13

Objective Bayes Inference and Markov Chain Monte Carlo

From its very beginnings, Bayesian inference exerted a powerful influence on statistical thinking. The notion of a single coherent methodology employing only the rules of probability to go from assumption to conclusion was and is immensely attractive. For 200 years, however, two impediments stood between Bayesian theory's philosophical attraction and its practical application.

1 In the absence of relevant past experience, the choice of a prior distribution introduces an unwanted subjective element into scientific inference.
2 Bayes' rule (3.5) looks simple enough, but carrying out the numerical calculation of a posterior distribution often involves intricate higher-dimensional integrals.

The two impediments fit neatly into the dichotomy of Chapter 1, the first being inferential and the second algorithmic.[1]

A renewed cycle of Bayesian enthusiasm took hold in the 1960s, at first concerned mainly with coherent inference. Building on work by Bruno de Finetti and L. J. Savage, a principled theory of *subjective probability* was constructed: the Bayesian statistician, by the careful elicitation of prior knowledge, utility, and belief, arrives at the correct *subjective prior distribution* for the problem at hand. Subjective Bayesianism is particularly appropriate for individual decision making, say for the business executive trying to choose the best investment in the face of uncertain information.

It is less appropriate for scientific inference, where the sometimes skeptical world of science puts a premium on objectivity. An answer came from the school of *objective Bayes inference*. Following the approach of Laplace and Jeffreys, as discussed in Section 3.2, their goal was to fashion objective, or "uninformative," prior distributions that in some sense were unbiased in their effects upon the data analysis.

[1] The exponential family material in this chapter provides technical support, but is not required in detail for a general understanding of the main ideas.

In what came as a surprise to the Bayes community, the objective school has been the most successful in bringing Bayesian ideas to bear on scientific data analysis. Of the 24 articles in the December 2014 issue of the *Annals of Applied Statistics*, 8 employed Bayesian analysis, predominantly based on objective priors.

This is where electronic computation enters the story. Commencing in the 1980s, dramatic steps forward were made in the numerical calculation of high-dimensional Bayes posterior distributions. *Markov chain Monte Carlo* (MCMC) is the generic name for modern posterior computation algorithms. These proved particularly well suited for certain forms of objective Bayes prior distributions.

Taken together, objective priors and MCMC computations provide an attractive package for the statistician faced with a complicated data analysis situation. Statistical inference becomes almost automatic, at least compared with the rigors of frequentist analysis. This chapter discusses both parts of the package, the choice of prior and the subsequent computational methods. Criticisms arise, both from the frequentist viewpoint and that of informative Bayesian analysis, which are brought up here and also in Chapter 21.

13.1 Objective Prior Distributions

A *flat*, or uniform, distribution over the space of possible parameter values seems like the obvious choice for an uninformative prior distribution, and has been so ever since Laplace's advocacy in the late eighteenth century. For a finite parameter space Ω, say

$$\Omega = \{\mu_{(1)}, \mu_{(2)}, \ldots, \mu_{(K)}\}, \tag{13.1}$$

"flat" has the obvious meaning

$$g^{\text{flat}}(\mu) = \frac{1}{K} \qquad \text{for all } \mu \in \Omega. \tag{13.2}$$

If K is infinite, or if Ω is continuous, we can still take

$$g^{\text{flat}}(\mu) = \text{constant}. \tag{13.3}$$

Bayes' rule (3.5) gives the same posterior distribution for any choice of the constant,

$$\begin{aligned} g^{\text{flat}}(\mu|x) &= g^{\text{flat}}(\mu)\, f_\mu(x)/f(x), \quad \text{with} \\ f(x) &= \int_\Omega f_\mu(x) g^{\text{flat}}(\mu)\, d\mu. \end{aligned} \tag{13.4}$$

Notice that $g^{\text{flat}}(\mu)$ cancels out of $g^{\text{flat}}(\mu|x)$. The fact that $g^{\text{flat}}(\mu)$ is "improper," that is, it integrates to infinity, doesn't affect the formal use of Bayes' rule in (13.4) as long as $f(x)$ is finite.

Notice also that $g^{\text{flat}}(\mu|x)$ amounts to taking the posterior density of μ to be proportional to the likelihood function $L_x(\mu) = f_\mu(x)$ (with x fixed and μ varying over Ω). This brings us close to Fisherian inference, with its emphasis on the direct interpretation of likelihoods, but Fisher was adamant in his insistance that likelihood was not probability.

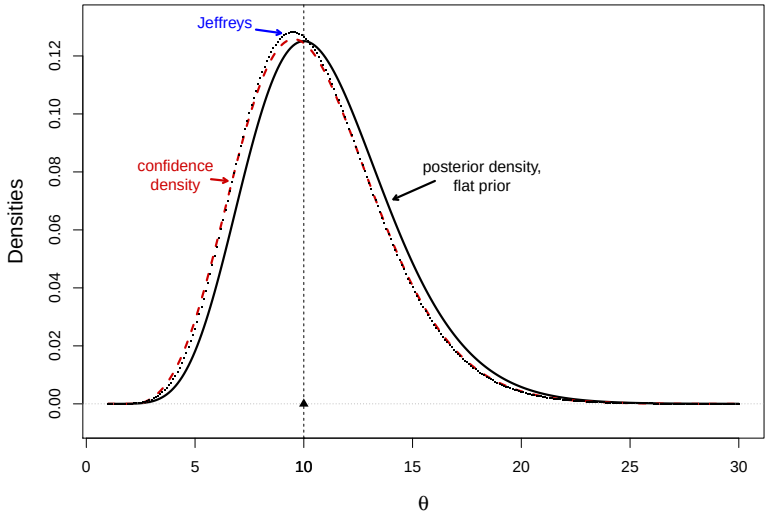

Figure 13.1 The solid curve is flat-prior posterior density (13.4) having observed $x = 10$ from Poisson model $x \sim \text{Poi}(\mu)$; it is shifted about 0.5 units right from the confidence density (dashed) of Figure 11.6. Jeffreys' prior gives a posterior density (dotted) nearly the same as the confidence density.

The solid curve in Figure 13.1 shows $g^{\text{flat}}(\mu|x)$ for the Poisson situation of Table 11.2,

$$x \sim \text{Poi}(\mu), \tag{13.5}$$

with $x = 10$ observed; $g^{\text{flat}}(\mu|x)$ is shifted almost exactly 0.5 units right of the confidence density from Figure 11.6. ("θ" is μ itself in this case.)[2]

Fisher's withering criticism of flat-prior Bayes inference focused on its

[2] The reader may wish to review Chapter 11, particularly Section 11.6, for these constructions.

lack of transformation invariance. If we were interested in $\theta = \log(\mu)$ rather than μ, $g^{\text{flat}}(\theta|x)$ would not be the transformation to the log scale of $g^{\text{flat}}(\mu|x)$. Jeffreys' prior, (3.17) or (11.72), which *does* transform correctly, is

$$g^{\text{Jeff}}(\mu) = 1/\sqrt{\mu} \tag{13.6}$$

for $x \sim \text{Poi}(\mu)$; $g^{\text{Jeff}}(\mu|x = 10)$ is then a close match to the confidence density in Figure 13.1.

Coverage Matching Priors

A variety of improvements and variations on Jeffreys' prior have been suggested for use as general-purpose uninformative prior distributions, as †1 briefly discussed in the chapter endnotes.† All share the drawback seen in Figure 11.7: the posterior distribution $g(\mu|x)$ can have unintended effects on the resulting inferences for a real-valued parameter of interest $\theta = t(\mu)$. This is unavoidable; it is mathematically impossible for any single prior to be uninformative for every choice of $\theta = t(\mu)$.

The label "uninformative" for a prior sometimes means "gives Bayes posterior intervals that closely match confidence intervals." Perhaps surprisingly, this definition has considerable resonance in the Bayes community. Such priors can be constructed for any given scalar parameter of interest $\theta = t(\mu)$, for instance the maximum eigenvalue parameter of Fig- †2 ure 11.7. In brief, the construction proceeds as follows.†

- The p-dimensional parameter vector μ is transformed to a form that makes θ the first coordinate, say

$$\mu \to (\theta, \nu), \tag{13.7}$$

  where ν is a $(p-1)$-dimensioned nuisance parameter.
- The transformation is chosen so that the Fisher information matrix (11.72) for (θ, ν) has the "diagonal" form

$$\begin{pmatrix} \mathcal{I}_{\theta\theta} & 0 \\ 0' & \boldsymbol{\mathcal{I}}_{\nu\nu} \end{pmatrix}. \tag{13.8}$$

  (This is always possible.)
- Finally, the prior for (θ, ν) is taken proportional to

$$g(\theta, \nu) = \mathcal{I}_{\theta\theta}^{1/2} h(\nu), \tag{13.9}$$

  where $h(\nu)$ is an arbitrary $(p-1)$-dimensional density. In other words,

$g(\theta, \nu)$ combines the one-dimensional Jeffreys' prior (3.16) for θ with an arbitrary independent prior for the orthogonal nuisance parameter vector ν.

The main thing to notice about (13.9) is that $g(\theta, \nu)$ represents different priors on the original parameter vector μ for different functions $\theta = t(\mu)$. No single prior $g(\mu)$ can be uninformative for all choices of the parameter of interest θ.

Calculating $g(\theta, \nu)$ can be difficult. One alternative is to go directly to the BCa confidence density (11.68)–(11.69), which can be interpreted as the posterior distribution from an uninformative prior (because its integrals agree closely with confidence interval endpoints).

Coverage matching priors are not much used in practice, and in fact none of the eight *Annals of Applied Statistics* objective Bayes papers mentioned earlier were of type (13.9). A form of "almost uninformative" priors, the conjugates, is more popular, mainly because of the simpler computation of their posterior distributions.

