Preface

Though there are many recent additions to graduate-level introductory books on Bayesian analysis, none has quite our blend of theory, methods, and applications. We believe a beginning graduate student taking a Bayesian course or just trying to find out what it means to be a Bayesian ought to have some familiarity with all three aspects. More specialization can come later.

Each of us has taught a course like this at Indian Statistical Institute or Purdue. In fact, at least partly, the book grew out of those courses. We would also like to refer to the review (Ghosh and Samanta (2002b)) that first made us think of writing a book. The book contains somewhat more material than can be covered in a single semester. We have done this intentionally, so that an instructor has some choice as to what to cover as well as which of the three aspects to emphasize. Such a choice is essential for the instructor. The topics include several results or methods that have not appeared in a graduate text before. In fact, the book can be used also as a second course in Bayesian analysis if the instructor supplies more details.

Chapter 1 provides a quick review of classical statistical inference. Some knowledge of this is assumed when we compare different paradigms. Following this, an introduction to Bayesian inference is given in Chapter 2 emphasizing the need for the Bayesian approach to statistics. Objective priors and objective Bayesian analysis are also introduced here. We use the terms *objective* and *nonsubjective* interchangeably. After briefly reviewing an axiomatic development of utility and prior, a detailed discussion on Bayesian robustness is provided in Chapter 3. Chapter 4 is mainly on convergence of posterior quantities and large sample approximations. In Chapter 5, we discuss Bayesian inference for problems with low-dimensional parameters, specifically objective priors and objective Bayesian analysis for such problems. This covers a whole range of possibilities including uniform priors, Jeffreys' prior, other invariant objective priors, and reference priors. After this, in Chapter 6 we discuss some aspects of testing and model selection, treating these two problems as equivalent. This mostly involves Bayes factors and bounds on these computed over large classes of priors. Comparison with classical P-value is also made whenever appropriate. Bayesian P-value and nonsubjective Bayes factors such as the intrinsic and fractional Bayes factors are also introduced.

Chapter 7 is on Bayesian computations. Analytic approximation and the E-M algorithm are covered here, but most of the emphasis is on Markov chain based Monte Carlo methods including the M-H algorithm and Gibbs sampler, which are currently the most popular techniques. Follwing this, in Chapter 8 we cover the Bayesian approach to some standard problems in statistics. The next chapter covers more complex problems, namely, hierarchical Bayesian (HB) point and interval estimation in high-dimensional problems and parametric empirical Bayes (PEB) methods. Superiority of HB and PEB methods to classical methods and advantages of HB methods over PEB methods are discussed in detail. Akaike information criterion (AIC), Bayes information criterion (BIC), and other generalized Bayesian model selection criteria, high-dimensional testing problems, microarrays, and multiple comparisons are also covered here. The last chapter consists of three major methodological applications along with the required methodology.

We have marked those sections that are either very technical or are very specialized. These may be omitted at first reading, and also they need not be part of a standard one-semester course.

Several problems have been provided at the end of each chapter. More problems and other material will be placed at `http://www.isical.ac.in/~tapas/book`

Many people have helped – our mentors, both friends and critics, from whom we have learnt, our family and students at ISI and Purdue, and the anonymous referees of the book. Special mention must be made of Arijit Chakrabarti for Sections 9.7 and 9.8, Sudipto Banerjee for Section 10.1, Partha P. Majumder for Appendix D, and Kajal Dihidar and Avranil Sarkar for help in several computations. We alone are responsible for our philosophical views, however tentatively held, as well as presentation.

Thanks to John Kimmel, whose encouragement and support, as well as advice, were invaluable.

Indian Statistical Institute and Purdue University *Jayanta K. Ghosh*
Indian Statistical Institute *Mohan Delampady*
Indian Statistical Institute *Tapas Samanta*
February 2006

9

High-dimensional Problems

Rather than begin by defining what is meant by high-dimensional, we begin with a couple of examples.

Example 9.1. (Stein's example) Let $N(\boldsymbol{\mu}_{p\times 1}, \Sigma_{p\times p} \equiv \sigma^2 I_{p\times p})$ be a p-variate normal population and $\boldsymbol{X}_i = (X_{i1}, \dots, X_{ip}),\ i = 1, \dots, n$ be n i.i.d. p-variate samples. Because $\Sigma = \sigma^2 I$, we may alternatively think of the data as p independent samples of size n from p univariate normal populations $N(\mu_j, \sigma^2)$, $j = 1, \dots, p$. The parameters of interest are the μ_j's. For convenience, we initially assume σ^2 is known. Usually, the number of parameters, p, is large and the sample size n is small compared with p. These have been called problems with large p, small n. Note that n in Stein's example is the sample size, if we think of the data as a p-variate sample of size n. However, we could also think of the data as univariate samples of size n from each of p univariate populations. Then the total sample size would be np. The second interpretation leads to a class of similar examples. Note that the observations are not exchangeable except in subgroups, in this sense one may call them partially exchangeable.

Example 9.2. Let $f(x|\mu_j), j = 1, \dots, p$, denote the densities for p populations, and $X_{ij}, i = 1, \dots, n, j = 1, \dots, p$ denote p samples of size n from these p populations. As in Example 9.1, $f(x|\mu_j)$ may contain additional common parameters. The object is to make inference about the μ_j's.

In several path-breaking papers Stein (1955), James and Stein (1960), Stein (1981), Robbins (1955, 1964), Efron and Morris (1971, 1972, 1973a, 1975) have shown classical objective Bayes or classical frequentist methods, e.g., maximum likelihood estimates, will usually be inappropriate here. See also Kiefer and Wolfowitz (1956) for applications to examples like those of Neyman and Scott (1948). These approaches are discussed in Sections 9.1 through 9.4, with stress on the parametric empirical Bayes (PEB) approach of Efron and Morris, as extended in Morris (1983).

It turns out that exchangeability of $\mu_1, \ldots, \mu_p$ plays a fundamental role in all these approaches. Under this assumption, there is a simple and natural Bayesian solution of the problem based on a hierarchical prior and MCMC. Much of the popularity of Bayesian methods is due to the fact that many new examples of this kind could be treated in a unified way.

Because of the fundamental role of exchangeability of μ_j's and the simplicity, at least in principle, of the Bayesian approach, we begin with these two topics in Section 9.1. This leads in a natural way to the PEB approach in Sections 9.2 and 9.3 and the frequentist approach in Section 9.4.

All the above sections deal with point or interval estimation. In Section 9.6 we deal with testing and multiple testing in high-dimensional problems, with an application to microarrays. High-dimensional inference is closely related to model selection in high-dimensional problems. A brief overview is presented in Sections 9.7 and 9.8. We discuss several general issues in Sections 9.5 and 9.9.

9.1 Exchangeability, Hierarchical Priors, Approximation to Posterior for Large p, and MCMC

We follow the notations of Example 9.1 and Example 9.2, i.e., we consider p similar but not identical populations with densities $f(x|\mu_1), \ldots, f(x|\mu_p)$. There is a sample of size n, $X_{1j}, \ldots, X_{nj}$, from the jth population. These p populations may correspond with p adjacent small areas with unknown per capita income $\mu_1, \ldots, \mu_p$, as in small area estimation, Ghosh and Meeden (1997, Chapters 4, 5). They could also correspond with p clinical trials in a particular hospital and μ_j, $j = 1, \ldots, p$, are the mean effects of the drug being tested. In all these examples, the different studied populations are related to each other. In Morris (1983), which we closely follow in Section 9.2, the p populations correspond to p baseball players and μ_j's are average scores. Other such studies are listed in Morris and Christiansen (1996).

In order to assign a prior distribution for the μ_j's, we model them as exchangeable rather than i.i.d. or just independent. An exchangeable, dependent structure is consistent with the assumption that the studies are similar in a broad sense, so they share many common elements.

On the other hand, independence may be unnecessarily restrictive and somewhat unintuitive in the sense that one would expect to have separate analysis for each sample if the μ_j's were independent and hence unrelated. However, to justify exchangeability one would need a particular kind of dependence. For example, Morris (1983) points out that the baseball players in his study were all hitters; his analysis would have been hard to justify if he had considered both hitters and pitchers.

