave log likelihood function greatly simplifies the calculation of $\boldsymbol{f}(\hat{\alpha})$ from $\boldsymbol{y}/N$. This comes at a price: the deconvolution step, from $\boldsymbol{f}(\hat{\alpha})$ to a prior distribution $\boldsymbol{g}(\hat{\alpha})$, is problematical, as discussed below.

This is only a problem if we want to know $\boldsymbol{g}$. The traditional applications of f-modeling apply to problems where the desired answer can be phrased directly in terms of $\boldsymbol{f}$. This was the case for Robbins' formula (6.5), the local false-discovery rate (15.38), and Tweedie's formula (20.37).

Nevertheless, f-modeling methodology for the estimation of the prior $g(\theta)$ does exist, an elegant example being the *Fourier method* described next. A function $f(x)$ and its Fourier transform $\phi(t)$ are related by

$$\phi(t) = \int_{-\infty}^{\infty} f(x)e^{itx}\,dx \quad \text{and} \quad f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi(t)e^{-itx}\,dt. \tag{21.59}$$

For the *normal case* where $X_i = \Theta_i + Z_i$ with $Z_i \sim \mathcal{N}(0,1)$, the Fourier transform of $f(x)$ is a multiple of that for $g(\theta)$,

$$\phi_f(t) = \phi_g(t)e^{-t^2/2}, \tag{21.60}$$

so, on the transform scale, estimating g from f amounts to removing the factor $\exp(t^2/2)$.

The Fourier method begins with the empirical density $\bar{f}(x)$ that puts probability $1/N$ on each observed value X_i, and then proceeds in three steps.

†3 1 $\bar{f}(x)$ is smoothed using the "sinc" kernel,†

$$\tilde{f}(x) = \frac{1}{N\lambda} \sum_{i=1}^{N} \operatorname{sinc}\left(\frac{X_i - x}{\lambda}\right), \qquad \operatorname{sinc}(x) = \frac{\sin(x)}{x}. \tag{21.61}$$

2 The Fourier transform of $\tilde{f}$, say $\tilde{\phi}(t)$, is calculated.

3 Finally, $\hat{g}(\theta)$ is taken to be the inverse Fourier transform of $\tilde{\phi}(t)e^{t^2/2}$, this last step eliminating the unwanted factor $e^{-t^2/2}$ in (21.60).

A pleasantly surprising aspect of the Fourier method is that $\hat{g}(\theta)$ can be expressed directly as a kernel estimate,

$$\hat{g}(\theta) = \frac{1}{N} \sum_{i=1}^{N} k_\lambda(X_i - \theta) = \int_{-\infty}^{\infty} k_\lambda(x - \theta)\bar{f}(x)\, dx, \tag{21.62}$$

where the kernel $k_\lambda(\cdot)$ is

$$k_\lambda(x) = \frac{1}{\pi} \int_0^{1/\lambda} e^{t^2/2} \cos(tx)\, dt. \tag{21.63}$$

Large values of λ smooth $\bar{f}(x)$ more in (21.61), reducing the variance of $\hat{g}(\theta)$ at the expense of increased bias.

Despite its compelling rationale, there are two drawbacks to the Fourier method. First of all, it applies only to situations $X_i = \Theta_i + Z_i$ where X_i is Θ_i plus iid noise. More seriously, the bias/variance trade-off in the choice of λ can be quite unfavorable.

This is illustrated in Figure 21.11 for the artificial example of Figure 21.1. The black curve is the standard deviation of the g-modeling estimate of $g(\theta)$ for θ in $[-3, 3]$, under specifications (21.41)–(21.42). The red curve graphs the standard deviation of the f-modeling estimate (21.62), with $\lambda = 1/3$, a value that produced roughly the same amount of bias as the g-modeling estimate (seen in Figure 21.3). The ratio of red to black standard deviations averages more than 20 over the range of θ.

This comparison is at least partly unfair: g-modeling is parametric while the Fourier method is almost nonparametric in its assumptions about $f(x)$ or $g(\theta)$. It can be greatly improved by beginning the three-step algorithm with a parametric estimate $\hat{f}(x)$ rather than $\bar{f}(x)$. The blue dotted curve in Figure 21.11 does this with $\hat{f}(x)$ a Poisson regression on the data $X_1, X_2, \dots, X_N$—as in Figure 10.5 but here using a natural spline basis `ns(df=5)` —giving the estimate

$$\hat{g}(\theta) = \int_{-\infty}^{\infty} k_\lambda(x - \theta)\hat{f}(x)\, dx. \tag{21.64}$$

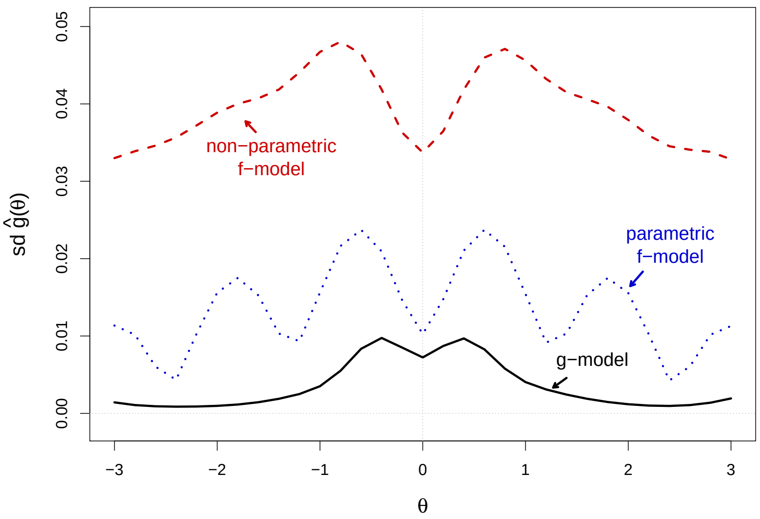


Figure 21.11 Standard deviations of estimated prior density $\hat{g}(\theta)$ for the artificial example of Figure 21.1, based on $N = 1000$ observations $X_i \sim \mathcal{N}(\Theta_i, 1)$; black curve using g-modeling under specifications (21.41)–(21.42); red curve nonparametric f-modeling (21.62), $\lambda = 1/3$; blue curve parametric f-modeling (21.64), with $\hat{f}(x)$ estimated from Poisson regression with a structure matrix having five degrees of freedom.

We see a substantial decrease in standard deviation, though still not attaining g-modeling rates.

As commented before, the great majority of empirical Bayes applications have been of the Robbins/fdr/Tweedie variety, where f-modeling is the natural choice. g-modeling comes into its own for situations like the **nodes** data analysis of Figures 6.4 and 6.5, where we really want an estimate of the prior $g(\theta)$. Twenty-first-century science is producing more such data sets, an impetus for the further development of g-modeling strategies.