13.2 Conjugate Prior Distributions

A mathematically convenient class of prior distributions, the *conjugate priors*, applies to samples from an exponential family,[3] Section 5.5,

$$f_\mu(x) = e^{\alpha x - \psi(\alpha)} f_0(x). \tag{13.10}$$

Here we have indexed the family with the expectation parameter

$$\mu = E_f\{x\}, \tag{13.11}$$

rather than the canonical parameter α. On the right-hand side of (13.10), α can be thought of as a one-to-one function of μ (the so-called "link function"), e.g., $\alpha = \log(\mu)$ for the Poisson family. The observed data is a random sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ from f_μ,

$$x_1, x_2, \ldots, x_n \overset{\text{iid}}{\sim} f_\mu, \tag{13.12}$$

having density function

$$f_\mu(\boldsymbol{x}) = e^{n[\alpha \bar{x} - \psi(\alpha)]} f_0(\boldsymbol{x}), \tag{13.13}$$

the average $\bar{x} = \sum x_i / n$ being sufficient.

[3] We will concentrate on one-parameter families, though the theory extends to the multiparameter case. Figure 13.2 relates to a two-parameter situation.

The family of conjugate priors for μ, $g_{n_0,x_0}(\mu)$, allows the statistician to choose two parameters, n_0 and x_0,

$$g_{n_0,x_0}(\mu) = ce^{n_0[x_0\alpha-\psi(\alpha)]}/V(\mu), \tag{13.14}$$

$V(\mu)$ the variance of an x from f_μ,

$$V(\mu) = \text{var}_f\{x\}; \tag{13.15}$$

c is the constant that makes $g_{n_0,x_0}(\mu)$ integrate to 1 with respect to Lebesgue measure on the interval of possible μ values. The interpretation is that x_0 represents the average of n_0 hypothetical prior observations from f_μ.

The utility of conjugate priors is seen in the following theorem.

†3 **Theorem 13.1** † *Define*

$$n_+ = n_0 + n \quad \textit{and} \quad \bar{x}_+ = \frac{n_0}{n_+}x_0 + \frac{n}{n_+}\bar{x}. \tag{13.16}$$

Then the posterior density of μ *given* $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ *is*

$$g(\mu|\boldsymbol{x}) = g_{n_+,\bar{x}_+}(\mu); \tag{13.17}$$

moreover, the posterior expectation of μ *given* $\boldsymbol{x}$ *is*

$$E\{\mu|\boldsymbol{x}\} = \frac{n_0}{n_+}x_0 + \frac{n}{n_+}\bar{x}. \tag{13.18}$$

The intuitive interpretation is quite satisfying: we begin with a hypothetical prior sample of size n_0, sufficient statistic x_0; observe $\boldsymbol{x}$, a sample of size n; and update our prior distribution $g_{n_0,x_0}(\mu)$ to a distribution $g_{n_+,\bar{x}_+}(\mu)$ of the same form. Moreover, $E\{\mu|\boldsymbol{x}\}$ equals the average of a hypothetical sample with n_0 copies of x_0,

$$(x_0, x_0, \ldots, x_0, x_1, x_2, \ldots, x_n). \tag{13.19}$$

As an example, suppose $x_i \overset{\text{iid}}{\sim} \text{Poi}(\mu)$, that is we have n i.i.d. observations from a Poisson distribution, Table 5.1. Formula (13.14) gives conju-
†4 gate prior †

$$g_{n_0,x_0}(\mu) = c\mu^{n_0x_0-1}e^{-n_0\mu}, \tag{13.20}$$

c not depending on μ. So in the notation of Table 5.1, $g_{n_0,x_0}(\mu)$ is a gamma distribution, $\text{Gam}(n_0x_0, 1/n_0)$. The posterior distribution is

$$\begin{aligned} g(\mu|\boldsymbol{x}) = g_{n_+,\bar{x}_+}(\mu) &\sim \text{Gam}(n_+\bar{x}_+, 1/n_+) \\ &\sim \frac{1}{n_+}G_{n_+\bar{x}_+}, \end{aligned} \tag{13.21}$$

†5 where G_ν indicates a standard gamma distribution,†

$$G_\nu = \text{Gam}(\nu, 1). \tag{13.22}$$

Table 13.1 *Conjugate priors* (13.14)–(13.16) *for four familiar one-parameter exponential families, using notation in Table 5.1; the last column shows the posterior distribution of μ given n observations x_i, starting from prior $g_{n_0,x_0}(\mu)$. In line 4, G_ν is the standard gamma distribution* $\text{Gam}(\nu, 1)$*, with μ the same as gamma parameter σ in Table 5.1. The chapter endnotes give the density of the inverse gamma distribution $1/G_\nu$, and corresponding results for chi-squared variates.*

	Name	x_i distribution	$g_{n_0,x_0}(\mu)$	$g(\mu\|\boldsymbol{x})$
1.	Normal	$\mathcal{N}(\mu,\sigma_1^2)$ (σ_1^2 known)	$\mathcal{N}(x_0,\sigma_1^2/n_0)$	$\mathcal{N}(\bar{x}_+,\sigma_1^2/n_+)$
2.	Poisson	$\text{Poi}(\mu)$	$\text{Gam}(n_0x_0, 1/n_0)$	$\text{Gam}(n_+\bar{x}_+, 1/n_+)$
3.	Binomial	$\text{Bi}(1,\mu)$	$\text{Be}(n_0x_0, n_0(1-x_0))$	$\text{Be}(n_+\bar{x}_+, n_+(1-\bar{x}_+))$
4.	Gamma	$\mu G_\nu/\nu$ (ν known)	$n_0x_0\nu/G_{n_0\nu+1}$	$n_+\bar{x}_+\nu/G_{n_+\nu+1}$

Table 13.1 describes the conjugate prior and posterior distributions for four familiar one-parameter families. The binomial case, where μ is the "success probability" π in Table 5.1, is particularly evocative: independent coin flips $x_1, x_2, \ldots, x_n$ give, say, $s = \sum_i x_i = n\bar{x}$ successes. Prior $g_{n_0,x_0}(\pi)$ amounts to assuming proportion $x_0 = s_0/n_0$ prior successes in n_0 flips. Formula (13.18) becomes

$$E\{\pi|\boldsymbol{x}\} = \frac{s_0 + s}{n_0 + n} \tag{13.23}$$

for the posterior expectation of π. The choice $(n_0, x_0) = (2, 1/2)$ for instance gives Bayesian estimate $(s+1)/(n+2)$ for π, pulling the MLE s/n a little bit toward 1/2.

The size of n_0, the number of hypothetical prior observations, determines how informative or uninformative the prior $g_{n_0,x_0}(\mu)$ is. Recent objective Bayes literature has favored choosing n_0 small, $n_0 = 1$ being popular. The hope here is to employ a *proper* prior (one that has a finite integral), while still not injecting much unwarranted information into the analysis. The choice of x_0 is also by convention. One possibility is to set

$x_0 = \bar{x}$, in which case the posterior expectation $E\{\mu|\boldsymbol{x}\}$ (13.18) equals the MLE $\bar{x}$. Another possibility is choosing x_0 equal to a "null" value, for instance $x_0 = 0$ for effect size estimation in (3.28).

Table 13.2 *Vasoconstriction data; volume of air inspired in 39 cases, 19 without vasoconstriction ($y = 0$) and 20 with vasoconstriction ($y = 1$).*

$y = 0$		$y = 1$	
60	98	85	115
74	98	88	120
78	104	88	126
78	104	90	126
78	113	90	128
88	118	93	136
90	120	104	143
95	123	108	151
95	137	110	154
98		111	157

†6 As a miniature example of objective Bayes inference, we consider the *vasoconstriction data*† of Table 13.2: $n = 39$ measurements of lung volume have been obtained, 19 without vasoconstriction ($y = 0$) and 20 with ($y = 1$). Here we will think of the y_i as binomial variates,

$$y_i \overset{\text{ind}}{\sim} \text{Bi}(1, \pi_i), \qquad i = 1, 2, \ldots, 39, \tag{13.24}$$

following logistic regression model (8.5),

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_0 + \alpha_1 x_i, \tag{13.25}$$

with the x_i as fixed covariates (the values in Table 13.2).

Letting $X_i = (1, x_i)'$, (13.24)–(13.25) results in a two-parameter exponential family (8.24),

$$f_\alpha(\boldsymbol{y}) = e^{n\left[\alpha'\hat{\beta} - \psi(\alpha)\right]} f_0(\boldsymbol{y}), \tag{13.26}$$

having

$$\hat{\beta} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i, \sum_{i=1}^{n} x_i y_i\right)' \quad \text{and} \quad \psi(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{\alpha' X_i}).$$

The MLE $\hat{\alpha}$ has approximate 2×2 covariance matrix $\hat{V}$ as given in (8.30).

In Figure 13.2, the posterior distributions are graphed in terms of

$$\gamma = \hat{V}^{-1/2}(\alpha - \hat{\alpha}) \tag{13.27}$$

rather than α or μ, making the contours of equal density roughly circular and centered at zero.

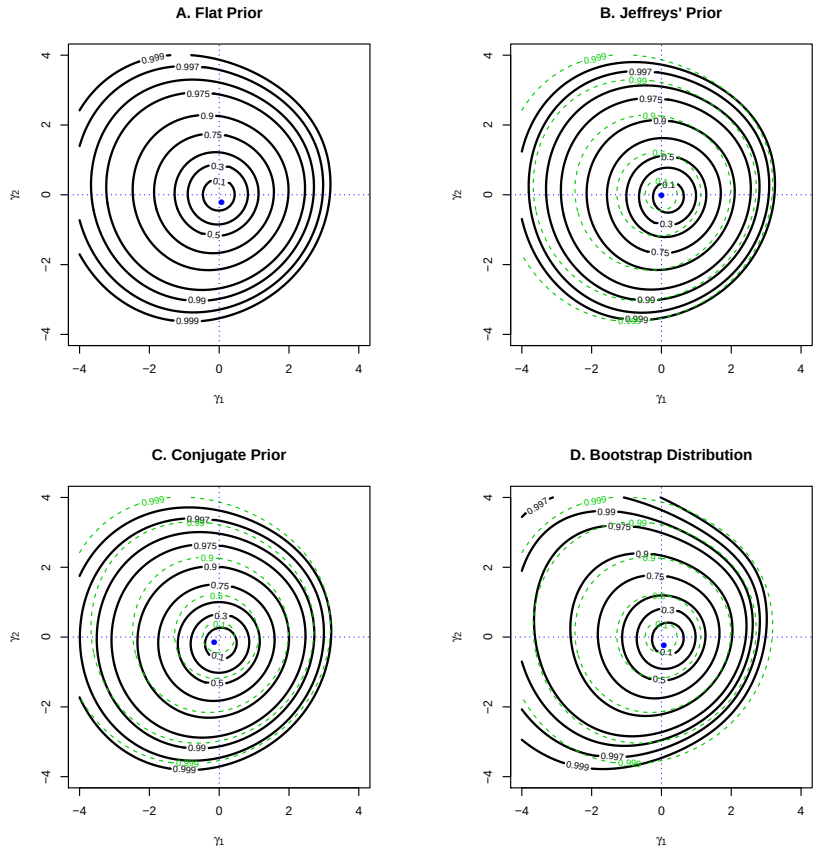

Figure 13.2 Vasoconstriction data; contours of equal posterior density of γ (13.27) from four uninformative priors, as described in the text. Numbers indicate probability content within contours; light dashed contours from Panel A, flat prior.