Using a standard way of generating exchangeable random variables, we introduce a vector of hyperparameters $\boldsymbol{\eta}$ and assume μ_j's are i.i.d. $\pi(\mu|\boldsymbol{\eta})$ given $\boldsymbol{\eta}$. Typically, if $f(x|\boldsymbol{\mu})$ belongs to an exponential family, it is convenient to take $\pi(\mu|\boldsymbol{\eta})$ to be a conjugate prior. It can be shown that for even moderately large p – in the baseball example of Morris (1983), $p = 18$ – there is a lot of information in the data on $\boldsymbol{\eta}$. Hence the choice of a prior for $\boldsymbol{\eta}$ does not have much influence on inference about μ_j's. It is customary to choose a uniform or one of the other objective priors (vide Chapter 5) for $\boldsymbol{\eta}$.

We illustrate these ideas by exploring in detail Example 9.1.

Example 9.3. (Example 9.1, continued) Let $f(x|\mu_j)$ be the density of $N(\mu_j, \sigma^2)$ where we initially assume σ^2 is known. We relax this assumption in Subsection 9.1.1.

The prior for μ_j is taken to be $N(\eta_1, \eta_2)$ where η_1 is the prior guess about the μ_j's and η_2 is a measure of the prior uncertainty about this guess, vide Example 2.1, Chapter 2. The prior for η_1, η_2 is $\pi(\eta_1, \eta_2)$, which we specify a bit later.

Because $(\bar{X}_j = \sum_{i=1}^n X_{ij}/n, j = 1, \ldots, p)$ form a sufficient statistic and $\bar{X}_j$'s are independent, the Bayes estimate for μ_j under squared error loss is

$$E(\mu_j|\boldsymbol{X}) = E(\mu_j|\bar{\boldsymbol{X}}) = \int E(\mu_j|\bar{\boldsymbol{X}}, \boldsymbol{\eta})\pi(\boldsymbol{\eta}|\bar{\boldsymbol{X}})d\boldsymbol{\eta}.$$

where $\boldsymbol{X} = (X_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, p)$, $\bar{\boldsymbol{X}} = (\bar{X}_1, \ldots, \bar{X}_p)$ and (vide Example 2.1)

$$E(\mu_j|\boldsymbol{X}, \boldsymbol{\eta}) = E(\mu_j|\bar{X}_j, \boldsymbol{\eta}) = \frac{\eta_2\bar{X}_j + (\sigma^2/n)\eta_1}{\eta_2 + (\sigma^2/n)} = (1 - B)\bar{X}_j + B\eta_1, \quad (9.1)$$

with $B = (\sigma^2/n)/(\eta_2 + \sigma^2/n)$, depends on data only through $\bar{X}_j$.

The Bayes estimate of μ_j, on the other hand, depends on $\bar{X}_j$ as above and also on $(\bar{X}_1, \ldots, \bar{X}_p)$ because the posterior distribution of $\boldsymbol{\eta}$ depends on all the $\bar{X}_j$'s. Thus the Bayes estimate learns from the full sufficient statistic justifying simultaneous estimation of all the μ_j's. This learning process is sometimes referred to as borrowing strength. If the modeling of μ_j's is realistic, we would expect the Bayes estimates to perform better than the $\bar{X}_j$'s. This is what is strikingly new in the case of large p, small n and follows from the exchangeability of μ_j's.

The posterior density $\pi(\boldsymbol{\eta}|\boldsymbol{X})$ is also easy to calculate in principle. For known σ^2, one can get it explicitly.

Integrating out the μ_j's and holding $\boldsymbol{\eta}$ fixed, we get $\bar{X}_j$'s are independent and

$$\bar{X}_j|\boldsymbol{\eta} \sim N(\eta_1, \eta_2 + \sigma^2/n). \quad (9.2)$$

Let the prior density of (η_1, η_2) be constant on $\mathcal{R} \times \mathcal{R}^+$. (See Problem 1 and Gelman et al. (1995) to find out why some other choices like $\pi(\eta_1, \eta_2) = 1/\eta_2$ are not suitable here.)

Given (9.2) and $\pi(\eta_1, \eta_2)$ as above,

$$\begin{aligned}\pi(\boldsymbol{\eta}|\boldsymbol{X}) &\propto \left\{2\pi(\eta_2 + \frac{\sigma^2}{n})\right\}^{-p/2} \exp\left\{-\frac{1}{2(\eta_2 + \frac{\sigma^2}{n})}\sum_{j=1}^{p}(\bar{X}_j - \eta_1)^2\right\}\pi(\boldsymbol{\eta}) \\ &\propto \left\{2\pi(\eta_2 + \frac{\sigma^2}{n})\right\}^{-1/2} \exp\left\{-\frac{p}{2(\eta_2 + \frac{\sigma^2}{n})}(\eta_1 - \bar{X})^2\right\} \\ &\quad\times \left(\eta_2 + \frac{\sigma^2}{n}\right)^{-(p-1)/2} \exp\left\{-\frac{1}{2(\eta_2 + \frac{\sigma^2}{n})}S\right\}, \end{aligned} \tag{9.3}$$

where $\bar{X} = \frac{1}{p}\sum_{j=1}^{p}\bar{X}_j$ and $S = \sum_{j=1}^{p}(\bar{X}_j - \bar{X})^2$.

In a similar way,

$$\pi(\boldsymbol{\mu}|\boldsymbol{X}) = \int \prod_{j=1}^{p} \pi(\mu_j|\bar{X}_j, \boldsymbol{\eta})\pi(\boldsymbol{\eta}|\bar{\boldsymbol{X}})\, d\boldsymbol{\eta}, \tag{9.4}$$

where $(\mu_j|\bar{X}_j, \boldsymbol{\eta})$ are independent normal with

$$\text{mean as in (9.1) and variance } \frac{\eta_2\sigma^2/n}{\eta_2 + \sigma^2/n} \tag{9.5}$$

and

$$\pi(\boldsymbol{\eta}|\bar{\boldsymbol{X}}) = \pi(\eta_1|\bar{\boldsymbol{X}}, \eta_2)\pi(\eta_2|\bar{\boldsymbol{X}}) \tag{9.6}$$

is the product of a normal and inverse-Gamma (as given in (9.3)).

Construction of a credible interval for μ_j is, in principle, simple. Consider $\pi(\mu_j|\bar{\boldsymbol{X}})$ and fix $0 < \alpha < 1$. Calculate the posterior quantiles $\underline{\mu}_j(\bar{\boldsymbol{X}}), \bar{\mu}_j(\bar{\boldsymbol{X}})$ of orders $100\alpha/2$ and $100(1-\alpha/2)$ for given data. Then

$$P\{\underline{\mu}_j(\bar{\boldsymbol{X}}) \le \mu_j \le \bar{\mu}_j(\bar{\boldsymbol{X}})|\bar{\boldsymbol{X}}\} = 1 - \alpha.$$

In general, to calculate $\underline{\mu}_j$ and $\bar{\mu}_j$ one would have to resort to MCMC which is explained in Subsection 9.1.1.

For large p, good approximations are available. To do this we anticipate to some extent Section 9.2.

Because p is large, we can invoke the theorem on posterior normality (Chapter 4). Thus the posterior for $\boldsymbol{\eta}$ is nearly normal with mean $\hat{\boldsymbol{\eta}}$ and variances of order $O(1/p)$, $\hat{\boldsymbol{\eta}}$ being the MLE of $\boldsymbol{\eta}$ based on the "likelihood"

$$\prod_{j=1}^{p} f(\bar{x}_j|\boldsymbol{\eta}).$$

Hence, $\pi(\boldsymbol{\eta}|\bar{\boldsymbol{x}})$ is approximately (in the sense of convergence in distribution) degenerate at $\hat{\boldsymbol{\eta}}$. This implies

$$\begin{aligned}\pi(\mu_j|\bar{\boldsymbol{X}}) &= \int \pi(\mu_j|\bar{X}_j, \boldsymbol{\eta})\pi(\boldsymbol{\eta}|\bar{\boldsymbol{X}})\, d\boldsymbol{\eta} \\ &= \pi(\mu_j|\bar{X}_j, \hat{\boldsymbol{\eta}}) \text{ (approximately) }. \end{aligned} \tag{9.7}$$

This in turn implies the Bayes estimate of μ_j is

$$E(\mu_j|\bar{\boldsymbol{X}}) = E(\mu_j|\bar{X}_j, \hat{\boldsymbol{\eta}}) \text{ (approximately)} \tag{9.8}$$

which, by (9.1), is a shrinkage estimate that shrinks $\bar{X}_j$ towards $\hat{\eta}_1$, and which depends on all the sample means.