Table 21.3 concerns the g-modeling estimation of $E_x = E\{\Theta|X = x\}$,

$$E_x = \int_{\mathcal{T}} \theta g(\theta) f_\theta(x)\, d\theta \Bigg/ \int_{\mathcal{T}} g(\theta) f_\theta(x)\, d\theta \qquad (21.65)$$

for the artificial example, under the same specifications as in Figure 21.11. Samples of size $N = 1000$ of $X_i \sim \mathcal{N}(\Theta_i, 1)$ were drawn from model (21.41)–(21.42), yielding MLE $\hat{g}(\theta)$ and estimates $\hat{E}_x$ for x between -4

Table 21.3 *Standard deviation of $\hat{E}\{\Theta|x\}$ computed from parametric bootstrap simulations of $\hat{g}(\theta)$. The g-modeling is as in Figure 21.11, with $N = 1000$ observations $X_i \sim \mathcal{N}(\Theta_i, 1)$ from the artificial example for each simulation. The column "info" is the implied empirical Bayes information for estimating $E\{\Theta|x\}$ obtained from one "other" observation X_i.*

x	$E\{\Theta\|x\}$	sd$(\hat{E})$	info
−3.5	−2.00	.10	.11
−2.5	−1.06	.10	.11
−1.5	−.44	.05	.47
−.5	−.13	.03	.89
.5	.13	.04	.80
1.5	.44	.05	.44
2.5	1.06	.10	.10
3.5	2.00	.16	.04

and 4. One thousand such estimates $\hat{E}_x$ were generated, averaging almost exactly E_x, with standard deviations as shown. Accuracy is reasonably good, the coefficient of variation $\text{sd}(\hat{E}_x)/E_x$ being about 0.05 for large values of $|x|$. (Estimate (21.65) is a favorable case: results are worse for other conditional estimates† such as $E\{\Theta^2|X = x\}$.) †4

Theorem 21.4 (page 429) implies that, for large values of the sample size N, the variance of $\hat{E}_x$ decreases as $1/N$, say

$$\text{var}\left\{\hat{E}_x\right\} \doteq c_x/N. \tag{21.66}$$

By analogy with the Fisher information bound (5.27), we can define the *empirical Bayes information* for estimating E_x in one observation to be

$$i_x = 1\Big/\left(N \cdot \text{var}\left\{\hat{E}_x\right\}\right), \tag{21.67}$$

so that $\text{var}\{\hat{E}_x\} \doteq i_x^{-1}/N$.

Empirical Bayes inference leads us directly into the world of indirect evidence, *learning from the experience of others* as in Sections 6.4 and 7.4. So, if $X_i = 2.5$, each " other" observation X_j provides 0.10 units of information for learning $E\{\Theta|X_i = 2.5\}$ (compared with the usual Fisher information value $\mathcal{I} = 1$ for the direct estimation of Θ_i from X_i). This is a favorable case, as mentioned, and i_x is often much smaller. The main point, perhaps, is that assuming a Bayes prior is not a casual matter, and can amount to the assumption of an enormous amount of relevant *other* information.

21.7 Notes and Details

Empirical Bayes and James–Stein estimation, Chapters 6 and 7, exploded onto the statistics scene almost simultaneously in the 1950s. They represented a genuinely new branch of statistical inference, unlike the computer-based extensions of classical methodology reviewed in previous chapters. Their development as practical tools has been comparatively slow. The pace has quickened in the twenty-first century, with false-discovery rates, Chapter 15, as a major step forward. A practical empirical Bayes methodology for use beyond traditional f-modeling venues such as fdr is the goal of the g-modeling approach.

†1 [p. 428] *Lemmas 21.1 and 21.2.* The derivations of Lemmas 21.1 and 21.2 are straightforward but somewhat involved exercises in differential calculus, carried out in Remark B of Efron (2016). Here we will present just a sample of the calculations. From (21.18), the gradient vector $\dot{f}_k(\alpha) = (\partial f_k(\alpha)/\partial\alpha_l)$ with respect to α is

$$\dot{f}_k(\alpha) = \dot{\boldsymbol{g}}(\alpha)'\boldsymbol{P}_k, \tag{21.68}$$

where $\dot{\boldsymbol{g}}(\alpha)$ is the $m \times p$ derivative matrix

$$\dot{\boldsymbol{g}}(\alpha) = (\partial g_j(\alpha)/\partial\alpha_l) = \boldsymbol{DQ}, \tag{21.69}$$

with $\boldsymbol{D}$ as in (21.36), the last equality following, after some work, by differentiation of $\log \boldsymbol{g}(\alpha) = \boldsymbol{Q}\alpha - \phi(\alpha)$.

Let $l_k = \log f_k$ (now suppressing α from the notation). The gradient with respect to α of l_k is then

$$\dot{l}_k = \dot{f}_k/f_k = \boldsymbol{Q}'\boldsymbol{DP}_k/f_k. \tag{21.70}$$

The vector $\boldsymbol{DP}_k/f_k$ has components

$$(g_j p_{kj} - g_j f_k)/f_k = w_{kj} \tag{21.71}$$

(21.27), using $\boldsymbol{g}'P_k = f_k$. This gives $\dot{l}_k = \boldsymbol{Q}'W_k(\alpha)$ (21.28). Adding up the independent score functions $\dot{l}_k$ over the full sample yields the overall score $\dot{l}_{\boldsymbol{y}}(\alpha) = \boldsymbol{Q}'\sum_1^n y_k W_k(\alpha)$, which is Lemma 21.1.

†2 [p. 428] *Lemma 2.* The penalized MLE $\hat{\alpha}$ satisfies

$$\boldsymbol{O} = \dot{m}(\hat{\alpha}) \doteq \dot{m}(\alpha_0) + \ddot{m}(\alpha_0)(\hat{\alpha} - \alpha_0), \tag{21.72}$$

where α_0 is the true value of α, or

$$\hat{\alpha} - \alpha_0 \doteq (-\ddot{m}(\alpha_0))^{-1}\dot{m}(\alpha_0)\left(-\ddot{l}_y(\alpha_0) + \ddot{s}(\alpha_0)\right)^{-1}\left(\dot{l}_y(\alpha_0) - \dot{s}(\alpha_0)\right). \tag{21.73}$$

Standard MLE theory shows that the random variable $\dot{l}_y(\alpha_0)$ has mean 0 and covariance Fisher information matrix $\mathcal{I}(\alpha_0)$, while $-\ddot{l}_y(\alpha_0)$ asymptotically approximates $\mathcal{I}(\alpha_0)$. Substituting in (21.73),

$$\hat{\alpha} - \alpha_0 \doteq (\mathcal{I}(\alpha_0) + \ddot{s}(\alpha_0))^{-1} Z, \tag{21.74}$$

where Z has mean $-\dot{s}(\alpha_0)$ and covariance $\mathcal{I}(\alpha_0)$. This gives Bias$(\hat{\alpha})$ and Var$(\hat{\alpha})$ as in Lemma 2. Note that the bias is with respect to a *true* parametric model (21.11), and is a consequence of the penalization.