Panel A of Figure 13.2 illustrates the flat prior posterior density of γ given the data $\boldsymbol{y}$ in model (13.24)–(13.25). The heavy lines are contours of equal density, with the one labeled "0.9" containing 90% of the posterior probability, etc. Panel B shows the corresponding posterior density

contours obtained from Jeffreys' multiparameter prior (11.72), in this case

$$g^{\text{Jeff}}(\alpha) = |\boldsymbol{V}_\alpha|^{1/2}, \tag{13.28}$$

$\boldsymbol{V}_\alpha$ the covariance matrix of $\hat{\alpha}$, as calculated from (8.30). For comparison purposes the light dashed curves show some of the flat prior contours from panel A. The effect of $g^{\text{Jeff}}(\alpha)$ is to reduce the flat prior bulge toward the upper left corner.

Panel C relates to the conjugate prior[4] $g_{1,0}(\alpha)$. Besides reducing the flat prior bulge, $g_{1,0}(\alpha)$ pulls the contours slightly downward.

Panel D shows the parametric bootstrap distribution: model (13.24)–(13.25), with $\hat{\alpha}$ replacing α, gave resamples $\boldsymbol{y}^*$ and MLE replications $\hat{\alpha}^*$. The contours of $\hat{\gamma}^* = \hat{V}^{-1/2}(\hat{\alpha}^* - \hat{\alpha})$ considerably accentuate the bulge toward the left.

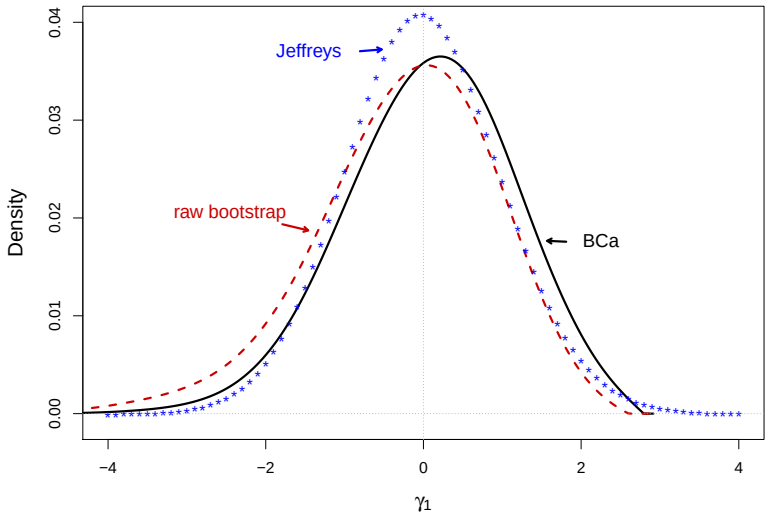

Figure 13.3 Posterior densities for γ_1, first coordinate of γ in (13.27), for the vasoconstriction data. Dashed red curve: raw (unweighted) distribution of $B = 8000$ parametric replications from model (13.24)–(13.25); solid black curve: BCa density (11.68) ($z_0 = 0.123$, $a = 0.053$); dotted blue curve: posterior density using Jeffreys multiparameter prior (11.72).

[4] The role of $\bar{x}$ in (13.13) is taken by $\hat{\beta}$ in (13.26), so $g_{1,0}$ has $\hat{\beta} = \mathbf{0}$, $n_0 = \mathbf{1}$. This makes $g_{1,0}(\boldsymbol{\alpha}) = \exp\{-\psi(\boldsymbol{\alpha})\}$. The factor $V(\mu)$ in (13.14) is absent in the conjugate prior for $\boldsymbol{\alpha}$ (as opposed to μ).

This doesn't necessarily imply that a bootstrap analysis would give much different answers than the three (quite similar) objective Bayes results. For any particular real-valued parameter of interest θ, the raw bootstrap distribution (equal weight on each replication) would be reweighted according to the BCa formula (11.68) in order to produce accurate confidence intervals. Figure 13.3 compares the raw bootstrap distribution, the BCa confidence density, and the posterior density obtained from Jeffreys' prior, for θ equal to γ_1, the first coordinate of γ in (13.27). The BCa density is shifted to the right of Jeffreys'.

Critique of Objective Bayes Inference

Despite its simplicity, or perhaps because of it, objective Bayes procedures are vulnerable to criticism from both ends of the statistical spectrum. From the subjectivist point of view, objective Bayes is only partially Bayesian: it employs Bayes' theorem but without doing the hard work of determining a convincing prior distribution. This introduces frequentist elements into its practice—clearly so in the case of Jeffreys' prior—along with frequentist incoherencies.

For the frequentist, objective Bayes analysis can seem dangerously untethered from the usual standards of accuracy, having only tenuous large-sample claims to legitimacy. This is more than a theoretical objection. The practical advantages claimed for Bayesian methods depend crucially on the fine structure of the prior. Can we safely ignore stopping rules or selective inference (e.g., choosing the largest of many estimated parameters for special attention) for a prior not based on some form of genuine experience?

In an era of large, complicated, and difficult data-analytic problems, objective Bayes methods are answering a felt need for relatively straightforward paths to solution. Granting their usefulness, it is still reasonable to hope for better justification,[5] or at least for more careful comparisons with competing methods as in Figure 13.3.

13.3 Model Selection and the Bayesian Information Criterion

Data-based model selection has become a major theme of modern statistical inference. In the problem's simplest form, the statistician observes data $\boldsymbol{x}$ and wishes to choose between a smaller model $\mathcal{M}_0$ and a larger model

[5] Chapter 20 discusses the frequentist assessment of Bayes and objective Bayes estimates.

$\mathcal{M}_1$. The classic textbook example takes $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)'$ as an independent normal sample,

$$x_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, 1) \qquad \text{for } i = 1, 2, \ldots, n, \tag{13.29}$$

with $\mathcal{M}_0$ the null hypothesis $\mu = 0$ and $\mathcal{M}_1$ the general two-sided alternative,

$$\mathcal{M}_0 : \mu = 0, \qquad \mathcal{M}_1 : \mu \neq 0. \tag{13.30}$$

(We can include $\mu = 0$ in $\mathcal{M}_1$ with no effect on what follows.) From a frequentist viewpoint, choosing between $\mathcal{M}_0$ and $\mathcal{M}_1$ in (13.29)–(13.30) amounts to running a hypothesis test of $H_0 : \mu = 0$, perhaps augmented with a confidence interval for μ.

Bayesian model selection aims for more: an evaluation of the posterior probabilities of $\mathcal{M}_0$ and $\mathcal{M}_1$ given $\boldsymbol{x}$. A full Bayesian specification requires prior probabilities for the two models,

$$\pi_0 = \Pr\{\mathcal{M}_0\} \quad \text{and} \quad \pi_1 = 1 - \pi_0 = \Pr\{\mathcal{M}_1\}, \tag{13.31}$$

and conditional prior densities for μ within each model,

$$g_0(\mu) = g(\mu|\mathcal{M}_0) \quad \text{and} \quad g_1(\mu) = g(\mu|\mathcal{M}_1). \tag{13.32}$$

Let $f_\mu(\boldsymbol{x})$ be the density of $\boldsymbol{x}$ given μ. Each model induces a marginal density for $\boldsymbol{x}$, say

$$f_0(\boldsymbol{x}) = \int_{\mathcal{M}_0} f_\mu(\boldsymbol{x}) g_0(\mu)\, d\mu \quad \text{and} \quad f_1(\boldsymbol{x}) = \int_{\mathcal{M}_1} f_\mu(\boldsymbol{x}) g_1(\mu)\, d\mu. \tag{13.33}$$

Bayes' theorem, in its ratio form (3.8), then gives posterior probabilities

$$\pi_0(\boldsymbol{x}) = \Pr\{\mathcal{M}_0|\boldsymbol{x}\} \quad \text{and} \quad \pi_1(\boldsymbol{x}) = \Pr\{\mathcal{M}_1|\boldsymbol{x}\} \tag{13.34}$$

satisfying

$$\frac{\pi_1(\boldsymbol{x})}{\pi_0(\boldsymbol{x})} = \frac{\pi_1}{\pi_0} B(\boldsymbol{x}), \tag{13.35}$$

where $B(\boldsymbol{x})$ is the *Bayes factor*

$$B(\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})}, \tag{13.36}$$

leading to the elegant statement that the posterior odds ratio is the prior odds ratio times the Bayes factor.

All of this is of more theoretical than applied use. Prior specifications (13.31)–(13.32) are usually unavailable in practical settings (which is why

standard hypothesis testing is so popular). The objective Bayes school has concentrated on estimating the Bayes factor $B(\boldsymbol{x})$, with the understanding that the prior odds ratio π_1/π_0 in (13.35) would be roughly evaluated depending on the specific circumstances—perhaps set to the Laplace choice $\pi_1/\pi_0 = 1$.