The approximation (9.8) is correct up to $O(1/p)$. A similar argument provides an approximation to the posterior s.d. but the accuracy is only $O(1/\sqrt{p})$.

Simulations indicate the approximation for the Bayes estimate is quite good but that for the posterior s.d. is much less accurate. It is known, vide Morris (1983), that the approximation is also inadequate for credible intervals.

As a final application of this approximation we indicate it is possible to check whether the prior $\pi(\mu_j|\boldsymbol{\eta})$ is consistent with data. More precisely, we check the consistency of $f(\bar{x}_j|\boldsymbol{\eta})$ with data, but a check for $f(\bar{x}_j|\boldsymbol{\eta})$ is indirectly a check for $\pi(\mu_j|\boldsymbol{\eta})$.

In the light of the data, $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}$ is the most likely value of the hyperparameter $\boldsymbol{\eta}$. Under $\hat{\boldsymbol{\eta}}$, $\bar{X}_j$'s are i.i.d normal with mean and variance given by (9.2) with $\boldsymbol{\eta}$ replaced by $\hat{\boldsymbol{\eta}}$. We can now check the fit of this model to the empirical distribution via the Q-Q plot. For each $0 < p < 1$, we plot the 100pth quantiles for the theoretical and empirical distributions as $(x(p), y(p))$. If the fit is good, the resulting curve $\{(x(p), y(p)), 0 < p < 1\}$ should scatter around an equiangular line passing through the origin.

9.1.1 MCMC and E-M Algorithm

We begin with p exponential densities of the same form, namely,

$$\exp\left\{nc(\boldsymbol{\theta}_j) + \sum_{i=1}^{k} t_{ji}(\boldsymbol{x}_j)\theta_{ji}\right\} h(\boldsymbol{x}_j), \quad j = 1, \ldots, p. \tag{9.9}$$

The conjugate prior density for the jth model is proportional to

$$\exp\{\eta_0 c(\boldsymbol{\theta}_j) + \sum_{i=1}^{k} \eta_i \theta_{ji}\}, \quad j = 1, \ldots, p. \tag{9.10}$$

Note that the hyperparameters $(\eta_0, \eta_1, \ldots, \eta_k)$ are the same for all j. Finally, consider a prior $\pi(\boldsymbol{\eta})$ for the hyperparameters.

Let $\boldsymbol{X} = (\boldsymbol{X_1}, \ldots, \boldsymbol{X}_p)$ and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_p)$. The conditional density of $\boldsymbol{\theta}$ given $\boldsymbol{X}, \boldsymbol{\eta}$ is

$$\pi(\boldsymbol{\theta}|\boldsymbol{X}, \boldsymbol{\eta}) \propto \prod_{j=1}^{p} \exp\{(\eta_0 + n)c(\boldsymbol{\theta}_j) + \sum_{i=1}^{k}(t_{ji}(\boldsymbol{x}_j) + \eta_i)\theta_{ji}\} \tag{9.11}$$

which shows conditionally $\boldsymbol{\theta}_j$'s remain independent. Also

$$\pi(\boldsymbol{\eta}|\boldsymbol{X},\boldsymbol{\theta}) \propto \exp\{pd(\boldsymbol{\eta}) + (\eta_0 + n)\sum_{j=1}^{p} c(\boldsymbol{\theta}_j) + \sum_{j=1}^{p}\sum_{i=1}^{k}(\eta_i + t_{ji}(\boldsymbol{x}_j))\theta_{ji}\}\pi(\boldsymbol{\eta}) \tag{9.12}$$

where $\exp(d(\boldsymbol{\eta}))$ is the normalizing constant of the expression in (9.10).

By (9.12), the Bayes formula and cancellation of some common terms in the numerator and denominator of the Bayes formula,

$$\pi(\boldsymbol{\eta}|\boldsymbol{X},\boldsymbol{\theta}) \propto \exp\{pd(\boldsymbol{\eta}) + \eta_0\sum_{j=1}^{p} c(\boldsymbol{\theta}_j) + \sum_{j=1}^{p}\sum_{i=1}^{k}\eta_i\theta_{ji}\}\pi(\boldsymbol{\eta}).$$

If $d(\boldsymbol{\eta})$ has an explicit form, as is often the case, one can apply Gibbs sampling to draw samples from the joint posterior of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ using the conditionals $\pi(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{\eta})$ and $\pi(\boldsymbol{\eta}|\boldsymbol{X},\boldsymbol{\theta})$. Otherwise one can apply Metropolis-Hastings.

In general, the approximations based on $\hat{\boldsymbol{\eta}}$ are still valid but computation of $\hat{\boldsymbol{\eta}}$ is non-trivial. It turns out that the E-M algorithm can be applied, vide Gelman et al. (1995, Chapter 9).

We illustrate the algorithms for MCMC and E-M in the case of $N(\mu_j,\sigma^2)$, $j=1,\ldots,p$, with $(\mu_1,\ldots,\mu_p)$ and σ^2 unknown. MCMC is quite straightforward here. Recall Example 7.13 from Chapter 7. The hierarchical Bayesian analysis of the usual one-way ANOVA was discussed there. With minimal modification, the same algorithm fits here to allow Gibbs sampling. Application of the E-M algorithm is also easy, as discussed in Gelman et al. (1995). We assume as before that μ_j are i.i.d. $N(\eta_1,\eta_2)$, and further take $\pi(\eta_1,\sigma^2,\eta_2) = 1/\sigma^2$. Then, recall from Section 7.2 that we need to apply the E and M steps to

$$\begin{aligned}\log\pi(\boldsymbol{\mu},\eta_1,\sigma^2,\eta_2|\boldsymbol{X}) = &-(\frac{n}{2}+1)\log\sigma^2 - \frac{p}{2}\log\eta_2 - \frac{1}{2\eta_2}\sum_{j=1}^{p}(\mu_j-\eta_1)^2\\ &-\frac{1}{2\sigma^2}\sum_{j=1}^{p}\sum_{i=1}^{n}(X_{ij}-\mu_j)^2 + \text{ constants }.\end{aligned} \tag{9.13}$$

The E-step requires the conditional distribution of $(\boldsymbol{\mu},\sigma^2)$ given $\boldsymbol{X}$ and the current value $(\eta_1^{(c)},\eta_2^{(c)})$ of (η_1,η_2). This is just the normal, inverse Gamma model. In the M-step we need to maximize $E^{(c)}(\log\pi(\boldsymbol{\mu},\eta_1,\sigma^2,\eta_2|\boldsymbol{X}))$ as a function of (η_1,η_2) which is straightforward.

9.2 Parametric Empirical Bayes

To explain the basic ideas, we consider once more the special case of $N(\mu_j,\sigma^2)$, σ^2 known. Explicit formulas are available in this special case for comparison with the estimates of Stein. Another interesting special case is discussed in Carlin and Louis (1996, Chapter 3).

The PEB approach was introduced by Efron and Morris in a series of papers including Efron and Morris (1971, 1972, 1973a, 1973b, 1975, 1976). In this section we generally follow Morris (1983).