†3 [p. 440] *The sinc kernel.* The Fourier transform $\phi_s(t)$ of the scaled sinc function $s(x) = \sin(x/\lambda)/(\pi x)$ is the indicator of the interval $[-1/\lambda, 1/\lambda]$, while that of $\bar{f}(x)$ is $(1/N)\sum_1^N \exp(itX_j)$. Formula (21.61) is the convolution $\bar{f} * s$, so $\tilde{f}$ has the product transform

$$\phi_{\tilde{f}}(t) = \left[\frac{1}{N}\sum_{j=1}^{N} e^{itX_j}\right] I_{[-1/\lambda,1/\lambda]}(t). \tag{21.75}$$

The effect of the sinc convolution is to censor the high-frequency (large t) components of $\bar{f}$ or $\phi_{\bar{f}}$. Larger λ yields more censoring. Formula (21.63) has upper limits $1/\lambda$ because of $\phi_s(t)$. All of this is due to Stefanski and Carroll (1990). Smoothers other than the sinc kernel have been suggested in the literature, but without substantial improvements on deconvolution performance.

†4 [p. 443] *Conditional expectation* (21.65). Efron (2014b) considers estimating $E\{\Theta^2|X = x\}$ and other such conditional expectations, both for f-modeling and for g-modeling. $E\{\Theta|X = x\}$ is by far the easiest case, as might be expected from the simple form of Tweedie's estimate (20.37).

Epilogue

Something important changed in the world of statistics in the new millennium. Twentieth-century statistics, even after the heated expansion of its late period, could still be contained within the classic Bayesian–frequentist–Fisherian inferential triangle (Figure 14.1). This is not so in the twenty-first century. Some of the topics discussed in Part III—false-discovery rates, post-selection inference, empirical Bayes modeling, the lasso—fit within the triangle but others seem to have escaped, heading south from the frequentist corner, perhaps in the direction of computer science.

The escapees were the large-scale prediction algorithms of Chapters 17–19: neural nets, deep learning, boosting, random forests, and support-vector machines. Notably missing from their development were parametric probability models, the building blocks of classical inference. Prediction algorithms are the media stars of the big-data era. It is worth asking why they have taken center stage and what it means for the future of the statistics discipline.

The *why* is easy enough: prediction is commercially valuable. Modern equipment has enabled the collection of mountainous data troves, which the "data miners" can then burrow into, extracting valuable information. Moreover, prediction is the simplest use of regression theory (Section 8.4). It can be carried out successfully without probability models, perhaps with the assistance of nonparametric analysis tools such as cross-validation, permutations, and the bootstrap.

A great amount of ingenuity and experimentation has gone into the development of modern prediction algorithms, with statisticians playing an important but not dominant role.[1] There is no shortage of impressive success stories. In the absence of optimality criteria, either frequentist or Bayesian, the prediction community grades algorithmic excellence on per-

[1] All papers mentioned in this section have their complete references in the bibliography. Footnotes will identify papers not fully specified in the text.

formance within a catalog of often-visited examples such as the spam and digits data sets of Chapters 17 and 18.[2] Meanwhile, "traditional statistics" —probability models, optimality criteria, Bayes priors, asymptotics—has continued successfully along on a parallel track. Pessimistically or optimistically, one can consider this as a bipolar disorder of the field or as a healthy duality that is bound to improve both branches. There are historical and intellectual arguments favoring the optimists' side of the story.

The first thing to say is that the current situation is not entirely unprecedented. By the end of the nineteenth century there was available an impressive inventory of statistical methods—Bayes' theorem, least squares, correlation, regression, the multivariate normal distribution—but these existed more as individual algorithms than as a unified discipline. Statistics as a distinct intellectual enterprise was not yet well-formed.

A small but crucial step forward was taken in 1914 when the astrophysicist Arthur Eddington[3] claimed that mean absolute deviation was superior to the familiar root mean square estimate for the standard deviation from a normal sample. Fisher in 1919 showed that this was wrong, and moreover, in a clear mathematical sense, the root mean square was the *best possible estimate*. Eddington conceded the point while Fisher went on to develop the theory of sufficiency and optimal estimation.[4]

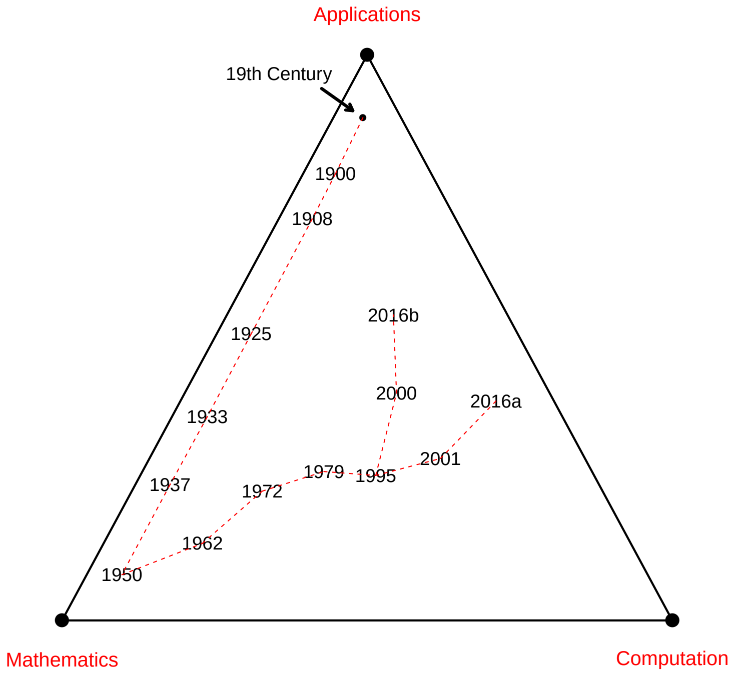
"Optimal" is the key word here. Before Fisher, statisticians didn't really understand estimation. The same can be said now about prediction. Despite their impressive performance on a raft of test problems, it might still be possible to do much better than neural nets, deep learning, random forests, and boosting—or perhaps they are coming close to some as-yet unknown theoretical minimum.