Table 13.3 *Jeffreys' scale of evidence for the interpretation of Bayes factors.*

Bayes factor	Evidence for M_1
< 1	negative
1–3	barely worthwhile
3–20	positive
20–150	strong
> 150	very strong

Jeffreys suggested a scale of evidence for interpreting Bayes factors, reproduced in Table 13.3;† $B(\boldsymbol{x}) = 10$ for instance constitutes *positive* but not *strong* evidence in favor of the bigger model. Jeffreys' scale is a Bayesian version of Fisher's interpretive scale for the outcome of a hypothetic test, with coverage value (one minus the significance level) 0.95 famously constituting "significant" evidence against the null hypothesis. Table 13.4 shows Fisher's scale, as commonly interpreted in the biomedical and social sciences. †7

Table 13.4 *Fisher's scale of evidence against null hypothesis $\mathcal{M}_0$ and in favor of $\mathcal{M}_1$, as a function of coverage level (1 minus the p-value).*

Coverage	(p-value)	Evidence for M_1
.80	(.20)	null
.90	(.10)	borderline
.95	(.05)	moderate
.975	(.025)	substantial
.99	(.01)	strong
.995	(.005)	very strong
.999	(.001)	overwhelming

Even if we accept the reduction of model selection to assessing the Bayes factor $B(\boldsymbol{x})$ in (13.35), and even if we accept Jeffreys' scale of interpretation, this still leaves a crucial question: how to compute $B(\boldsymbol{x})$ in

practice, without requiring informative choices of the priors g_0 and g_1 in (13.32).

A popular objective Bayes answer is provided by the *Bayesian information criterion*[†] (BIC). For a given model $\mathcal{M}$ we define

†8

$$\mathrm{BIC}(\mathcal{M}) = \log\left\{f_{\hat{\mu}}(\boldsymbol{x})\right\} - \frac{p}{2}\log(n), \tag{13.37}$$

where $\hat{\mu}$ is the MLE, p the degrees of freedom (number of free parameters) in $\mathcal{M}$, and n the sample size. Then the BIC approximation to Bayes factor $B(\boldsymbol{x})$ (13.36) is

$$\begin{aligned}\log B_{\mathrm{BIC}}(\boldsymbol{x}) &= \mathrm{BIC}(\mathcal{M}_1) - \mathrm{BIC}(\mathcal{M}_0)\\ &= \log\left\{f_{\hat{\mu}_1}(\boldsymbol{x})/f_{\hat{\mu}_0}(\boldsymbol{x})\right\} - \frac{p_1 - p_0}{2}\log(n),\end{aligned} \tag{13.38}$$

the subscripts indexing the MLEs and degrees of freedom in $\mathcal{M}_1$ and $\mathcal{M}_0$.

This can be restated in somewhat more familiar terms. Letting $W(\boldsymbol{x})$ be *Wilks' likelihood ratio statistic*,

$$W(\boldsymbol{x}) = 2\log\left\{f_{\hat{\mu}_1}(\boldsymbol{x})/f_{\hat{\mu}_0}(\boldsymbol{x})\right\}, \tag{13.39}$$

we have

$$\log B_{\mathrm{BIC}}(\boldsymbol{x}) = \frac{1}{2}\left\{W(\boldsymbol{x}) - d\,\log(n)\right\}, \tag{13.40}$$

with $d = p_1 - p_0$. $W(\boldsymbol{x})$ approximately follows a χ^2_d distribution under model $\mathcal{M}_0$, $E_0\{W(\boldsymbol{x})\} \doteq d$, implying $B_{\mathrm{BIC}}(\boldsymbol{x})$ will tend to be less than one, favoring $\mathcal{M}_0$ if it is true, ever more strongly as n increases.

We can apply BIC selection to the vasoconstriction data of Table 13.2, taking $\mathcal{M}_1$ to be model (13.24)–(13.25), and $\mathcal{M}_0$ to be the submodel having $\alpha_1 = 0$. In this case $d = 1$ in (13.40). Direct calculation gives $W = 7.07$ and

$$B_{\mathrm{BIC}} = 5.49, \tag{13.41}$$

positive but not *strong* evidence against $\mathcal{M}_0$ according to Jeffreys' scale. By comparison, the usual frequentist z-value for testing $\alpha_1 = 0$ is 2.36, coverage level 0.982, between *substantial* and *strong* evidence against $\mathcal{M}_0$ on Fisher's scale.

The BIC was named in reference to Akaike's information criterion (AIC),

$$\mathrm{AIC}(\mathcal{M}) = \log\left\{f_{\hat{\mu}}(\boldsymbol{x})\right\} - p, \tag{13.42}$$

which suggests, as in (12.73), basing model selection on the sign of

$$\mathrm{AIC}(\mathcal{M}_1) - \mathrm{AIC}(\mathcal{M}_0) = \frac{1}{2}\left\{W(\boldsymbol{x}) - 2d\right\}. \tag{13.43}$$

The BIC penalty $d\log(n)$ in (13.40) grows more severe than the AIC penalty $2d$ as n gets larger, increasingly favoring selection of $\mathcal{M}_0$ rather than $\mathcal{M}_1$. The distinction is rooted in Bayesian notions of coherent behavior, as discussed in what follows.

Where does the BIC penalty term $d\log(n)$ in (13.40) come from? A first answer uses the simple normal model $x_i \sim \mathcal{N}(\mu, 1)$, (13.29)–(13.30). $\mathcal{M}_0$ has prior $g_0(\mu) = g(\mu|\mathcal{M}_0)$ equal a delta function at zero. Suppose we take $g_1(\mu) = g(\mu|\mathcal{M}_1)$ in (13.32) to be the Gaussian conjugate prior

$$g_1(\mu) \sim \mathcal{N}(M, A). \tag{13.44}$$

The discussion following (13.23) in Section 13.2 suggests setting $M = 0$ and $A = 1$, corresponding to prior information equivalent to one of the n actual observations. In this case we can calculate the actual Bayes factor $B(\boldsymbol{x})$,

$$\log B(\boldsymbol{x}) = \frac{1}{2}\left\{\frac{n}{n+1}W(\boldsymbol{x}) - \log(n+1)\right\}, \tag{13.45}$$

nearly equaling $\log B_{\text{BIC}}(\boldsymbol{x})$ ($d = 1$), for large n. Justifications of the BIC formula as an approximate Bayes factor follow generalizations of this kind of argument, as discussed in the chapter endnotes.

The difference between BIC and frequentist hypothesis testing grows more drastic for large n. Suppose $\mathcal{M}_0$ is a regression model and $\mathcal{M}_1$ is $\mathcal{M}_0$ augmented with one additional covariate (so $d = 1$). Let z be a standard z-value for testing the hypothesis that $\mathcal{M}_1$ is no improvement over $\mathcal{M}_0$,

$$z \doteq \mathcal{N}(0, 1) \text{ under } \mathcal{M}_0. \tag{13.46}$$

Table 13.5 shows $B_{\text{BIC}}(\boldsymbol{x})$ as a function of z and n. At $n = 15$ Fisher's and Jeffreys' scales give roughly similar assessments of the evidence against $\mathcal{M}_0$ (though Jeffreys' nomenclature is more conservative). At the other end of the table, at $n = 10,000$, the inferences are contradictory: $z = 3.29$, with p-value 0.001 and coverage level 0.999, is *overwhelming* evidence for $\mathcal{M}_1$ on Fisher's scale, but *barely worthwhile* for Jeffreys'. Bayesian coherency, the axiom that inferences should be consistent over related situations, lies behind the contradiction.

Suppose $n = 1$ in the simple normal model (13.29)–(13.30). That is, we observe only the single variable

$$x \sim \mathcal{N}(\mu, 1), \tag{13.47}$$

and wish to decide between $\mathcal{M}_0 : \mu = 0$ and $\mathcal{M}_1 : \mu \neq 0$. Let $g_1^{(1)}(\mu)$ denote our $\mathcal{M}_1$ prior density (13.32) for this situation.

Table 13.5 *BIC Bayes factors corresponding to z-values for testing one additional covariate; coverage value (1 minus the significance level) of a two-sided hypothesis test as interpreted by Fisher's scale of evidence, right. Jeffreys' scale of evidence, Table 13.3, is in rough agreement with Fisher for $n = 15$, but favors the null much more strongly for larger sample sizes.*

		n							
Cover	z-value	15	50	250	1000	2500	5000	10000	Fisher
.80	1.28	.59	.32	.14	.07	.05	.03	.02	null
.90	1.64	1.00	.55	.24	.12	.08	.05	.04	borderline
.95	1.96	1.76	.97	.43	.22	.14	.10	.07	moderate
.975	2.24	3.18	1.74	.78	.39	.25	.17	.12	substantial
.99	2.58	7.12	3.90	1.74	.87	.55	.39	.28	strong
.995	2.81	13.27	7.27	3.25	1.63	1.03	.73	.51	very strong
.999	3.29	57.96	31.75	14.20	7.10	4.49	3.17	2.24	overwhelming

The case $n > 1$ in (13.29) is logically identical to (13.47). Letting $x^{(n)} = \sqrt{n}(\sum x_i/n)$ and $\mu^{(n)} = \sqrt{n}\mu$ gives

$$x^{(n)} \sim \mathcal{N}\left(\mu^{(n)}, 1\right), \tag{13.48}$$

with (13.30) becoming $\mathcal{M}_0 : \mu^{(n)} = 0$ and $\mathcal{M}_1 : \mu^{(n)} \neq 0$. Coherency requires that $\mu^{(n)}$ in (13.48) have the same $\mathcal{M}_1$ prior as μ in (13.47). Since $\mu = \mu^{(n)}/\sqrt{n}$, this implies that $g_1^{(n)}(\mu)$, the $\mathcal{M}_1$ prior for sample size n, satisfies

$$g_1^{(n)}(\mu) = g_1^{(1)}\left(\mu/\sqrt{n}\right)/\sqrt{n}, \tag{13.49}$$

this being "sample size coherency."

The effect of (13.49) is to spread the $\mathcal{M}_1$ prior density $g_1^{(n)}(\mu)$ farther away from the null value $\mu = 0$ at rate $\sqrt{n}$, while the $\mathcal{M}_0$ prior $g_0^{(n)}(\mu)$ stays fixed. For any fixed value of the sufficient statistic $x^{(n)}$ ($x^{(n)}$ being "z" in Table 13.5), this results in the Bayes factor $B(x^{(n)})$ decreasing at rate $1/\sqrt{n}$; the frequentist/Bayesian contradiction seen in Table 13.5 goes beyond the specifics of the BIC algorithm.