Efron and Morris tried to take an intermediate position between a fully Bayes and a fully frequentist approach by treating the likelihood as given by $f(\bar{x}_j|\boldsymbol{\eta})$ obtained by integrating out the μ_j's as in (9.2). The $\boldsymbol{\eta}$'s are treated as unknown parameters as in frequentist analysis whereas the μ_j's are treated as random variables as in Bayesian analysis. This leads to a reduction of a high-dimensional frequentist problem about μ_j's to a low-dimensional semi-frequentist problem about $\boldsymbol{\eta}$, about which there is a lot of information in the data. The fully Bayesian and the PEB approach differ in that no prior is assigned to $\boldsymbol{\eta}$, and $\boldsymbol{\eta}$ is estimated by MLE or by a suitable unbiased estimate. So one may, if one wishes, think of the PEB approach as an approximation to the hierarchical Bayes approach of Section 9.1. A disadvantage of PEB is that accounting for the uncertainty about $\boldsymbol{\eta}$ is more difficult than in hierarchical Bayes – a point that would be discussed again in subsection 9.2.1. An advantage is that one gets an explicit estimate for μ_j, namely, (9.1) with $\boldsymbol{\eta}$ replaced by an estimate of $\boldsymbol{\eta}$.

Note that under the likelihood (9.2), the complete sufficient statistic is the pair

$$(\bar{X} = \frac{1}{p}\sum_{j=1}^{p} \bar{X}_j, \quad S = \sum_{j=1}^{p}(\bar{X}_j - \bar{X})^2). \tag{9.14}$$

Also, $\bar{X}$ and

$$\hat{B} = (p-3)\frac{\sigma^2/n}{S} \tag{9.15}$$

are unbiased estimates of η_1 and

$$B = \frac{\sigma^2/n}{\sigma^2/n + \eta_2}. \tag{9.16}$$

Then the best unbiased predictor of the RHS of (9.1) is

$$\hat{\mu}_j = (1-\hat{B})\bar{X}_j + \hat{B}\bar{X} \tag{9.17}$$

which is the famous James-Stein-Lindley estimate of μ_j. It shrinks $\bar{X}_j$ towards the overall mean $\bar{X}$.

The amount of shrinkage is determined by $\hat{B}$, which is close to 1 if $S/(p-3)$ is close to σ^2/n and close to zero if $S/(p-3)$ is much larger than σ^2/n. Note that if $S/(p-3)$ is small, as in the first case, then the $\bar{X}_j$'s are close to $\bar{X}$ indicating μ_j's are close to each other. This justifies a fairly strong shrinkage towards the grand mean. On the other hand, a large $S/(p-3)$ indicates heterogeneity among the μ_j's, suggesting relatively large weight for $\bar{X}_j$.

Because

$$E(S/(p-1)) = \frac{\sigma^2}{n} + \eta_2, \tag{9.18}$$

an unbiased estimate of η_2 is $\hat{\eta_2} = S/(p-1) - \sigma^2/n$. Because $\eta_2 \geq 0$, a more plausible estimate is $\hat{\eta}_2^+ = \max(0, \hat{\eta}_2)$, the positive part of the unbiased estimate. This suggests changing the estimate of B to

$$\widetilde{B} = \frac{(p-3)}{(p-1)} \frac{\sigma^2}{n} \Big/ (\frac{\sigma^2}{n} + \hat{\eta}_2^+), \tag{9.19}$$

which is the James-Stein-Lindley positive part estimate.

If we take $\eta_1 = 0$, i.e., μ_j's are i.i.d $N(0, \eta_2)$, then the two estimates are of the form

$$\hat{\mu}_j = (1 - \hat{B})\bar{X}_j \text{ and } \widetilde{\mu}_j = (1 - \widetilde{B})\bar{X}_j. \tag{9.20}$$

These are the James-Stein and James-Stein positive part estimates. They shrink the estimate towards an arbitrary point zero and so do not seem attractive in the exchangeable case. But they have turned out to be quite useful in estimating coefficients in an orthogonal expansion of an unknown function with white noise as error, vide Cai et al. (2000). We study frequentist properties of these two estimates in Section 9.4.

9.2.1 PEB and HB Interval Estimates

Morris defines a random confidence interval $(\underline{\mu}_j(\bar{\boldsymbol{X}}), \bar{\mu}_j(\bar{\boldsymbol{X}}))$ for μ_j to have PEB confidence coefficient $(1-\alpha)$ if

$$P\boldsymbol{\eta}\{\underline{\mu}_j \leq \mu_j \leq \bar{\mu}_j\} \geq 1 - \alpha. \tag{9.21}$$

Let $S_j^2 = [1 - ((p-1)/p)\hat{B}]\sigma^2/n + (2/(p-3))\hat{B}^2(\bar{X}_j - \bar{X})^2$. Morris has conjectured

$$\bar{X}_j \pm z_{\alpha/2} S_j \tag{9.22}$$

is a PEB confidence interval with confidence coefficient $1 - \alpha$.

It is shown in Basu et al. (2003) that the conjecture is not true but the violations are so rare and so small in magnitude that it hardly matters. Basu et al. (2003) suggest an adjustment that would make (9.22) true up to $O(p^{-2})$. It is also shown there that asymptotically the adjusted interval is equivalent to a PEB interval proposed by Carlin and Louis (1996, Chapter 3).

A trouble with Morris's interval is that it is somewhat ad hoc. We are not told how exactly it is derived. It seems he puts a noninformative prior on η_1, η_2 and adjusts somewhat the HB credible interval to get a conservative frequentist coverage probability.

There is a natural alternative that does not require additional adjustment and ensures (9.21) with the inequality replaced by an equality up to $O(p^{-2})$. To do this, one has to choose a prior for $\boldsymbol{\eta}$ that is probability matching in the sense of

$$P_{\boldsymbol{\eta}}\{\underline{\mu}_j \le \mu_j \le \bar{\mu}_j\} = 1 - \alpha + O(p^{-2}), \tag{9.23}$$

where

$$\begin{aligned} P\{\mu_j > \bar{\mu}_j | \bar{\boldsymbol{X}}\} &= \alpha/2, \\ P\{\mu_j < \underline{\mu}_j | \bar{\boldsymbol{X}}\} &= \alpha/2, \end{aligned} \tag{9.24}$$

and the probabilities in (9.24) are the posterior probabilities under the prior for $\boldsymbol{\eta}$. This leads to probability matching differential equations. A solution is

$$\pi(\boldsymbol{\eta}) = \frac{\sigma^2/n}{\eta_2 + \sigma^2/n}, \tag{9.25}$$

vide Datta, Ghosh, and Mukerjee (2000).

9.3 Linear Models for High-dimensional Parameters

We can extend the HB and PEB approach to a more general setup using covariates and linear models. The parameters are no longer exchangeable but are affected by a common set of low-dimensional hyperparameters assuming the role of $\boldsymbol{\eta}$. The model in Sections 9.1 and 9.2 is a special case of the linear model below.

Following Morris (1983), we change our notations slightly

$$Y_j|\theta_j \sim N(\theta_j, V), \quad j = 1, \ldots, p \quad \text{independent}, \tag{9.26}$$

and given $\boldsymbol{\beta}$, A,

$$\boldsymbol{\theta}_{p\times 1} = Z_{p\times r}\boldsymbol{\beta}_{r\times 1} + \boldsymbol{\epsilon}_{p\times 1}, \tag{9.27}$$

ϵ_j's are i.i.d $N(0, A)$. Here p is at least moderately large, r is relatively small. Morris allows the variance of ϵ_j to depend on j, which is often a more realistic assumption. Keeping the same variance A for all j simplifies the algebra considerably.

In the HB analysis we need to put a further prior on $\boldsymbol{\beta}$. A conjugate prior for $\boldsymbol{\beta}$ given A is

$$\boldsymbol{\beta} \sim N(\gamma_1, \gamma_2(Z'Z)^{-1}). \tag{9.28}$$

Finally, A is given an inverse Gamma or a uniform or the standard prior $1/A$ for scale parameters. Assuming V is known, our specification of priors is complete.

To do MCMC we partition the parameters and (random) hyperparameters into three sets $(\boldsymbol{\theta}, \boldsymbol{\beta}, A)$. The conditionals are as follows.
(1) Given $\boldsymbol{\beta}, A$ (and $\boldsymbol{Y}$), $\boldsymbol{\theta}$ is multivariate normal.
(2) Given $\boldsymbol{\theta}, A$ (and $\boldsymbol{Y}$), $\boldsymbol{\beta}$ is multivariate normal.
(3) Given $\boldsymbol{\theta}, \boldsymbol{\beta}$ (and $\boldsymbol{Y}$), A has an inverse Gamma distribution. You are asked to find the parameters of these conditionals in Problem 6.