It is the job of statistical inference to connect "dangling algorithms" to the central core of well-understood methodology. The connection process is already underway. Section 17.4 showed how **`Adaboost`**, the original machine learning algorithm, could be restated as a close cousin of logistic regression. Purely empirical approaches like the Common Task Framework are ultimately unsatisfying without some form of principled justification. Our optimistic scenario has the big-data/data-science prediction world rejoining the mainstream of statistical inference, to the benefit of both branches.

[2] This empirical approach to optimality is sometimes codified as the *Common Task Framework* (Liberman, 2015 and Donoho, 2015).

[3] Eddington became world-famous for his 1919 empirical verification of Einstein's relativity theory.

[4] See Stigler (2006) for the full story.



Development of the statistics discipline since the end of the nineteenth century, as discussed in the text.

Whether or not we can predict the future of statistics, we can at least examine the past to see how we've gotten where we are. The next figure does so in terms of a new triangle diagram, this time with the poles labeled *Applications*, *Mathematics*, and *Computation*. "Mathematics" here is shorthand for the mathematical/logical justification of statistical methods. "Computation" stands for the empirical/numerical approach.

Statistics is a branch of applied mathematics, and is ultimately judged by how well it serves the world of applications. Mathematical logic, *à la* Fisher, has been the traditional vehicle for the development and understanding of statistical methods. Computation, slow and difficult before the 1950s, was only a bottleneck, but now has emerged as a competitor to (or perhaps a seven-league boots enabler of) mathematical analysis. At any one time the discipline's energy and excitement is directed unequally toward the three poles. The figure attempts, in admittedly crude fashion, to track the changes in direction over the past 100+ years.

The tour begins at the end of the nineteenth century. Mathematicians of the caliber of Gauss and Laplace had contributed to the available methodology, but the subsequent development was almost entirely applications-driven. Quetelet[5] was especially influential, applying the Gauss–Laplace formulation to census data and his “Average Man.” A modern reader will search almost in vain for any mathematical symbology in nineteenth-century statistics journals.

1900

Karl Pearson’s chi-square paper was a bold step into the new century, applying a new mathematical tool, matrix theory, in the service of statistical methodology. He and Weldon went on to found *Biometrika* in 1901, the first recognizably modern statistics journal. Pearson’s paper, and *Biometrika*, launched the statistics discipline on a fifty-year march toward the mathematics pole of the triangle.

1908

Student’s t statistic was a crucial first result in small-sample “exact” inference, and a major influence on Fisher’s thinking.

1925

Fisher’s great estimation paper—a more coherent version of its 1922 predecessor. It introduced a host of fundamental ideas, including sufficiency, efficiency, Fisher information, maximum likelihood theory, and the notion of optimal estimation. Optimality is a mark of maturity in mathematics, making 1925 the year statistical inference went from a collection of ingenious techniques to a coherent discipline.

1933

This represents Neyman and Pearson’s paper on optimal hypothesis testing. A logical completion of Fisher’s program, it nevertheless aroused his strong antipathy. This was partly personal, but also reflected Fisher’s concern that mathematization was squeezing intuitive correctness out of statistical thinking (Section 4.2).

1937

Neyman’s seminal paper on confidence intervals. His sophisticated mathematical treatment of statistical inference was a harbinger of decision theory.

[5] Adolphe Quetelet was a tireless organizer, helping found the Royal Statistical Society in 1834, with the American Statistical Association following in 1839.

1950

The publication of Wald's *Statistical Decision Functions*. Decision theory completed the full mathematization of statistical inference. This date can also stand for Savage's and de Finetti's decision-theoretic formulation of Bayesian inference. We are as far as possible from the Applications corner of the triangle now, and it is fair to describe the 1950s as a nadir of the influence of the statistics discipline on scientific applications.

1962

The arrival of electronic computation in the mid 1950s began the process of stirring statistics out of its inward-gazing preoccupation with mathematical structure. Tukey's paper "The future of data analysis" argued for a more application- and computation-oriented discipline. Mosteller and Tukey later suggested changing the field's name to *data analysis*, a prescient hint of today's *data science*.

1972

Cox's proportional hazards paper. Immensely useful in its own right, it signaled a growing interest in biostatistical applications and particularly survival analysis, which was to assert its scientific importance in the analysis of AIDS epidemic data.

1979

The bootstrap, and later the widespread use of MCMC: electronic computation used for the extension of classic statistical inference.

1995

This stands for false-discovery rates and, a year later, the lasso.[6] Both are computer-intensive algorithms, firmly rooted in the ethos of statistical inference. They lead, however, in different directions, as indicated by the split in the diagram.

2000

Microarray technology inspires enormous interest in large-scale inference, both in theory and as applied to the analysis of microbiological data.

[6] Benjamini and Hochberg (1995) and Tibshirani (1996).

2001

Random forests; it joins boosting[7] and the resurgence of neural nets in the ranks of *machine learning* prediction algorithms.

2016a

Data science: a more popular successor to Tukey and Mosteller's "data analysis," at one extreme it seems to represent a statistics discipline without parametric probability models or formal inference. The Data Science Association defines a practitioner as one who "... uses scientific methods to liberate and create meaning from raw data." In practice the emphasis is on the algorithmic processing of large data sets for the extraction of useful information, with the prediction algorithms as exemplars.

2016b

This represents the traditional line of statistical thinking, of the kind that could be located within Figure 14.1, but now energized with a renewed focus on applications. Of particular applied interest are biology and genetics. Genome-wide association studies (GWAS) show a different face of big data. Prediction is important here,[8] but not sufficient for the scientific understanding of disease.

A cohesive inferential theory was forged in the first half of the twentieth century, but unity came at the price of an inwardly focused discipline, of reduced practical utility. In the century's second half, electronic computation unleashed a vast expansion of useful—and much used—statistical methodology. Expansion accelerated at the turn of the millennium, further increasing the reach of statistical thinking, but now at the price of intellectual cohesion.

It is tempting but risky to speculate on the future of statistics. What will the Mathematics–Applications–Computation diagram look like, say 25 years from now? The appetite for statistical analysis seems to be always increasing, both from science and from society in general. Data science has blossomed in response, but so has the traditional wing of the field. The data-analytic initiatives represented in the diagram by 2016a and 2016b are in actuality not isolated points but the centers of overlapping distributions.

[7] Breiman (1996) for random forests, Freund and Schapire (1997) for boosting.

[8] "Personalized medicine" in which an individual's genome predicts his or her optimal treatment has attracted grail-like attention.

A hopeful scenario for the future is one of an increasing overlap that puts data science on a solid footing while leading to a broader general formulation of statistical inference.

References

Abu-Mostafa, Y. 1995. Hints. *Neural Computation*, **7**, 639–671.

Achanta, R., and Hastie, T. 2015. *Telugu OCR Framework using Deep Learning*. Tech. rept. Statistics Department, Stanford University.