A *general information criterion* takes the form

$$\mathrm{GIC}(\mathcal{M}) = \log f_{\hat{\mu}}(\boldsymbol{x}) - p\, c_n, \tag{13.50}$$

where c_n is any sequence of positive numbers; $c_n = \log(n)/2$ for BIC (13.37) and $c_n = 1$ for AIC (13.42). The difference

$$\Delta \equiv \text{GIC}(\mathcal{M}_1) - \text{GIC}(\mathcal{M}_0) = \frac{1}{2}\,(W(\boldsymbol{x}) - 2c_n d)\,, \tag{13.51}$$

$d = p_1 - p_0$, will be positive if $W(\boldsymbol{x}) > 2c_n d$. For $d = 1$, as in Table 13.5, Δ will favor $\mathcal{M}_1$ if $W(\boldsymbol{x}) \geq 2c_n$, with approximate probability, if $\mathcal{M}_0$ is actually true,

$$\Pr\{\chi^2_1 \geq 2c_n\}. \tag{13.52}$$

This equals 0.157 for the AIC choice $c_n = 1$; for BIC, $n = 10,000$, it equals 0.0024. The choice

$$c_n = 1.92 \tag{13.53}$$

makes $\Pr\{\Delta > 0|\mathcal{M}_0\} \doteq 0.05$, agreeing with the usual frequentist 0.05 rejection level.

The BIC is *consistent*: $\Pr\{\Delta > 0\}$ goes to zero as $n \to \infty$ if $\mathcal{M}_0$ is true. This isn't true of (13.53) for instance, where we will have $\Pr\{\Delta > 0\} \doteq 0.05$ no matter how large n may be, but consistency is seldom compelling as a practical argument.

Confidence intervals help compensate for possible frequentist overfitting. With $z = 3.29$ and $n = 10,000$, the 95% confidence interval for μ in model $\mathcal{M}_1$ (13.30) is $(0.013, 0.053)$. Whether or not such a small effect is interesting depends on the scientific context. The fact that BIC says "not interesting" speaks to its inherent small-model bias.

The prostate cancer study data of Section 3.3 provides a more challenging model selection problem. Figure 3.4 shows the histogram of $N = 6033$ observations x_i, each measuring the effects of one gene. The histogram has 49 bins, each of width 0.2, with centers c_j ranging from -4.4 to 5.2; y_j, the height of the histogram at c_j, is the number of x_i in bin j,

$$y_j = \#\{x_i \in \text{bin } j\} \qquad \text{for } j = 1, 2, \ldots, 49. \tag{13.54}$$

We assume that the y_j follow a Poisson regression model as in Section 8.3,

$$y_j \overset{\text{ind}}{\sim} \text{Poi}(\nu_j), \qquad j = 1, 2, \ldots, 49, \tag{13.55}$$

and wish to fit a log polynomial GLM model to the ν_j. The model selection question is "What degree polynomial?" Degree 2 corresponds to normal densities, but the long tails seen in Figure 3.4 suggest otherwise.

Models of degree 2 through 8 are assessed in Figure 13.4. Four model selection measures are compared: AIC (13.42); BIC (13.37) with $n = 49$,

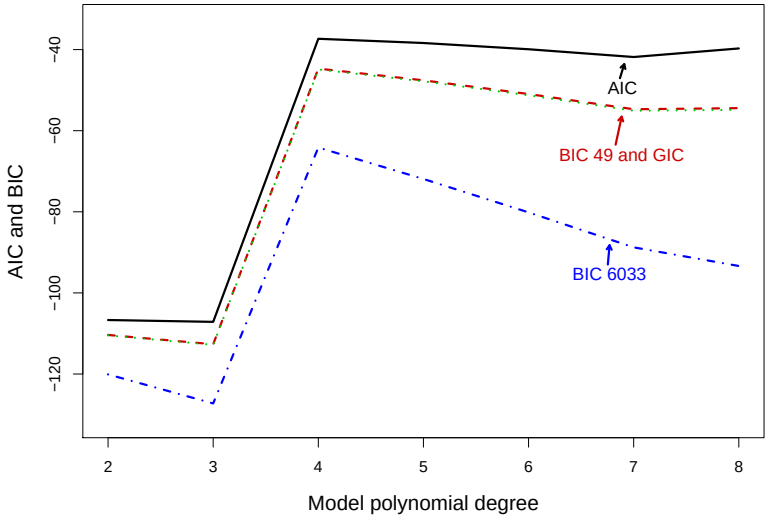

Figure 13.4 Log polynomial models of degree 2 through 8 applied to the prostate study histogram of Figure 3.4. Model selection criteria: AIC (13.42); BIC (13.37) with $n = 49$, number of bins, or 6033, number of genes; GIC (13.50) using classic Fisher hypothesis choice $c_n = 1.92$. All four selected the fourth-degree model as best.

the number of y_j values (bins), and also $n = 6033$, the number of genes; and GIC (13.50), with $c_n = 1.92$ (13.53), the choice based on classic Fisherian hypothesis testing. (This is almost the same as BIC $n = 49$, since $\log(49)/2 = 1.95$.) A fourth-degree polynomial model was the winner under all four criteria.

The "untethered" criticism made against objective Bayes methods in general is particularly applicable to BIC. The concept of "sample size" is not well defined, as the prostate study example shows. Sample size coherency (13.49), the rationale for BIC's strong bias toward smaller models, is less convincing in the absence of priors based on genuine experience (especially if there is no prospect of the sample size changing). Whatever its vulnerabilities, BIC model selection has nevertheless become a mainstay of objective Bayes model selection, not least because of its freedom from the choice of Bayesian priors.

13.4 Gibbs Sampling and MCMC

Miraculously blessed with visions of the future, a Bayesian statistician of the 1970s would certainly be pleased with the prevalence of Bayes methodology in twenty-first-century applications. But his pleasure might be tinged with surprise that the applications were mostly of the objective, "uninformative" type, rather than taken from the elegant de Finetti–Savage school of subjective inference.

The increase in Bayesian applications, and the change in emphasis from subjective to objective, had more to do with computation than philosophy. Better computers and algorithms facilitated the calculation of formerly intractable Bayes posterior distributions. Technology determines practice, and the powerful new algorithms encouraged Bayesian analyses of large and complicated models where subjective priors (or those based on actual past experience) were hard to come by. Add in the fact that the algorithms worked most easily with simple "convenience" priors like the conjugates of Section 13.2, and the stage was set for an objective Bayes renaissance.

At first glance it's hard to see why Bayesian computations should be daunting. From parameter vector $\boldsymbol{\theta}$, data $\boldsymbol{x}$, density function $f_{\boldsymbol{\theta}}(\boldsymbol{x})$, and prior density $g(\boldsymbol{\theta})$, Bayes' rule (3.5)–(3.6) directly produces the posterior density

$$g(\boldsymbol{\theta}|\boldsymbol{x}) = g(\boldsymbol{\theta})\, f_{\boldsymbol{\theta}}(\boldsymbol{x})/f(\boldsymbol{x}), \tag{13.56}$$

where $f(\boldsymbol{x})$ is the marginal density

$$f(\boldsymbol{x}) = \int_{\Omega} g(\boldsymbol{\theta})\, f_{\boldsymbol{\theta}}(\boldsymbol{x})\, d\boldsymbol{\theta}. \tag{13.57}$$

The posterior probability of any set A in the parameter space Ω is then

$$P\{A|\boldsymbol{x}\} = \int_{A} g(\boldsymbol{\theta})\, f_{\boldsymbol{\theta}}(\boldsymbol{x})\, d\boldsymbol{\theta} \Big/ \int_{\Omega} g(\boldsymbol{\theta})\, f_{\boldsymbol{\theta}}(\boldsymbol{x})\, d\boldsymbol{\theta}. \tag{13.58}$$

This is easy to write down but usually difficult to evaluate if $\boldsymbol{\theta}$ is multidimensional.

Modern Bayes methods attack the problem through the application of computer power. Even if we can't integrate $g(\boldsymbol{\theta}|\boldsymbol{x})$, perhaps we can *sample* from it. If so, a sufficiently large sample, say

$$\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(B)} \sim g(\boldsymbol{\theta}|\boldsymbol{x}) \tag{13.59}$$

would provide estimates

$$\hat{P}\{A|\boldsymbol{x}\} = \#\left\{\boldsymbol{\theta}^{(j)} \in A\right\} \Big/ B, \tag{13.60}$$

and similarly for posterior moments, correlations, etc. We would in this way be employing the same general tactic as the bootstrap, applied now for Bayesian rather than frequentist purposes—toward the same goal as the bootstrap, of freeing practical applications from the constraints of mathematical tractability.

The two most popular computational methods,[6] *Gibbs sampling* and *Markov chain Monte Carlo* (MCMC), are based on Markov chain algorithms; that is, the posterior samples $\boldsymbol{\theta}^{(b)}$ are produced in sequence, each one depending only on $\boldsymbol{\theta}^{(b-1)}$ and not on its more distant predecessors. We begin with Gibbs sampling.

The central idea of Gibbs sampling is to reduce the generation of multidimensional vectors $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ to a series of univariate calculations. Let $\boldsymbol{\theta}_{(k)}$ denote $\boldsymbol{\theta}$ with component k removed, and $g_{(k)}$ the conditional density of θ_k given $\boldsymbol{\theta}_{(k)}$ and the data $\boldsymbol{x}$,

$$\theta_k|\boldsymbol{\theta}_{(k)}, \boldsymbol{x} \sim g_{(k)}\Big(\theta_k|\boldsymbol{\theta}_{(k)}, \boldsymbol{x}\Big). \tag{13.61}$$

The algorithm begins at some arbitrary initial value $\boldsymbol{\theta}^{(0)}$. Having computed $\boldsymbol{\theta}^{(1)}$, $\boldsymbol{\theta}^{(2)}$, $\ldots$, $\boldsymbol{\theta}^{(b-1)}$, the components of $\boldsymbol{\theta}^{(b)}$ are generated according to conditional distributions (13.61),

$$\theta_k^{(b)} \sim g_{(k)}\left(\theta_k\big|\boldsymbol{\theta}_{(k)}^{(b-1)}, \boldsymbol{x}\right) \qquad \text{for } k = 1, 2, \ldots, K. \tag{13.62}$$

As an example, we take $\boldsymbol{x}$ to be the $n = 20$ observations for $y = 1$ in the vasoconstriction data of Table 13.2, and assume that these are a normal sample,

$$x_i \overset{\text{iid}}{\sim} \mathcal{N}(\mu, \tau), \qquad i = 1, 2, \ldots, n = 20. \tag{13.63}$$

The sufficient statistics for estimating the bivariate parameter $\theta = (\mu, \tau)$ are the sample mean and variance

$$\bar{x} = \sum_1^n x_i/n \quad \text{and} \quad T = \sum_1^n (x_i - \bar{x})^2/(n-1), \tag{13.64}$$

having independent normal and gamma distributions,

$$\bar{x} \sim \mathcal{N}(\mu, \tau/n) \quad \text{and} \quad T \sim \tau G_\nu/\nu, \tag{13.65}$$

with $\nu = \frac{n-1}{2}$, the latter being $\text{Gam}(\nu, \tau/\nu)$ in the notation of Table 5.1.