In the PEB approach, one first notes

$$\theta_i|Y_i, \boldsymbol{\beta}, A \sim N(\theta_i^*, V(1-B)), \tag{9.29}$$

where

$$\theta_i^* = (1-B)Y_i + BZ_i'\boldsymbol{\beta} \tag{9.30}$$

with $B = V/(V+A)$. Here Z_i is the ith column vector of Z. This is the shrinkage estimate corresponding with (9.1) of Section 9.1.

In the PEB approach one has to estimate $\boldsymbol{\beta}$ and B either by maximizing the likelihood of the independent Y_i's with

$$Y_i|\boldsymbol{\beta}, A \sim N(Z_i'\beta, V+A) \tag{9.31}$$

or by finding a suitable unbiased estimate as in (9.18). Let

$$\hat{\boldsymbol{\beta}} = (Z'Z)^{-1}(Z'\mathbf{Y}).$$

The statistic $\hat{\boldsymbol{\beta}}$ and

$$S = (\boldsymbol{Y} - Z\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - Z\hat{\boldsymbol{\beta}})$$

are independent, complete sufficient statistics for $(\boldsymbol{\beta}, A)$. Hence the best unbiased estimates for $\boldsymbol{\beta}$ and B are $\hat{\boldsymbol{\beta}}$ and

$$\hat{B} = (p-r-2)V/S$$

(vide Problem 10). Substituting in the shrinkage estimate θ_i^* for θ_i, one gets

$$\hat{\theta}_i = (1-\hat{B})Y_i + \hat{B}Z'\hat{\boldsymbol{\beta}}.$$

This is the analogue of James-Stein-Lindley estimate under the regression model.

In Problem 8, you are asked to show that the PEB risk of $\hat{\theta}_i$, namely $E_{\boldsymbol{\beta},A}(\hat{\theta}_i - \theta_i)^2$ is smaller than the PEB risk of Y_i, namely, $E_{\boldsymbol{\beta},A}(Y_i - \theta_i)^2$. The relative strength of the PEB estimate comes through the use of $\hat{\boldsymbol{\beta}}$, which is based on the full data set $\boldsymbol{Y}$.

In Section 8.3, linear regression is discussed as a common statistical problem where an objective Bayesian analysis is done. You may wish to explore how similar ideas are used in this section to model a high-dimensional problem.

9.4 Stein's Frequentist Approach to a High-dimensional Problem

Once again we study Example 9.1. Let $\bar{X}_j$'s be independent, $\bar{X}_j \sim N(\mu_j, \sigma^2/n)$. Classical criteria like maximum likelihood, minimaxity or minimizing variance among unbiased estimates, all lead to $(\bar{X}_1, \ldots, \bar{X}_p)$ as estimate of $(\mu_1, \ldots, \mu_p)$. Let $p \geq 3$. Stein, in a series of papers, Stein (1956), James and Stein (1960), Stein (1981), showed that if we define a loss function

$$L(\bar{\boldsymbol{X}}, \boldsymbol{\mu}) = \sum_{j=1}^{p} (\bar{X}_j - \mu_j)^2 \tag{9.32}$$

and generally

$$L(\boldsymbol{T}, \boldsymbol{\mu}) = \sum_{j=1}^{p} (T_j(\bar{\boldsymbol{X}}) - \mu_j)^2 \tag{9.33}$$

for a general estimate $\boldsymbol{T}$, it is possible to choose a $\boldsymbol{T}$ that is better than $\bar{\boldsymbol{X}}$ in the sense

$$E_{\boldsymbol{\mu}}(L(\boldsymbol{T}, \boldsymbol{\mu})) < E_{\boldsymbol{\mu}}(L(\bar{\boldsymbol{X}}, \boldsymbol{\mu})) \text{ for all } \boldsymbol{\mu}. \tag{9.34}$$

Stein (1956) provides a heuristic motivation that suggests $\bar{\boldsymbol{X}}$ is too large in a certain sense explained below. To see this compare the expectation of the squared norm of $\bar{\boldsymbol{X}}$ with the squared norm of $\boldsymbol{\mu}$.

$$E_{\boldsymbol{\mu}}(\|\bar{\boldsymbol{X}}\|^2) = \|\boldsymbol{\mu}\|^2 + p\sigma^2/n > \|\boldsymbol{\mu}\|^2. \tag{9.35}$$

The larger the p the bigger the deviation between the LHS and RHS. So one would expect at least for sufficiently large p, one can get a better estimate by shrinking each coordinate of $\bar{\boldsymbol{X}}$ suitably towards zero. We present below two of the most well-known shrinkage estimates, namely, the James-Stein and the positive part James-Stein estimate. Both have already appeared in Section 9.2 as PEB estimates. It seems to us that the PEB approach provides the most insight about Stein's estimates, even though the PEB interpretation came much later.

Morris points out that there is no exchangeability or prior in Stein's approach but summing the individual losses produces a similar effect. Moreover, pooling the individual losses would be a natural thing to do only when the different μ_j's are related in some way. If they are totally unrelated, Stein's result would be merely a curious fact with no practical significance, not a profound new data analytic tool. It is in the case of exchangeable high-dimensional problems that it provides substantial improvement.

We present two approaches to proving that the Stein-James estimate is superior to the classical estimate. One is based on Stein (1981) with details as in Ibragimov and Has'minskii (1981). The other is an interesting variation on this due to Schervish (1995).

Stein's Identity. *Let $X \sim N(\mu, \sigma^2)$ and $\phi(x)$ be a real valued function differentiable on $\mathcal{R}$ with $\int_a^x \phi'(u)du = \phi(x) - \phi(a)$. Then*

$$\sigma^2 E(\phi'(X)) = E((X - \mu)\phi(X)).$$

Proof. Integrating by parts or changing the order of integration

$$\begin{aligned} E(\phi'(X)) &= \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\phi'(x)\exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}dx \\ &= -\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty}\phi(x)\frac{d}{dx}\exp\left\{\frac{-(x-\mu)^2}{2\sigma^2}\right\}dx \\ &= \sigma^{-2}E(\phi(X)(X-\mu)).\square \end{aligned} \tag{9.36}$$

For more details see the proof in Stein (1981).

As a corollary we have the following result.

Corollary. *Let $(X_1, X_2, \ldots, X_p)$ be a random vector $\sim N_p(\boldsymbol{\mu}, \frac{\sigma^2}{n}I)$. Let $\phi = (\phi_1, \phi_2, \ldots, \phi_p) : \mathcal{R}^p \to \mathcal{R}^p$ be differentiable, $E|\frac{\partial\phi_j}{\partial X_j}| < \infty$, $\phi_j(x_1, \ldots, x_{j-1}, x, x_{j+1}, \ldots, x_p) = \int_a^x \frac{\partial\phi_j}{\partial x_j}dx_j$ and assume that $\phi_j(x_1, \ldots, x_{j-1}, x, x_{j+1}, \ldots)\exp\{\frac{-(x-\mu_j)^2}{2\sigma^2/n}\} \to 0$ as $|x| \to \infty$. Then*

$$E\left\{\sigma^2\frac{\partial\phi_j}{\partial X_j}\right\} = E((X_j - \mu_j)\phi_j). \tag{9.37}$$

We now return to Example 9.1. The classical estimate for $\boldsymbol{\mu}$ is $\bar{\boldsymbol{X}} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_p)$. Consider an alternative estimate of the form

$$\widetilde{\boldsymbol{\mu}} = \bar{\boldsymbol{X}} + n^{-1}\sigma^2 g(\bar{\boldsymbol{X}}), \tag{9.38}$$

where $g(\boldsymbol{x}) = (g_1, g, \ldots, g_p) : \mathcal{R}^p \to \mathcal{R}^p$ with g satisfying the conditions in the corollary.