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. Pages 267–281 of: *Second International Symposium on Information Theory (Tsahkadsor, 1971)*. Akadémiai Kiadó, Budapest.

Anderson, T. W. 2003. *An Introduction to Multivariate Statistical Analysis*. Third edn. Wiley Series in Probability and Statistics. Wiley-Interscience.

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I. J., Bergeron, A., Bouchard, N., and Bengio, Y. 2012. *Theano: new features and speed improvements*. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

Becker, R., Chambers, J., and Wilks, A. 1988. *The New S Language: A Programming Environment for Data Analysis and Graphics*. Pacific Grove, CA: Wadsworth and Brooks/Cole.

Bellhouse, D. R. 2004. The Reverend Thomas Bayes, FRS: A biography to celebrate the tercentenary of his birth. *Statist. Sci.*, **19**(1), 3–43. With comments and a rejoinder by the author.

Bengio, Y., Courville, A., and Vincent, P. 2013. Representation learning: a review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**(8), 1798–1828.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, **57**(1), 289–300.

Benjamini, Y., and Yekutieli, D. 2005. False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.*, **100**(469), 71–93.

Berger, J. O. 2006. The case for objective Bayesian analysis. *Bayesian Anal.*, **1**(3), 385–402 (electronic).

Berger, J. O., and Pericchi, L. R. 1996. The intrinsic Bayes factor for model selection and prediction. *J. Amer. Statist. Assoc.*, **91**(433), 109–122.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. 2010 (June). Theano: a CPU and GPU math expression compiler. In: *Proceedings of the Python for Scientific Computing Conference (SciPy)*.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. 2013. Valid post-selection inference. *Ann. Statist.*, **41**(2), 802–837.

Berkson, J. 1944. Application of the logistic function to bio-assay. *J. Amer. Statist. Assoc.*, **39**(227), 357–365.

Bernardo, J. M. 1979. Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B*, **41**(2), 113–147. With discussion.

Birch, M. W. 1964. The detection of partial association. I. The 2×2 case. *J. Roy. Statist. Soc. Ser. B*, **26**(2), 313–324.

Bishop, C. 1995. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Boos, D. D., and Serfling, R. J. 1980. A note on differentials and the CLT and LIL for statistical functions, with application to M-estimates. *Ann. Statist.*, **8**(3), 618–624.

Boser, B., Guyon, I., and Vapnik, V. 1992. A training algorithm for optimal margin classifiers. In: *Proceedings of COLT II*.

Breiman, L. 1996. Bagging predictors. *Mach. Learn.*, **24**(2), 123–140.

Breiman, L. 1998. Arcing classifiers (with discussion). *Annals of Statistics*, **26**, 801–849.

Breiman, L. 2001. Random forests. *Machine Learning*, **45**, 5–32.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. 1984. *Classification and Regression Trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software.

Carlin, B. P., and Louis, T. A. 1996. *Bayes and Empirical Bayes Methods for Data Analysis*. Monographs on Statistics and Applied Probability, vol. 69. Chapman & Hall.

Carlin, B. P., and Louis, T. A. 2000. *Bayes and Empirical Bayes Methods for Data Analysis*. 2 edn. Texts in Statistical Science. Chapman & Hall/CRC.

Chambers, J. M., and Hastie, T. J. (eds). 1993. *Statistical Models in S*. Chapman & Hall Computer Science Series. Chapman & Hall.

Cleveland, W. S. 1981. LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *Amer. Statist.*, **35**(1), 54.

Cox, D. R. 1958. The regression analysis of binary sequences. *J. Roy. Statist. Soc. Ser. B*, **20**, 215–242.

Cox, D. R. 1970. *The Analysis of Binary Data*. Methuen's Monographs on Applied Probability and Statistics. Methuen & Co.

Cox, D. R. 1972. Regression models and life-tables. *J. Roy. Statist. Soc. Ser. B*, **34**(2), 187–220.

Cox, D. R. 1975. Partial likelihood. *Biometrika*, **62**(2), 269–276.

Cox, D. R., and Hinkley, D. V. 1974. *Theoretical Statistics*. Chapman & Hall.

Cox, D. R., and Reid, N. 1987. Parameter orthogonality and approximate conditional inference. *J. Roy. Statist. Soc. Ser. B*, **49**(1), 1–39. With a discussion.

Crowley, J. 1974. Asymptotic normality of a new nonparametric statistic for use in organ transplant studies. *J. Amer. Statist. Assoc.*, **69**(348), 1006–1011.

de Finetti, B. 1972. *Probability, Induction and Statistics. The Art of Guessing*. John Wiley & Sons, London-New York-Sydney.

Dembo, A., Cover, T. M., and Thomas, J. A. 1991. Information-theoretic inequalities. *IEEE Trans. Inform. Theory*, **37**(6), 1501–1518.

Dempster, A. P., Laird, N. M., and Rubin, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, **39**(1), 1–38.

Diaconis, P., and Ylvisaker, D. 1979. Conjugate priors for exponential families. *Ann. Statist.*, **7**(2), 269–281.

DiCiccio, T., and Efron, B. 1992. More accurate confidence intervals in exponential families. *Biometrika*, **79**(2), 231–245.

Donoho, D. L. 2015. 50 years of data science. *R-bloggers*. `www.r-bloggers.com/50-years-of-data-science-by-david-donoho/`.

Edwards, A. W. F. 1992. *Likelihood*. Expanded edn. Johns Hopkins University Press. Revised reprint of the 1972 original.

Efron, B. 1967. The two sample problem with censored data. Pages 831–853 of: *Proc. 5th Berkeley Symp. Math. Statist. and Prob., Vol. 4*. University of California Press.

Efron, B. 1975. Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.*, **3**(6), 1189–1242. With discussion and a reply by the author.

Efron, B. 1977. The efficiency of Cox's likelihood function for censored data. *J. Amer. Statist. Assoc.*, **72**(359), 557–565.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, **7**(1), 1–26.

Efron, B. 1982. *The Jackknife, the Bootstrap and Other Resampling Plans*. CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 38. Society for Industrial and Applied Mathematics (SIAM).

Efron, B. 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.*, **78**(382), 316–331.

Efron, B. 1985. Bootstrap confidence intervals for a class of parametric problems. *Biometrika*, **72**(1), 45–58.

Efron, B. 1986. How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.*, **81**(394), 461–470.

Efron, B. 1987. Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.*, **82**(397), 171–200. With comments and a rejoinder by the author.

Efron, B. 1988. Logistic regression, survival analysis, and the Kaplan–Meier curve. *J. Amer. Statist. Assoc.*, **83**(402), 414–425.

Efron, B. 1993. Bayes and likelihood calculations from confidence intervals. *Biometrika*, **80**(1), 3–26.