[6] The two methods are often referred to collectively as MCMC because of mathematical connections, with "Metropolis-Hasting algorithm" referring to the second type of procedure.

For our Bayes prior distribution we take the conjugates

$$\tau \sim k_1\tau_1/G_{k_1+1} \quad \text{and} \quad \mu|\tau \sim \mathcal{N}(\mu_0, \tau/n_0). \tag{13.66}$$

In terms of Table 13.1, $(x_0, n_0\nu) = (\tau_1, k_1)$ for the gamma, while $(x_0, \sigma_1^2) = (\mu_0, \tau)$ for the normal. (A simple specification would take $\mu \sim \mathcal{N}(\mu_0, \tau_1/n_0)$.)

Multiplying the normal and gamma functional forms in Table 5.1 yields density function

$$f_{\mu,\tau}(\bar{x}, T) = c\tau^{-(\nu+\frac{1}{2})} \exp\left\{-\frac{1}{\tau}\left[\nu T + \frac{n}{2}(\bar{x}-\mu)^2\right]\right\} \tag{13.67}$$

and prior density

$$g(\mu, \tau) = c\tau^{-(k_1+2.5)} \exp\left\{-\frac{1}{\tau}\left[k_1\tau_1 + \frac{n_0}{2}(\mu-\mu_0)^2\right]\right\}, \tag{13.68}$$

c indicating positive constants that do not affect the posterior computations. The posterior density $cg(\mu, \tau) f_{\mu,\tau}(\bar{x}, T)$ is then calculated to be

$$\begin{gathered} g(\mu, \tau|\bar{x}, T) = c\tau^{-(\nu+k_1+3)} \exp\{-Q/\tau\}, \\ \text{where } Q = (k_1\tau_1 + T) + \frac{n_+}{2}(\mu-\bar{\mu}_+)^2 + \frac{n_0 n}{2n_+}(\mu_0-\bar{x})^2. \end{gathered} \tag{13.69}$$

Here $n_+ = n_0 + n$ and $\bar{\mu}_+ = (n_0\mu_0 + n\bar{x})/n_+$.

In order to make use of Gibbs sampling we need to know the *full conditional* distributions $g(\mu|\tau, \bar{x}, T)$ and $g(\tau|\mu, \bar{x}, T)$, as in (13.62). (In this case, $k = 2$, $\theta_1 = \mu$, and $\theta_2 = \tau$.) This is where the conjugate expressions in Table 13.1 come into play. Inspection of density (13.69) shows that

$$\mu|\tau, \bar{x}, T \sim \mathcal{N}\left(\bar{\mu}_+, \frac{\tau}{n_+}\right) \quad \text{and} \quad \tau|\mu, \bar{x}, T \sim \frac{Q}{G_{\nu+k_1+2}}. \tag{13.70}$$

$B = 10,000$ Gibbs samples $\boldsymbol{\theta}^{(b)} = (\mu^{(b)}, \tau^{(b)})$ were generated starting from $\boldsymbol{\theta}^{(0)} = (\bar{x}, T) = (116, 554)$. The prior specifications were chosen to be (presumably) uninformative or mildly informative,

$$n_0 = 1, \quad \mu_0 = \bar{x}, \quad k_1 = 1 \text{ or } 9.5, \quad \text{and} \quad \tau_1 = T. \tag{13.71}$$

(In which case $\bar{\mu}_+ = \bar{x}$ and $Q = (\nu + k_1)T + n_+(\mu - \bar{x})^2$. From $\nu = (n-1)/2$, we see that k_1 corresponds to about $2k_1$ hypothetical prior observations.) The resulting posterior distributions for τ are shown by the histograms in Figure 13.5.

As a point of frequentist comparison, $B = 10,000$ parametric bootstrap replications (which involve no prior assumptions),

$$\hat{\tau}^* \sim \hat{\tau}G_\nu/\nu, \qquad \hat{\tau} = T, \tag{13.72}$$

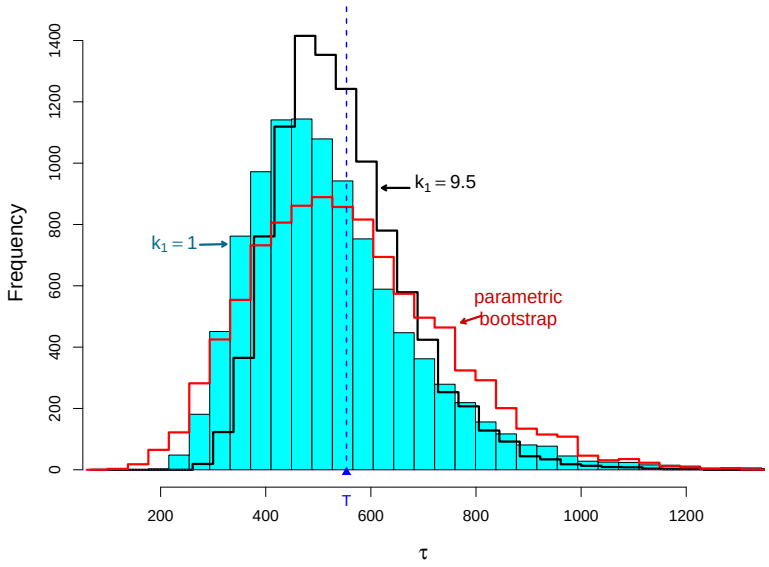

Figure 13.5 Posterior distributions for variance parameter τ, model (13.63)–(13.65), volume of air inspired for vasoconstriction group $y = 1$ from Table 13.2. Solid teal histogram: $B = 10,000$ Gibbs samples with $k_1 = 1$; black line histogram: $B = 10,000$ samples with $k_1 = 9.5$; red line histogram: 10,000 parametric bootstrap samples (13.72) suggests even the $k_1 = 1$ prior has substantial posterior effect.

are seen to be noticeably more dispersed than even the $k_1 = 1$ Bayes posterior distribution, the likely choice for an objective Bayes analysis. Bayes techniques, even objective ones, have regularization effects that may or may not be appropriate.

A similar, independent Gibbs sample of size 10,000 was obtained for the 19 $y = 0$ vasoconstriction measurements in Table 13.2, with specifications as in (13.71), $k = 1$. Let

$$\delta^{(b)} = \frac{\mu_1^{(b)} - \mu_0^{(b)}}{\left(\tau_1^{(b)} + \tau_0^{(b)}\right)^{1/2}}, \tag{13.73}$$

where $(\mu_1^{(b)}, \tau_1^{(b)})$ and $(\mu_0^{(b)}, \tau_0^{(b)})$ denote the bth Gibbs samples from the $y = 1$ and $y = 0$ runs.

Figure 13.6 shows the posterior distribution of δ. Twenty-eight of the

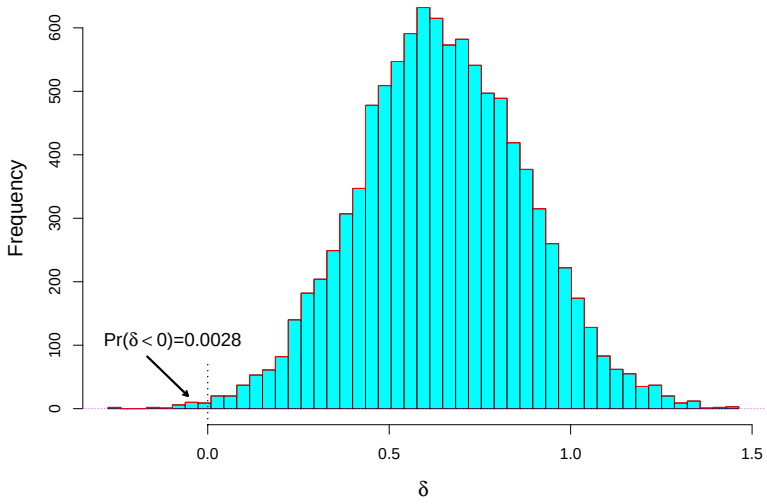

Figure 13.6 $B = 10{,}000$ Gibbs samples for "Bayes t-statistic" (13.73) comparing $y = 1$ with $y = 0$ values for vasoconstriction data.

$B = 10{,}000$ values $\delta^{(b)}$ were less than 0, giving a "Bayesian t-test" estimate

$$P\{\delta < 0 | \bar{x}_1, \bar{x}_0, T_1, T_0\} = 0.0028. \tag{13.74}$$

(The usual t-test yielded one-sided p-value 0.0047 against the null hypothesis $\mu_0 = \mu_1$.) An appealing feature of Gibbs sampling is that having obtained $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(B)}$ (13.59) the posterior distribution of any parameter $\gamma = t(\boldsymbol{\theta})$ is obtained directly from the B values $\gamma^{(b)} = t(\boldsymbol{\theta}^{(b)})$.