Then, by the corollary,

$$\begin{aligned} &E_{\boldsymbol{\mu}}\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}\|^2 - E_{\boldsymbol{\mu}}\|\bar{\boldsymbol{X}} + n^{-1}\sigma^2 g(\bar{\boldsymbol{X}}) - \boldsymbol{\mu}\|^2 \\ &\quad = -2n^{-1}\sigma^2 E_{\boldsymbol{\mu}}\{(\bar{\boldsymbol{X}} - \boldsymbol{\mu})' g(\bar{\boldsymbol{X}})\} - n^{-2}\sigma^4 E_{\boldsymbol{\mu}}\|g(\bar{\boldsymbol{X}})\|^2 \\ &\quad = -2n^{-2}\sigma^4 E_{\boldsymbol{\mu}}\{\sum_1^p \frac{\partial g_j}{\partial X_j}\} - n^{-2}\sigma^4 E_{\boldsymbol{\mu}}\|g(\bar{\boldsymbol{X}})\|^2. \end{aligned} \tag{9.39}$$

Now suppose $g(\boldsymbol{x}) = \text{grad}(\log\phi(\boldsymbol{x}))$, where ϕ is a twice continuously differentiable function from $\mathcal{R}^p$ into $\mathcal{R}$. Then

$$\sum_1^p \frac{\partial g_j}{\partial x_j} = -\|g\|^2 + \frac{1}{\phi}\Delta\phi \tag{9.40}$$

where $\Delta = \sum_1^p \frac{\partial^2}{\partial x_j^2}$. Hence

$$E_{\boldsymbol{\mu}}\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}\|^2 - E_{\boldsymbol{\mu}}\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = n^{-2}\sigma^4 E_{\boldsymbol{\mu}}\|g\|^2 - n^{-2}\sigma^4 E_{\boldsymbol{\mu}}\left\{\frac{1}{\phi(\bar{\boldsymbol{X}})}\Delta\phi(\bar{\boldsymbol{X}})\right\} \tag{9.41}$$

which is positive if $\phi(x)$ is a positive non-constant superharmonic function, i.e,

$$\Delta\phi \leq 0. \tag{9.42}$$

Thus $\widetilde{\boldsymbol{\mu}}$ is superior to $\bar{\boldsymbol{X}}$ if (9.42) holds. It is known that such functions exist if and only if $p \geq 3$.

To show the superiority of the James-Stein positive part estimate for $p \geq 3$, take

$$\phi(\boldsymbol{x}) = \begin{cases} \|\boldsymbol{x}\|^{-(p-2)} & \text{if } \|\boldsymbol{x}\| \geq \sqrt{p-2} \\ (p-2)^{-(p-2)/2} \exp\left\{\frac{1}{2}(p-2) - \|\boldsymbol{x}\|^2)\right\} & \text{otherwise.} \end{cases} \tag{9.43}$$

Then grad$(\log\phi)$ is easily verified to be the James-Stein positive part estimate.

To show the superiority of the James-Stein estimate, take

$$\phi(\boldsymbol{x}) = \|\boldsymbol{x}\|^{p-2}. \tag{9.44}$$

We observed earlier that shrinking towards zero is natural if one modeled μ_j's as exchangeable with common mean equal to zero. We expect substantial improvement if $\boldsymbol{\mu} = \mathbf{0}$.

Calculation shows

$$E_{\boldsymbol{\mu}}\|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = \frac{2}{p} E_{\boldsymbol{\mu}}\|\bar{\boldsymbol{X}} - \boldsymbol{\mu}\|^2 = 2 \tag{9.45}$$

if $\boldsymbol{\mu} = \mathbf{0}, \sigma^2 = 1, n = 1$.

It appears that borrowing strength in the frequentist formulation is possible because Stein's loss adds up the losses of the component decision problems. Such addition would make sense only when the different problems are connected in a natural way, in which case the exchangeability assumption and the PEB or hierarchical models are also likely to hold. It is natural to ask how good are the James-Stein estimates in the frequentist sense. They are certainly minimax since they dominate minimax estimates. Are they admissible? Are they Bayes (not just PEB)? For the James-Stein positive part estimate the answer to both questions is no, see Berger (1985a, pp. 542, 543). On the other hand, Strawderman (1971) constructs a proper Bayes minimax estimate for $p \geq 5$. Berger (1985a, pp. 364, 365) discusses the question of which among the various minimax estimates to choose. Note that the PEB approach leads in a natural way to James-Stein positive part estimate, suggesting that it can't be substantially improved even though it is not Bayes. See in this connection Robert (1994, p. 66). There is a huge literature on Stein estimates as well as questions of admissibility in multidimensional problems. Berger (1985a) and Robert (1994) provide excellent reviews of the literature. There are intriguing connections between admissibility and recurrence of suitably constructed Markov processes, see Brown (1971), Srinivasan (1981), and Eaton (1992, 1997, 2004).

When extreme μ's may occur, the Stein estimates do not offer much improvement. Stein (1981) and Berger and Dey (1983) suggest how this problem can be solved by suitably truncating the sample means. For Stein type results for general ridge regression estimates see Strawderman (1978) where several other references are given.

Of course, instead of zero we could shrink towards an arbitrary $\boldsymbol{\mu}_0$. Then a substantial improvement will occur near $\boldsymbol{\mu}_0$. Exactly similar results hold for the James-Stein-Lindley estimate and its positive part estimate if $p \geq 4$.

For the James-Stein estimate, Schervish (1995, pp. 163–165) uses Stein's identity as well as (9.40) but then shows directly (with $\sigma^2 = 1, n = 1$)

$$\|g\|^2 + 2\sum_{j=1}^{p} \frac{\partial}{\partial x_j} g_j = \frac{-(p-2)^2}{\sum_1^p x_j^2} < 0.$$

Clearly for $\widetilde{\boldsymbol{\mu}} =$ James-Stein estimate,

$$E_{\boldsymbol{\mu}} \|\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}\|^2 = p - E_{\boldsymbol{\mu}} \left\{ \frac{(p-2)^2}{\sum \bar{X}_j^2} \right\},$$

which shows how the risk can be evaluated by simulating a noncentral χ^2 -distribution.

9.5 Comparison of High-dimensional and Low-dimensional Problems

In the low-dimensional case, where n is large or moderate and p small, the prior is washed away by the data, the likelihood influences the posterior more than the prior. This is not so when p is much larger than n – the so-called high-dimensional case. The prior is important, so elicitation, if possible, is important. Checking the prior against data is possible and should be explored. We discuss this below.

In the high-dimensional cases examined in Sections 9.2 and 9.3 some aspects of the prior, namely $\pi(\mu_j|\hat{\boldsymbol{\eta}})$, can be checked against the empirical distribution. We have discussed this earlier mathematically, but one can approach this from a more intuitive point of view. Because we have many μ_j's as sample from $\pi(\mu_j|\hat{\boldsymbol{\eta}})$ and $\bar{X}_j$'s provide approximate estimates of μ_j's, the empirical distribution of the $\bar{X}_j$'s should provide a check on the appropriateness of $\pi(\mu_j|\hat{\boldsymbol{\eta}})$.

Thus there is a curious dichotomy. In the low-dimensional case, the data provide a lot of information about the parameters but not much information about their distribution, i.e., the prior. The opposite is true in high-dimensional problems. The data don't tell us much about the parameters but there is information about the prior.

This general fact suggests that the smoothed empirical distribution of estimates could be used to generate a tentative prior if the likelihood is not exponential and so conjugate priors cannot be used. Adding a location-scale hyperparameter η could provide a family of priors as a starting point of objective high-dimensional Bayesian analysis.

Bernardo (1979) has shown that at least for Example 9.1 a sensible Bayesian analysis can be based on a reference prior with a suitable reparameterization. It does seem very likely that this example is not an exception but a general theory of the right reparameterization needs to be developed.

9.6 High-dimensional Multiple Testing (PEB)

Multiple tests have become very popular because of application in many areas including microarrays where one searches for genes that have been expressed. We provide a minimal amount of modeling that covers a variety of such applications arising in bioinformatics, statistical genetics, biology, etc. Microarrays are discussed in Appendix D. Whereas PEB or HB high-dimensional estimation has been around for some time, PEB or HB high-dimensional multiple testing is of fairly recent origin, e.g., Efron et al. (2001a), Newton et al. (2003), etc.