Efron, B. 1998. R. A. Fisher in the 21st Century (invited paper presented at the 1996 R. A. Fisher Lecture). *Statist. Sci.*, **13**(2), 95–122. With comments and a rejoinder by the author.

Efron, B. 2004. The estimation of prediction error: Covariance penalties and cross-validation. *J. Amer. Statist. Assoc.*, **99**(467), 619–642. With comments and a rejoinder by the author.

Efron, B. 2010. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs, vol. 1. Cambridge University Press.

Efron, B. 2011. Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.*, **106**(496), 1602–1614.

Efron, B. 2014a. Estimation and accuracy after model selection. *J. Amer. Statist. Assoc.*, **109**(507), 991–1007.

Efron, B. 2014b. Two modeling strategies for empirical Bayes estimation. *Statist. Sci.*, **29**(2), 285–301.

Efron, B. 2015. Frequentist accuracy of Bayesian estimates. *J. Roy. Statist. Soc. Ser. B*, **77**(3), 617–646.

Efron, B. 2016. Empirical Bayes deconvolution estimates. *Biometrika*, **103**(1), 1–20.

Efron, B., and Feldman, D. 1991. Compliance as an explanatory variable in clinical trials. *J. Amer. Statist. Assoc.*, **86**(413), 9–17.

Efron, B., and Gous, A. 2001. Scales of evidence for model selection: Fisher versus Jeffreys. Pages 208–256 of: *Model Selection*. IMS Lecture Notes Monograph Series, vol. 38. Beachwood, OH: Institute of Mathematics and Statististics. With discussion and a rejoinder by the authors.

Efron, B., and Hinkley, D. V. 1978. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, **65**(3), 457–487. With comments and a reply by the authors.

Efron, B., and Morris, C. 1972. Limiting the risk of Bayes and empirical Bayes estimators. II. The empirical Bayes case. *J. Amer. Statist. Assoc.*, **67**, 130–139.

Efron, B., and Morris, C. 1977. Stein's paradox in statistics. *Scientific American*, **236**(5), 119–127.

Efron, B., and Petrosian, V. 1992. A simple test of independence for truncated data with applications to redshift surveys. *Astrophys. J.*, **399**(Nov), 345–352.

Efron, B., and Stein, C. 1981. The jackknife estimate of variance. *Ann. Statist.*, **9**(3), 586–596.

Efron, B., and Thisted, R. 1976. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, **63**(3), 435–447.

Efron, B., and Tibshirani, R. 1993. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability, vol. 57. Chapman & Hall.

Efron, B., and Tibshirani, R. 1997. Improvements on cross-validation: The .632+ bootstrap method. *J. Amer. Statist. Assoc.*, **92**(438), 548–560.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. 2004. Least angle regression. *Annals of Statistics*, **32**(2), 407–499. (with discussion, and a rejoinder by the authors).

Finney, D. J. 1947. The estimation from individual records of the relationship between dose and quantal response. *Biometrika*, **34**(3/4), 320–334.

Fisher, R. A. 1915. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, **10**(4), 507–521.

Fisher, R. A. 1925. Theory of statistical estimation. *Math. Proc. Cambridge Phil. Soc.*, **22**(7), 700–725.

Fisher, R. A. 1930. Inverse probability. *Math. Proc. Cambridge Phil. Soc.*, **26**(10), 528–535.

Fisher, R. A., Corbet, A., and Williams, C. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.*, **12**, 42–58.

Fithian, W., Sun, D., and Taylor, J. 2014. Optimal inference after model selection. *ArXiv e-prints*, Oct.

Freund, Y., and Schapire, R. 1996. Experiments with a new boosting algorithm. Pages 148–156 of: *Machine Learning: Proceedings of the Thirteenth International Conference*. Morgan Kauffman, San Francisco.

Freund, Y., and Schapire, R. 1997. A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences*, **55**, 119–139.

Friedman, J. 2001. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, **29**(5), 1189–1232.

Friedman, J., and Popescu, B. 2005. *Predictive Learning via Rule Ensembles*. Tech. rept. Stanford University.

Friedman, J., Hastie, T., and Tibshirani, R. 2000. Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics*, **28**, 337–307.

Friedman, J., Hastie, T., and Tibshirani, R. 2009. *glmnet: Lasso and elastic-net regularized generalized linear models*. R package version 1.1-4.

Friedman, J., Hastie, T., and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33**(1), 1–22.

Geisser, S. 1974. A predictive approach to the random effect model. *Biometrika*, **61**, 101–107.

Gerber, M., and Chopin, N. 2015. Sequential quasi Monte Carlo. *J. Roy. Statist. Soc. B*, **77**(3), 509–580. with discussion, doi: 10.1111/rssb.12104.

Gholami, S., Janson, L., Worhunsky, D. J., Tran, T. B., Squires, Malcolm, I., Jin, L. X., Spolverato, G., Votanopoulos, K. I., Schmidt, C., Weber, S. M., Bloomston, M., Cho, C. S., Levine, E. A., Fields, R. C., Pawlik, T. M., Maithel, S. K., Efron, B., Norton, J. A., and Poultsides, G. A. 2015. Number of lymph nodes removed and survival after gastric cancer resection: An analysis from the US Gastric Cancer Collaborative. *J. Amer. Coll. Surg.*, **221**(2), 291–299.

Good, I., and Toulmin, G. 1956. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, **43**, 45–63.

Hall, P. 1988. Theoretical comparison of bootstrap confidence intervals. *Ann. Statist.*, **16**(3), 927–985. with discussion and a reply by the author.

Hampel, F. R. 1974. The influence curve and its role in robust estimation. *J. Amer. Statist. Assoc.*, **69**, 383–393.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. 1986. *Robust Statistics: The approach based on influence functions*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons.

Harford, T. 2014. Big data: A big mistake? *Significance*, **11**(5), 14–19.

Hastie, T., and Loader, C. 1993. Local regression: automatic kernel carpentry (with discussion). *Statistical Science*, **8**, 120–143.

Hastie, T., and Tibshirani, R. 1990. *Generalized Additive Models*. Chapman and Hall.

Hastie, T., and Tibshirani, R. 2004. Efficient quadratic regularization for expression arrays. *Biostatistics*, **5**(3), 329–340.

Hastie, T., Tibshirani, R., and Friedman, J. 2009. *The Elements of Statistical Learning. Data mining, Inference, and Prediction*. Second edn. Springer Series in Statistics. Springer.

Hastie, T., Tibshirani, R., and Wainwright, M. 2015. *Statistical Learning with Sparsity: the Lasso and Generalizations*. Chapman and Hall, CRC Press.

Hoeffding, W. 1952. The large-sample power of tests based on permutations of observations. *Ann. Math. Statist.*, **23**, 169–192.