Gibbs sampling requires the ability to sample from the full conditional distributions (13.61). A more general Markov chain Monte Carlo method, commonly referred to as MCMC, makes clearer the basic idea. Suppose the space of possible $\boldsymbol{\theta}$ values is finite, say $\{\boldsymbol{\theta}(1), \boldsymbol{\theta}(2), \ldots, \boldsymbol{\theta}(M)\}$, and we wish to simulate samples from a posterior distribution putting probability $p(i)$ on $\boldsymbol{\theta}(i)$,

$$\boldsymbol{p} = (p(1), p(2), \ldots, p(M)). \tag{13.75}$$

The MCMC algorithm begins with the choice of a "candidate" probability distribution $q(i, j)$ for moving from $\boldsymbol{\theta}(i)$ to $\boldsymbol{\theta}(j)$; in theory $q(i, j)$ can be almost anything, for instance $q(i, j) = 1/(M - 1)$ for $j \neq i$. The simulated samples $\boldsymbol{\theta}^{(b)}$ are obtained by a random walk: if $\boldsymbol{\theta}^{(b)}$ equals $\boldsymbol{\theta}(i)$,

then $\boldsymbol{\theta}^{(b+1)}$ equals $\boldsymbol{\theta}(j)$ with probability[7]

$$Q(i,j) = q(i,j) \cdot \min\left\{\frac{p(j)q(j,i)}{p(i)q(i,j)}, 1\right\} \tag{13.76}$$

for $j \neq i$, while with probability

$$Q(i,i) = 1 - \sum_{j \neq i} Q(i,j) \tag{13.77}$$

$\boldsymbol{\theta}^{(b+1)} = \boldsymbol{\theta}^{(b)} = \boldsymbol{\theta}(i)$. Markov chain theory then says that, under quite general conditions, the empirical distribution of the random walk values $\boldsymbol{\theta}^{(b)}$ will approach the desired distribution $\boldsymbol{p}$ as b gets large.

A heuristic argument for why this happens begins by supposing that $\boldsymbol{\theta}^{(1)}$ was in fact generated by sampling from the target distribution $\boldsymbol{p}$, $\Pr\{\boldsymbol{\theta}^{(1)} = i\} = p(i)$, and then $\boldsymbol{\theta}^{(2)}$ was obtained according to transition probabilities (13.76)–(13.77). A little algebra shows that (13.76) implies

$$p(i)Q(i,j) = p(j)Q(j,i), \tag{13.78}$$

the so-called *balance equations*. This results in

$$\begin{aligned}\Pr\left\{\boldsymbol{\theta}^{(2)} = i\right\} &= p(i)Q(i,i) + \sum_{j \neq i} p(j)Q(j,i) \\ &= p(i)\sum_{j=1}^{M} Q(i,j) = p(i).\end{aligned} \tag{13.79}$$

In other words, if $\boldsymbol{\theta}^{(1)}$ has distribution $\boldsymbol{p}$ then so will $\boldsymbol{\theta}^{(2)}$, and likewise $\boldsymbol{\theta}^{(3)}, \boldsymbol{\theta}^{(4)}, \ldots$; $\boldsymbol{p}$ is the *equilibrium distribution* of the Markov chain random walk defined by transition probabilities Q. Under reasonable conditions,† $\boldsymbol{\theta}^{(b)}$ must asymptotically attain distribution $\boldsymbol{p}$ no matter how $\boldsymbol{\theta}^{(1)}$ is initially selected.

†9

13.5 Example: Modeling Population Admixture

MCMC has had a big impact in statistical genetics, where Bayesian modeling is popular and useful for representing the complex evolutionary processes. Here we illustrate its use in demography and modeling *admixture*—estimating the contributions from ancestral populations in an individual

[7] In Bayes applications, $p(i) = g(\boldsymbol{\theta}(i)|\boldsymbol{x}) = g(\boldsymbol{\theta}(i))f_{\boldsymbol{\theta}(i)}(\boldsymbol{x})/f(\boldsymbol{x})$ (13.56). However, $f(\boldsymbol{x})$ is not needed since it cancels out of (13.76), a considerable advantage in complicated situations when $f(\boldsymbol{x})$ is often unavailable, and a prime reason for the popularity of MCMC.

genome. For example, we might consider human ancestry, and for each individual wish to estimate the proportion of their genome coming from **European**, **African**, and **Asian** origins. The procedure we describe here is unsupervised—a type of soft clustering—but we will see it can be very informative with regard to such questions. We have a sample of n individuals, and we assume each arose from possible admixture among J parent populations, each with their own characteristic vector of allele frequencies. For us $J = 3$, and let $Q_i \in \mathcal{S}_3$ denote a probability vector for individual i representing the proportions of their heritage coming from populations $j \in \{1, 2, 3\}$ (see Section 5.4). We have genomic measurements for each individual, in our case SNPs (single-nucleotide polymorphisms) at each of M well-spaced loci, and hence can assume they are in linkage equilibrium. At each SNP we have a measurement that identifies the two alleles (one per chromosome), where each can be either the wild-type A or the mutation a. That is, we have the genotype G_{im} at SNP m for individual i: a three-level factor with levels {**AA**, **Aa**, **aa**} which we code as $0, 1, 2$. Table 13.6 shows some examples.

Table 13.6 *A subset of the genotype data on 197 individuals, each with genotype measurements at 100 SNPs. In this case the* **ethnicity** *is known for each individual, one of* **Japanese**, **African**, **European**, *or* **African American**. *For example, individual* **NA12239** *has* **genotype Aa** *for SNP1,* **NA19247** *has* **AA**, *and* **NA20126** *has* **aa**.

Subject	SNP_1	SNP_2	SNP_3	$\cdots$	SNP_{97}	SNP_{98}	SNP_{99}	SNP_{100}
NA10852	1	1	0	$\cdots$	1	1	0	0
NA12239	1	1	0	$\cdots$	1	1	0	0
NA19072	0	0	0	$\cdots$	0	0	0	0
NA19247	0	0	2	$\cdots$	0	0	0	2
NA20126	2	0	0	$\cdots$	2	0	0	0
NA18868	0	0	1	$\cdots$	0	0	0	1
NA19257	0	0	0	$\cdots$	0	0	0	0
NA19079	0	1	0	$\cdots$	0	1	0	0
NA19067	0	0	0	$\cdots$	0	0	0	0
NA19904	0	0	1	$\cdots$	0	0	0	1

Let P_j be the (unknown) M-vector of minor allele frequencies (proportions actually) in population j. We have available a sample of n individuals, and for each sample we have their genomic information measured at each of the M loci. Some of the individuals might appear to have pure ancestral origins, but many do not. Our goal is to estimate $Q_i,\ i = 1, \ldots, n$, and $P_j,\ j \in \{1, 2, 3\}$.

For this purpose it is useful to pose a generative model. We first create a pair of variables $X_{im} = (X_{im}^{(1)}, X_{im}^{(2)})$ corresponding to each G_{im}, to which we allocate the two alleles (in arbitrary order). For example, if $G_{im} = 1$ (corresponding to Aa), then we might set $X_{im}^{(1)} = 0$ and $X_{im}^{(2)} = 1$ (or vice versa). If $G_{im} = 0$ they are both 0, and if $G_{im} = 2$, they are both 1. Let $Z_{im} \in \{1, 2, 3\}^2$ represent the ancestral origin for individual i of each of these allele copies X_{im} at locus m, again a two-vector with elements $Z_{im} = (Z_{im}^{(1)}, Z_{im}^{(2)})$. Then our generative model goes as follows.

1 $Z_{im}^{(c)} \sim \text{Mult}(1, Q_i)$, independently at each m, for each copy $c = 1, 2$. That is, we select the ancestral origin of each chromosome at locus m according to the individual's mixture proportions Q_i.
2 $X_{im}^{(c)} \sim \text{Bi}(1, P_{jm})$ if $Z_{im}^{(c)} = j$, for each copy $c = 1, 2$. What this means is that, for each of the two ancestral picks at locus m (one for each arm of the chromosome), we draw a binomial with the appropriate allele frequency.

To complete the Bayesian specification, we need to supply priors for the Q_i and also for P_{jm}. Although one can get fancy here, we resort to the recommended flat priors, which are

- $Q_i \sim D(\lambda, \lambda, \lambda)$, a flat three-component Dirichlet, independently for each subject i [†10]
- $P_{jm} \sim D(\gamma, \gamma)$ independently for each population j, and each locus m (the beta distribution; see †10 in the end notes).

We use the least-informative values $\lambda = \gamma = 1$. In practice, these could get updated as well, but for the purposes of this demonstration we leave them fixed at these values.

Let $\boldsymbol{X}$ be the $n \times M \times 2$ array of observed alleles for all n samples. We wish to estimate the posterior distribution $\Pr(P, Q|\boldsymbol{X})$, referring collectively to all the elements of P and Q.

For this purpose we use Gibbs sampling, which amounts to the following sequence.

0 Initialize $Z^{(0)}$, $P^{(0)}$, $Q^{(0)}$.
1 Sample $Z^{(b)}$ from the conditional distribution $\Pr(Z|\boldsymbol{X}, P^{(b-1)}, Q^{(b-1)})$.
2 Sample $P^{(b)}, Q^{(b)}$ from the conditional distribution $\Pr(P, Q|\boldsymbol{X}, Z^{(b)})$.

Gibbs is effective when one can sample efficiently from these conditional distributions, which is the case here.

In step 2, we can sample P and Q separately. It can be seen that for each (j, m) we should sample P_{jm} from

$$P_{jm}|\boldsymbol{X}, Z \sim D(\lambda + n_{jm}^{(0)},\ \lambda + n_{jm}^{(1)}), \tag{13.80}$$

where $Z = Z^{(b)}$ and

$$\begin{aligned} n_{jm}^{(0)} &= \#\{(i,c)\ :\ X_{im}^{(c)} = 0 \text{ and } Z_{im}^{(c)} = j\}, \\ n_{jm}^{(1)} &= \#\{(i,c)\ :\ X_{im}^{(c)} = 1 \text{ and } Z_{im}^{(c)} = j\}. \end{aligned} \tag{13.81}$$

This follows from the conjugacy of the two-component Dirichlet (beta) with the binomial distribution, Table 13.1.