We have p samples, each of size n, from p normal populations. In the simplest case we assume the populations are homoscedastic. Let σ^2 be the common unknown variance, and the means $\mu_1, \ldots, \mu_p$.

For μ_j, consider the hypotheses $H_{0j} : \mu_j = 0, H_{1j} : \mu_j \sim N(\eta_1, \eta_2), j = 1, \ldots, p$. The data are $X_{ij}, i = 1, \ldots n, j = 1, \ldots, p$. In the gene expression problem, $X_{ij}, i = 1, \ldots n$ are n i.i.d. observations on the expression of the jth gene. The value of $|X_{ij}|$ may be taken as a measure of observed intensity of expression. If one accepts H_{0j}, it amounts to saying the jth gene is not expressed in this experiment. On the other hand, accepting H_{1j} is to assert that the jth gene has been expressed. Roughly speaking, a gene is said to be expressed when the gene has some function in the cell or cells being studied, which could be a malignant tumor. For more details, see the appendix. In addition to H_{0j} and H_{1j}, the model involves $\pi_0 =$ probability that H_{0j} is true and $\pi_1 = 1 - \pi_0 =$ probability that H_{1j} is true. If

$$I_j = \begin{cases} 1 & \text{if } H_{1j} \text{ is true;} \\ 0 & \text{if } H_{0j} \text{ is true,} \end{cases}$$

then we assume $I_1, \ldots, I_p$ are i.i.d. $\sim B(1, \pi_1)$.

The interpretation of π_1 has a subjective and a frequentist aspect. It represents our uncertainty about expression of each particular gene as well as approximate proportion of expression among p genes.

If $\sigma^2, \pi_1, \eta_1, \eta_2$ are all known, $\bar{X}_j$ is sufficient for μ_j and a Bayes test is available for each j. Calculate the posterior probability of H_{1j}:

$$\pi_{1j} = \frac{\pi_1 f_1(\bar{X}_j)}{\pi_1 f_1(\bar{X}_j) + \pi_0 f_0(\bar{X}_j)}$$

which is a function of $\bar{X}_j$ only. Here f_0 and f_1 are densities of $\bar{X}_j$ under H_{0j} and H_{1j}.

$$\begin{aligned} &\text{If } \pi_{1j} > \frac{1}{2} \quad \text{accept } H_{1j} \quad \text{and} \\ &\text{if } \pi_{1j} < \frac{1}{2} \quad \text{accept } H_{0j}. \end{aligned}$$

This test is based only on the data for the jth gene.

In practice, we do not know π_1, η_1, η_2. In PEB testing, we have to estimate all three. In HB testing, we have to put a prior on (π_1, η_1, η_2). To us a natural prior would be a uniform for π_1 on some range $(0, \delta), \delta$ being upper bound to π_1, uniform prior for η_1 on $\mathcal{R}$ and uniform or some other objective prior for η_2.

In the PEB approach, we have to estimate π_1, η_1, η_2. If σ^2 is also unknown, we have to put a prior on σ^2 also or estimate it from data. An estimate of σ^2 is $\sum_i \sum_j (X_{ij} - \bar{X}_j)^2 / \{p(n-1)\}$.

For fixed π_1, we can estimate η_1 and η_2 by the method of moments using the equations,

$$\bar{X} \equiv \frac{1}{p} \sum \bar{X}_j = \pi_1 \eta_1, \tag{9.46}$$

$$\frac{1}{p} \sum (\bar{X}_j - \bar{X})^2 = \frac{\sigma^2}{n} + \pi_1 \eta_2 + \pi_1 (1 - \pi_1) \eta_1^2, \tag{9.47}$$

from which it follows that

$$\hat{\eta}_1 = \frac{1}{\pi_1} \bar{X}, \tag{9.48}$$

$$\hat{\eta}_2 = \frac{1}{\pi_1} \left\{ \frac{1}{p} \sum (\bar{X}_j - \bar{X})^2 - \frac{\sigma^2}{n} - \frac{1 - \pi_1}{\pi_1} \left(\bar{X}\right)^2 \right\}^+ . \tag{9.49}$$

Alternatively, if it is felt that $\eta_1 = 0$, then the estimate for η_2 is given by

$$\hat{\eta}_2 = \frac{1}{\pi_1} \left\{ \frac{1}{p} \sum (\bar{X}_j - \bar{X})^2 - \frac{\sigma^2}{n} \right\}^+ . \tag{9.50}$$

Now we may maximize the joint likelihood of $\bar{X}_j$'s with respect to π_1.

Using these estimates, we can carry out the Bayes test for each j, provided we know π_1 or put a prior on π_1. We do not know of good PEB estimates of π_1.

Scott and Berger (2005) provide a very illuminating fully Bayesian analysis for microarrays.

9.6.1 Nonparametric Empirical Bayes Multiple Testing

Nonparametric empirical Bayes (NPEB) solutions were introduced by Robbins (1951, 1955, 1964). It is a Bayes solution based on a nonparametric estimate of the prior. Robbins applied these ideas in an ingenious way in several problems. It was regarded as a breakthrough, but the method never became popular because the nonparametric methods did not perform well even in moderately large samples and were somewhat unstable.

Recently Efron et al. (2001a, b) have made a successful application to a microarray with p equal to several thousands. The data are massive enough for NPEB to be stable and perform well.

After some reductions the testing problem takes the following form.
For $j = 1, 2, \ldots, p$, we have random variables Z_j. $Z_j \sim f_0(z)$ under H_{0j} and $Z_j \sim f_1(z)$ under H_{1j} where f_0 is completely specified but $f_1(z) \neq f_0(z)$ is completely unknown. This is what makes the problem nonparametric. Finally, as in the case of parametric empirical Bayes, the indicator of H_{1j} is $I_j = 1$ with probability π_1 and $= 0$ with probability $\pi_0 = 1 - \pi_1$. If π_1 and f_1 were known we could use the Bayes test of H_{0j} based on the posterior probability of H_{1j}

$$P(H_{1j}|z_j) = \frac{\pi_1 f_1(z_j)}{\pi_1 f_1(z_j) + (1-\pi_1) f_0(z_j)}.$$

Let $f(z) = \pi_1 f_1(z) + (1-\pi_1) f_0(z)$. We know $f_0(z)$. Also we can estimate $f(z)$ using any standard method – kernel, spline, nonparametric Bayes, vide Ghosh and Ramamoorthi (2003) – from the empirical distribution of the z_j's. But since π_1 and f_1 are both unknown, there is an identifiability problem and hence estimation of π_1, f_1 is difficult. The two papers, Efron et al. (2001a, b), provide several methods for bounding π_1.

One bound follows from

$$\pi_0 \leq \min_z [f(z)/f_0(z)],$$

$$\pi_1 \geq 1 - \min_z [f(z)/f_0(z)].$$

So the posterior probability of H_{1j} is

$$P\{H_{1j}|z_j\} = 1 - \frac{\pi_0 f_0(z_j)}{f(z_j)} \geq 1 - \left\{ \min_z \frac{f(z)}{f_0(z)} \right\} \frac{f_0(z_j)}{f(z_j)}$$

which is estimated by $1 - \left\{ \min_z \frac{\hat{f}(z)}{f_0(z)} \right\} \frac{f_0(z_j)}{\hat{f}(z_j)}$, where $\hat{f}$ is an estimate of f as mentioned above. The minimization will usually be made over observed values of z.

Another bound is given by

$$\pi_0 \leq \frac{\int_A f(z) dz}{\int_A f_0(z) dz}.$$

Now minimize the RHS over different choices of A. Intuition suggests a good choice would be an interval centered at the mode of $f_0(z)$, which will usually be at zero. A fully Bayesian nonparametric approach is yet to be worked out. Other related papers are Efron (2003, 2004). For an interesting discussion of microarrays and the application of nonparametric empirical Bayes methodology, see Young and Smith (2005).