Hoeffding, W. 1965. Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.*, **36**(2), 369–408.

Hoerl, A. E., and Kennard, R. W. 1970. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**(1), 55–67.

Huber, P. J. 1964. Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101.

Jaeckel, L. A. 1972. Estimating regression coefficients by minimizing the dispersion of the residuals. *Ann. Math. Statist.*, **43**, 1449–1458.

James, W., and Stein, C. 1961. Estimation with quadratic loss. Pages 361–379 of: *Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. I. University of California Press.

Jansen, L., Fithian, W., and Hastie, T. 2015. Effective degrees of freedom: a flawed metaphor. *Biometrika*, **102**(2), 479–485.

Javanmard, A., and Montanari, A. 2014. Confidence intervals and hypothesis testing for high-dimensional regression. *J. of Machine Learning Res.*, **15**, 2869–2909.

Jaynes, E. 1968. Prior probabilities. *IEEE Trans. Syst. Sci. Cybernet.*, **4**(3), 227–241.

Jeffreys, H. 1961. *Theory of Probability*. Third ed. Clarendon Press.

Johnson, N. L., and Kotz, S. 1969. *Distributions in Statistics: Discrete Distributions*. Houghton Mifflin Co.

Johnson, N. L., and Kotz, S. 1970a. *Distributions in Statistics. Continuous Univariate Distributions. 1*. Houghton Mifflin Co.

Johnson, N. L., and Kotz, S. 1970b. *Distributions in Statistics. Continuous Univariate Distributions. 2*. Houghton Mifflin Co.

Johnson, N. L., and Kotz, S. 1972. *Distributions in Statistics: Continuous Multivariate Distributions*. John Wiley & Sons.

Kaplan, E. L., and Meier, P. 1958. Nonparametric estimation from incomplete observations. *J. Amer. Statist. Assoc.*, **53**(282), 457–481.

Kass, R. E., and Raftery, A. E. 1995. Bayes factors. *J. Amer. Statist. Assoc.*, **90**(430), 773–795.

Kass, R. E., and Wasserman, L. 1996. The selection of prior distributions by formal rules. *J. Amer. Statist. Assoc.*, **91**(435), 1343–1370.

Kuffner, R., Zach, N., Norel, R., Hawe, J., Schoenfeld, D., Wang, L., Li, G., Fang, L., Mackey, L., Hardiman, O., Cudkowicz, M., Sherman, A., Ertaylan, G., Grosse-Wentrup, M., Hothorn, T., van Ligtenberg, J., Macke, J. H., Meyer, T., Scholkopf, B., Tran, L., Vaughan, R., Stolovitzky, G., and Leitner, M. L. 2015. Crowdsourced analysis of clinical trial data to predict amyotrophic lateral sclerosis progression. *Nat Biotech*, **33**(1), 51–57.

LeCun, Y., and Cortes, C. 2010. *MNIST Handwritten Digit Database*. http://yann.lecun.com/exdb/mnist/.

LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. *Nature*, **521**(7553), 436–444.

Lee, J., Sun, D., Sun, Y., and Taylor, J. 2016. Exact post-selection inference, with application to the Lasso. *Annals of Statistics*, **44**(3), 907–927.

Lehmann, E. L. 1983. *Theory of Point Estimation*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons.

Leslie, C., Eskin, E., Cohen, A., Weston, J., and Noble, W. S. 2003. Mismatch string kernels for discriminative pretein classification. *Bioinformatics*, **1**, 1–10.

Liaw, A., and Wiener, M. 2002. Classification and regression by randomForest. *R News*, **2**(3), 18–22.

Liberman, M. 2015 (April). *"Reproducible Research and the Common Task Method"*. Simons Foundation Frontiers of Data Science Lecture, April 1, 2015; video available.

Lockhart, R., Taylor, J., Tibshirani, R., and Tibshirani, R. 2014. A significance test for the lasso. *Annals of Statistics*, **42**(2), 413–468. With discussion and a rejoinder by the authors.

Lynden-Bell, D. 1971. A method for allowing for known observational selection in small samples applied to 3CR quasars. *Mon. Not. Roy. Astron. Soc.*, **155**(1), 95–18.

Mallows, C. L. 1973. Some comments on C_p. *Technometrics*, **15**(4), 661–675.

Mantel, N., and Haenszel, W. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. *J. Natl. Cancer Inst.*, **22**(4), 719–748.

Mardia, K. V., Kent, J. T., and Bibby, J. M. 1979. *Multivariate Analysis*. Academic Press.

McCullagh, P., and Nelder, J. 1983. *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall.

McCullagh, P., and Nelder, J. 1989. *Generalized Linear Models*. Second edn. Monographs on Statistics and Applied Probability. Chapman & Hall.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. 1953. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, **21**(6), 1087–1092.

Miller, Jr, R. G. 1964. A trustworthy jackknife. *Ann. Math. Statist*, **35**, 1594–1605.

Miller, Jr, R. G. 1981. *Simultaneous Statistical Inference*. Second edn. Springer Series in Statistics. New York: Springer-Verlag.

Nesterov, Y. 2013. Gradient methods for minimizing composite functions. *Mathematical Programming*, **140**(1), 125–161.

Neyman, J. 1937. Outline of a theory of statistical estimation based on the classical theory of probability. *Phil. Trans. Roy. Soc.*, **236**(767), 333–380.

Neyman, J. 1977. Frequentist probability and frequentist statistics. *Synthese*, **36**(1), 97–131.

Neyman, J., and Pearson, E. S. 1933. On the problem of the most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc. A*, **231**(694-706), 289–337.

Ng, A. 2015. *Neural Networks*. `http://deeplearning.stanford.edu/wiki/index.php/Neural_Networks`. Lecture notes.

Ngiam, J., Chen, Z., Chia, D., Koh, P. W., Le, Q. V., and Ng, A. 2010. Tiled convolutional neural networks. Pages 1279–1287 of: Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A. (eds), *Advances in Neural Information Processing Systems 23*. Curran Associates, Inc.

O'Hagan, A. 1995. Fractional Bayes factors for model comparison. *J. Roy. Statist. Soc. Ser. B*, **57**(1), 99–138. With discussion and a reply by the author.

Park, T., and Casella, G. 2008. The Bayesian lasso. *J. Amer. Statist. Assoc.*, **103**(482), 681–686.

Pearson, K. 1900. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag.*, **50**(302), 157–175.

Pritchard, J., Stephens, M., and Donnelly, P. 2000. Inference of Population Structure using Multilocus Genotype Data. *Genetics*, **155**(June), 945–959.

Quenouille, M. H. 1956. Notes on bias in estimation. *Biometrika*, **43**, 353–360.

R Core Team. 2015. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ridgeway, G. 2005. *Generalized boosted models: A guide to the gbm package*. Available online.