Updating Q_i involves simulating from

$$Q_i|\boldsymbol{X}, Z \sim D(\gamma + m_{i1}, \gamma + m_{i2}, \gamma + m_{i3}), \tag{13.82}$$

where m_{ij} is the number of allele copies in individual i that originated (according to $Z = Z^{(b)}$) in population j:

$$m_{ij} = \#\{(c,m)\ :\ Z_{im}^{(c)} = j\}. \tag{13.83}$$

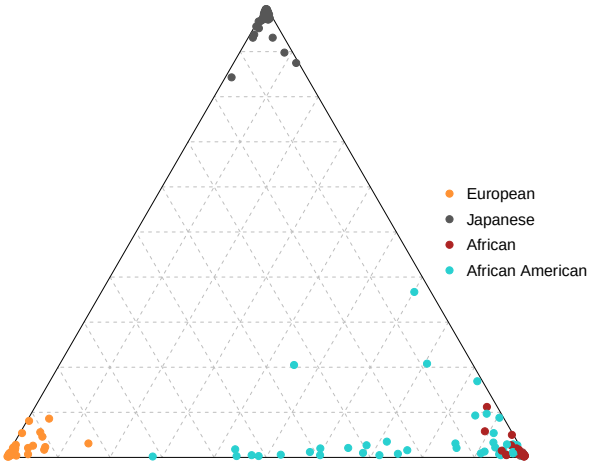

Figure 13.7 Barycentric coordinate plot for the estimated posterior means of the Q_i based on MCMC sampling.

Step 1 can be performed by simulating $Z_{im}^{(c)}$ independently, for each i, m,

and c from

$$\Pr(Z_{im}^{(c)} = j \,|\, \boldsymbol{X}, P, Q) = \frac{Q_{ij} \Pr(X_{im}^{(c)} | P, Z_{im}^{(c)} = j)}{\sum_{\ell=1}^{3} Q_{i\ell} \Pr(X_{im}^{(c)} | P, Z_{im}^{(c)} = \ell)}. \quad (13.84)$$

The probabilities on the right refer back to our generative distribution described earlier.

Figure 13.7 shows a triangle plot that summarizes the result of running the MCMC algorithm on our 197 subjects. We used a burn in of 1000 complete iterations, and then a further 2000 to estimate the distribution of the parameters of interest, in this case the Q_i. Each dot in the figure represents a three-component probability vector, and is the posterior mean of the sampled Q_i for each subject. The points are colored according to the known ethnicity. Although this algorithm is unsupervised, we see that the ethnic groups cluster nicely in the corners of the simplex, and allow us to identify these clusters. The `African American` group is spread between the `African` and `European` clusters (with a little movement toward the `Japanese`).

Markov chain methods are versatile tools that have proved their value in Bayesian applications. There are some drawbacks.

- The algorithms are not universal in the sense of maximum likelihood, requiring some individual ingenuity with each application.
- As a result, applications, especially of Gibbs sampling, have favored a small set of convenient priors, mainly Jeffreys and conjugates, that simplify the calculations. This can cast doubt on the relevance of the resulting Bayes inferences.
- Successive realizations $\boldsymbol{\theta}^{(b)}$ are highly correlated with each other, making the convergence of estimates such as $\bar{\boldsymbol{\theta}} = \sum \boldsymbol{\theta}^{(b)}/B$ slow.
- The correlation makes it difficult to assign a standard error to $\bar{\boldsymbol{\theta}}$. Actual applications ignore an initial B_0 of the $\boldsymbol{\theta}^{(b)}$ values (as a "burn-in" period) and go on to large enough B such that estimates like $\bar{\boldsymbol{\theta}}$ appear to settle down. However, neither the choice of B_0 nor that of B may be clear.

Objective Bayes offers a paradigm of our book's theme, the effect of electronic computation on statistical inference: ingenious new algorithms facilitated Bayesian applications over a wide class of applied problems and, in doing so, influenced the dominant philosophy of the whole area.

13.6 Notes and Details

The books by Savage (1954) and de Finetti (1972), summarizing his earlier work, served as foundational texts for the subjective Bayesian school of inference. Highly influential, they championed a framework for Bayesian applications based on coherent behavior and the careful elucidation of personal probabilities. A current leading text on Bayesian methods, Carlin and Louis (2000), does not reference either Savage or de Finetti. Now Jeffreys (1961), again following earlier works, claims foundational status. The change of direction has not gone without protest from the subjectivists—see Adrian Smith's discussion of O'Hagan (1995)—but is nonetheless almost a complete rout.

Metropolis *et al.* (1953), as part of nuclear weapons research, developed the first MCMC algorithm. A vigorous line of work on Markov chain methods for solving difficult probability problems has continued to flourish under such names as particle filtering and sequential Monte Carlo; see Gerber and Chopin (2015) and its enthusiastic discussion.

Modeling population admixture (Pritchard *et al.*, 2000) is one of several applications of hierarchical Bayesian models and MCMC in genetics. Other applications include haplotype estimation and motif finding, as well as estimation of phylogenetic trees. The examples in this section were developed with the kind help of Hua Tang and David Golan, both from the Stanford Genetics department. Hua suggested the example and provided helpful guidance; David provided the data, and ran the MCMC algorithm using the STRUCTURE program in the Pritchard lab.

†1 [p. 236] *Uninformative priors.* A large catalog of possible uninformative priors has been proposed, thoroughly surveyed by Kass and Wasserman (1996). One approach is to use the likelihood from a small part of the data, say just one or two data points out of n, as the prior, as with the "intrinsic priors" of Berger and Pericchi (1996), or O'Hagan's (1995) "fractional Bayes factors." Another approach is to minimize some mathematical measure of prior information, as with Bernardo's (1979) "reference priors" or Jaynes' (1968) "maximum entropy" criterion. Kass and Wasserman list a dozen more possibilities.

†2 [p. 236] *Coverage matching priors.* Welch and Peers (1963) showed that, for a multiparameter family $f_\mu(x)$ and real-valued parameter of interest $\theta = t(\mu)$, there exist priors $g(\mu)$ such that the Bayes credible interval of coverage α has frequentist coverage $\alpha + O(1/n)$, with n the sample size. In other words, the credible intervals are "second-order accurate" confidence intervals. Tibshirani (1989), building on Stein's (1985) work, produced the

nice formulation (13.9). Stein's paper developed the *least-favorable family*, the one-parameter subfamily of $f_\mu(x)$ that does not inappropriately increase the amount of Fisher information for estimating θ. Cox and Reid's (1987) *orthogonal parameters* form (13.8) is formally equivalent to the least favorable family construction.

Least favorable family versions of reference priors and intrinsic priors have been proposed to avoid the difficulty with general-purpose uninformative priors seen in Figure 11.7. They do so, but at the price of requiring a different prior for each choice of $\theta = t(\mu)$—which begins to sound more frequentistic than Bayesian.

†3 [p. 238] *Conjugate families theorem.* Theorem 13.1, (13.16)–(13.18), is rigorously derived in Diaconis and Ylvisaker (1979). Families other than (13.14) have conjugate-like properties, but not the neat posterior expectation result (13.18).

†4 [p. 238] *Poisson formula* (13.20). This follows immediately from (13.14), using $\alpha = \log(\mu)$, $\psi(\alpha) = \mu$, and $V(\mu) = \mu$ for the Poisson.

†5 [p. 239] *Inverse gamma and chi-square distributions.* A G_ν variate (13.22) has density $\mu^{\nu-1}e^{-\mu}/\Gamma(\nu)$. An *inverse gamma* variate $1/G_\nu$ has density $\mu^{-(\nu+1)}\, e^{-1/\mu}/\Gamma(\nu)$, so

$$g_{n_0,x_0}(\mu) = c\mu^{-(n_0x_0+2)}e^{-n_0x_0\nu/\mu} \tag{13.85}$$

is the gamma conjugate density in Table 13.1. The gamma results can be restated in terms of chi-squared variates:

$$x_i \sim \mu\frac{\chi^2_m}{m} = \mu\frac{G_{m/2}}{m/2} \tag{13.86}$$

has conjugate prior

$$g_{n_0,x_0}(\mu) \sim n_0x_0m/\chi^2_{n_0m+2}, \tag{13.87}$$

an *inverse chi-squared* distribution.

†6 [p. 240] *Vasoconstriction data.* Efron and Gous (2001) use this data to illustrate a theory connecting Bayes factors with Fisherian hypothesis testing. It is part of a larger data set appearing in Finney (1947), also discussed in Kass and Raftery (1995).

†7 [p. 245] *Jeffreys' and Fisher's scales of evidence.* Jeffreys' scale as it appears in Table 13.3 is taken from the slightly amended form in Kass and Raftery (1995). Efron and Gous (2001) compare it with Fisher's scale for the contradictory results of Table 13.5. Fisher and Jeffreys worked in different scientific contexts—small-sample agricultural experiments versus

hard-science geostatistics—which might explain Jeffreys' more stringent conception of what constitutes significant evidence.

†8 [p. 246] *The Bayesian information criterion.* The BIC was proposed by Schwarz (1978). Kass and Wasserman (1996) provide an extended discussion of the BIC and model selection. "Proofs" of (13.37) ultimately depend on sample size coherency (13.49), as in Efron and Gous (2001). Quotation marks are used here to indicate the basically qualitative nature of BIC: if we think of the data points as being collected in pairs then n becomes $n/2$ in (13.38), etc., so it doesn't pay to put too fine a point on the criterion.

†9 [p. 256] *MCMC convergence.* Suppose we begin the MCMC random walk (13.76)–(13.77) by choosing $\boldsymbol{\theta}^{(1)}$ according to some arbitrary starting distribution $\boldsymbol{p}^{(1)}$. Let $\boldsymbol{p}^{(b)}$ be the distribution of $\boldsymbol{\theta}^{(b)}$, obtained after b steps of the random walk. Markov chain theory says that, under certain broad conditions on $Q(i, j)$, $\boldsymbol{p}^{(b)}$ will converge to the target distribution $\boldsymbol{p}$ (13.75). Moreover, the convergence is geometric in the L_1 norm $\sum |p_k^{(b)} - p_k|$, successive discrepancies eventually decreasing by a multiplicative factor. A proof appears in Tanner and Wong (1987). Unfortunately, the factor won't be known in most applications, and the actual convergence may be quite slow.

†10 [p. 258] *Dirichlet distribution.* The Dirichlet is a multivariate generalization of the beta distribution (Section 5.1), typically used to represent prior distributions for the multinomial distribution. For $x = (x_1, x_2, \ldots, x_k)'$, with $x_j \in (0, 1)$, $\sum_j x_j = 1$, the $D(\nu)$ density is defined as

$$f_\nu(x) = \frac{1}{B(\nu)} \prod_{j=1}^{k} x_j^{\nu_j - 1}, \tag{13.88}$$

where $B(\nu) = \prod_j \Gamma(\nu_j) / \Gamma(\sum_j \nu_j)$.