Ridgeway, G., and MacDonald, J. M. 2009. Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *J. Amer. Statist. Assoc.*, **104**(486), 661–668.

Ripley, B. D. 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press.

Robbins, H. 1956. An empirical Bayes approach to statistics. Pages 157–163 of: *Proc. 3rd Berkeley Symposium on Mathematical Statistics and Probability*, vol. I. University of California Press.

Rosset, S., Zhu, J., and Hastie, T. 2004. Margin maximizing loss functions. In: Thrun, S., Saul, L., and Schölkopf, B. (eds), *Advances in Neural Information Processing Systems 16*. MIT Press.

Rubin, D. B. 1981. The Bayesian bootstrap. *Ann. Statist.*, **9**(1), 130–134.

Savage, L. J. 1954. *The Foundations of Statistics*. John Wiley & Sons; Chapman & Hill.

Schapire, R. 1990. The strength of weak learnability. *Machine Learning*, **5**(2), 197–227.

Schapire, R., and Freund, Y. 2012. *Boosting: Foundations and Algorithms*. MIT Press.

Scheffé, H. 1953. A method for judging all contrasts in the analysis of variance. *Biometrika*, **40**(1-2), 87–110.

Schölkopf, B., and Smola, A. 2001. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. MIT Press.

Schwarz, G. 1978. Estimating the dimension of a model. *Ann. Statist.*, **6**(2), 461–464.

Senn, S. 2008. A note concerning a selection "paradox" of Dawid's. *Amer. Statist.*, **62**(3), 206–210.

Soric, B. 1989. Statistical "discoveries" and effect-size estimation. *J. Amer. Statist. Assoc.*, **84**(406), 608–610.

Spevack, M. 1968. *A Complete and Systematic Concordance to the Works of Shakespeare*. Vol. 1–6. Georg Olms Verlag.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. of Machine Learning Res.*, **15**, 1929–1958.

Stefanski, L., and Carroll, R. J. 1990. Deconvoluting kernel density estimators. *Statistics*, **21**(2), 169–184.

Stein, C. 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Pages 197–206 of: *Proc. 3rd Berkeley Symposium on Mathematical Statististics and Probability*, vol. I. University of California Press.

Stein, C. 1981. Estimation of the mean of a multivariate normal distribution. *Ann. Statist.*, **9**(6), 1135–1151.

Stein, C. 1985. On the coverage probability of confidence sets based on a prior distribution. Pages 485–514 of: *Sequential Methods in Statistics*. Banach Center Publication, vol. 16. PWN, Warsaw.

Stigler, S. M. 2006. How Ronald Fisher became a mathematical statistician. *Math. Sci. Hum. Math. Soc. Sci.*, **176**(176), 23–30.

Stone, M. 1974. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. B*, **36**, 111–147. With discussion and a reply by the author.

Storey, J. D., Taylor, J., and Siegmund, D. 2004. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *J. Roy. Statist. Soc. B*, **66**(1), 187–205.

Tanner, M. A., and Wong, W. H. 1987. The calculation of posterior distributions by data augmentation. *J. Amer. Statist. Assoc.*, **82**(398), 528–550. With discussion and a reply by the authors.

Taylor, J., Loftus, J., and Tibshirani, R. 2015. Tests in adaptive regression via the Kac-Rice formula. *Annals of Statistics*, **44**(2), 743–770.

Thisted, R., and Efron, B. 1987. Did Shakespeare write a newly-discovered poem? *Biometrika*, **74**(3), 445–455.

Tibshirani, R. 1989. Noninformative priors for one parameter of many. *Biometrika*, **76**(3), 604–608.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. B*, **58**(1), 267–288.

Tibshirani, R. 2006. A simple method for assessing sample sizes in microarray experiments. *BMC Bioinformatics*, **7**(Mar), 106.

Tibshirani, R., Bien, J., Friedman, J., Hastie, T., Simon, N., Taylor, J., and Tibshirani, R. 2012. Strong rules for discarding predictors in lasso-type problems. *J. Roy. Statist. Soc. B*, **74**.

Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J., and Reid, S. 2016. *selectiveInference: Tools for Post-Selection Inference*. R package version 1.1.3.

Tukey, J. W. 1958. "Bias and confidence in not-quite large samples" in Abstracts of Papers. *Ann. Math. Statist.*, **29**(2), 614.

Tukey, J. W. 1960. A survey of sampling from contaminated distributions. Pages 448–485 of: *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (I. Olkin, et. al, ed.). Stanford University Press.

Tukey, J. W. 1962. The future of data analysis. *Ann. Math. Statist.*, **33**, 1–67.

Tukey, J. W. 1977. *Exploratory Data Analysis*. Behavioral Science Series. Addison-Wesley.

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, **42**(3), 1166–1202.

Vapnik, V. 1996. *The Nature of Statistical Learning Theory*. Springer.

Wager, S., Wang, S., and Liang, P. S. 2013. Dropout training as adaptive regularization. Pages 351–359 of: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds), *Advances in Neural Information Processing Systems 26*. Curran Associates, Inc.

Wager, S., Hastie, T., and Efron, B. 2014. Confidence intervals for random forests: the jacknife and the infintesimal jacknife. *J. of Machine Learning Res.*, **15**, 1625–1651.

Wahba, G. 1990. *Spline Models for Observational Data*. SIAM.

Wahba, G., Lin, Y., and Zhang, H. 2000. GACV for support vector machines. Pages 297–311 of: Smola, A., Bartlett, P., Schölkopf, B., and Schuurmans, D. (eds), *Advances in Large Margin Classifiers*. MIT Press.

Wald, A. 1950. *Statistical Decision Functions*. John Wiley & Sons; Chapman & Hall.

Wedderburn, R. W. M. 1974. Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. *Biometrika*, **61**(3), 439–447.

Welch, B. L., and Peers, H. W. 1963. On formulae for confidence points based on integrals of weighted likelihoods. *J. Roy. Statist. Soc. B*, **25**, 318–329.

Westfall, P., and Young, S. 1993. *Resampling-based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley Series in Probability and Statistics. Wiley-Interscience.

Xie, M., and Singh, K. 2013. Confidence distribution, the frequentist distribution estimator of a parameter: A review. *Int. Statist. Rev.*, **81**(1), 3–39. with discussion.

Ye, J. 1998. On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.*, **93**(441), 120–131.

Zhang, C.-H., and Zhang, S. 2014. Confidence intervals for low-dimensional parameters with high-dimensional data. *J. Roy. Statist. Soc. B*, **76**(1), 217–242.

Zou, H., Hastie, T., and Tibshirani, R. 2007. On the “degrees of freedom” of the lasso. *Ann. Statist.*, **35**(5), 2173–2192.

Author Index