9.6.2 False Discovery Rate (FDR)

The false discovery rate (FDR) was introduced by Benjamini and Hochberg (1995). Controlling it has become an important frequentist concept and method in multiple testing, specially in high-dimensional problems. We provide a brief review, because it has interesting similarities with NPEB, as noted, e.g., in Efron et al. (2001a, b). We consider the multiple testing scenario introduced earlier in this section. Consider a fixed test. The (random) FDR for the test is defined as $\frac{U(\boldsymbol{z})}{V(\boldsymbol{z})} I_{\{V(\boldsymbol{z})>0\}}$, where $U =$ total number of false discoveries, i.e., number of true H_{0j}'s that are rejected by the test for a $\boldsymbol{z}$, and $V =$ total number of discoveries, i.e., number of H_{0j}'s that are rejected by a test. The (expected) FDR is

$$FDR = E_{\boldsymbol{\mu}} \left(\frac{U}{V} I_{\{V>0\}} \right).$$

To fix ideas suppose all H_{0j}'s are true, i.e., all μ_j's are zero, then $U = V$ and so

$$\frac{U}{V} I_{\{V>0\}} = I_{\{V>0\}}$$

and

$$\begin{aligned} FDR &= P_{\boldsymbol{\mu}=0}(\text{ at least one } H_{0j} \text{ is rejected }) \\ &= \text{ Type 1 error probability under the full null.} \end{aligned}$$

This is usually called family wise error rate (FWER). The Benjamini-Hochberg (BH) algorithm (see Benjamini and Hochberg (1995)) for controlling FDR is to define

$$j_0 = \max\{j : P_{(j)} \le \frac{j}{p} \alpha\}$$

where $P_j =$ the P-value corresponding with the test for jth null and $P_{(j)} =$ jth order statistic among the P-values with $P_{(1)} =$ the smallest, etc.

The algorithm requires rejecting all H_{0j} for which $P_j \le P_{(j_0)}$. Benjamini and Hochberg (1995) showed this ensures

$$E_{\boldsymbol{\mu}} \left(\frac{U}{V} I_{\{V>0\}} \right) \le \frac{p_0}{p+1} \alpha \le \alpha \ \forall \boldsymbol{\mu}$$

where p_0 is the number of true H_{0j}'s. It is a remarkable result because it is valid for all $\boldsymbol{\mu}$. This exact result has been generalized by Sarkar (2003).

Benjamini and Liu (1999) have provided another algorithm. See also Benjamini and Yekutieli (2001). Genovese and Wasserman (2001) provide a test based on an asymptotic evaluation of j_0 and a less conservative rejection rule. An asymptotic evaluation is also available in Genovese and Wasserman (2002). See also Storey (2002) and Donoho and Jin (2004). Scott and Berger (2005) discuss FDR from a Bayesian point of view.

Controlling FDR leads to better performance under alternatives than controlling FWER. Many successful practical applications of FDR control are known. On the other hand, from a decision theoretic point of view it seems more reasonable to control the sum of false discoveries and false negatives rather than FDR and proportion of false negatives.

9.7 Testing of a High-dimensional Null as a Model Selection Problem[1]

Selection from among nested models is one way of handling testing problems as we have seen in Chapter 6. Parsimony is taken care of to some extent by the prior on the additional parameters of the more complex model. As in estimation or multiple testing, consider samples of size r from p normal populations $N(\mu_i, \sigma^2)$. For simplicity σ^2 is assumed known. Usually σ^2 will be unknown. Because $S^2 = \sum_i \sum_j (X_{ij} - \bar{X}_i)^2/p(r-1)$ is an unbiased estimate of σ^2 with lots of degrees of freedom, it does not matter much whether we put one of the usual objective priors for σ^2 or pretend that σ^2 is known to be S^2.

We wish to test $H_0 : \mu_i = 0 \ \forall i$ versus H_1: at least one $\mu \neq 0$. This is sometimes called Stone's problem, Berger et al. (2003), Stone (1979). We may treat this as a model selection problem with $M_0 \equiv H_0 : \mu_i = 0 \ \forall i$ and $M_1 = H_0 \cup H_1$, i.e., $M_1 : \boldsymbol{\mu} \in \mathcal{R}^p$. In this formulation, $M_0 \subset M_1$ whereas H_0 and H_1 are disjoint. On grounds of parsimony, H_0 is favored if both M_0 and M_1 are equally plausible.

To test a null or select a model, we have to define a prior $\pi(\boldsymbol{\mu})$ under M_1 and calculate the Bayes factor

$$B_{01} = \frac{\prod_{i=1}^p f_0(\boldsymbol{X}_i)}{\int_{\mathcal{R}^p} \prod_{i=1}^p f_1(\boldsymbol{X}_i|\mu_i)\pi(\boldsymbol{\mu})d\boldsymbol{\mu}}.$$

There is no well developed theory of objective priors, specially for testing problems. However as in estimation it appears natural to treat μ_j's as exchangeable rather than independent. A popular prior in this context is the Zellner and Siow (1980) multivariate Cauchy prior

$$\pi(\boldsymbol{\mu}) = \frac{\Gamma(\frac{(p+1)}{2})}{\pi^{\frac{p+1}{2}}\sigma^p}(1 + \frac{\boldsymbol{\mu}'\boldsymbol{\mu}}{\sigma^2})^{-\frac{(p+1)}{2}}$$

[1] Section 9.7 may be omitted at first reading.

$$= \int_0^\infty \frac{t^{\frac{p}{2}}}{(2\pi)^{\frac{p}{2}}\sigma^p} e^{-\frac{t}{2\sigma^2}\boldsymbol{\mu}'\boldsymbol{\mu}} \frac{1}{\sqrt{2\pi}} e^{-\frac{t}{2}} t^{-\frac{1}{2}} dt. \tag{9.51}$$

Another plausible prior is the smooth Cauchy prior given by

$$\pi_{sc}(\boldsymbol{\mu}) = \frac{\Gamma(\frac{p+1}{2})}{\Gamma(\frac{p+2}{2})\Gamma(\frac{1}{2})(2\pi\sigma^2)^{\frac{p}{2}}} e^{-\frac{\boldsymbol{\mu}'\boldsymbol{\mu}}{2\sigma^2}} M(\frac{1}{2}, \frac{p+2}{2}, \frac{\boldsymbol{\mu}'\boldsymbol{\mu}}{2\sigma^2})$$
$$= \int_0^1 \frac{t^{\frac{p}{2}}}{(2\pi)^{\frac{p}{2}}\sigma^p} e^{-\frac{t}{2\sigma^2}\boldsymbol{\mu}'\boldsymbol{\mu}} \frac{dt}{\pi\sqrt{t(1-t)}},$$

where $M(\frac{1}{2}, \frac{p+2}{2}, \frac{\boldsymbol{\mu}'\boldsymbol{\mu}}{2\sigma^2})$ is the hypergeometric ${}_1F_1$ function of Abramowitz and Stegun (1970).

It is tempting to use the difference (between the two models) of BIC as an approximation to the logarithm of Bayes factor (BF) even though it was developed by Schwarz for low-dimensional problems. Stone was the first to point out that the use of BIC is problematic in high-dimensional problems. Berger et al. (2003) have developed a generalization of BIC called GBIC, which provides a good approximation to the integrated likelihood for priors like the above Cauchy priors which are obtained by integrating the scale parameter for $N(\mu_i, \sigma^2)$. In Stone's problem one has the normal linear model setup

$$X_{ij} = \mu_i + \epsilon_{ij};\ i = 1, \ldots, p;\ j = 1, \ldots, r;\ n = pr. \tag{9.52}$$

It is assumed that as $n \to \infty$, $p \to \infty$ and r is fixed. Under these assumptions, Berger et al. (2003) provide a Laplace approximation and a GBIC. The GBIC also approximates the BIC for low-dimensional problems. The formula for ΔGBIC (the difference of GBIC for the comparison of M_1 and M_0) is given by

$$\Delta\text{GBIC} = (\frac{r}{2}\bar{\mathbf{X}}'\bar{\mathbf{X}} - \frac{p}{2}\log(rc_p) - \frac{p}{2})^+ - \frac{\log p}{2}, \tag{9.53}$$

where $c_p = \frac{1}{p}\sum_{i=1}^p \bar{X}_i{}^2$. Table 9.1, taken from Berger et al. (2003) provides some idea of the accuracy of BIC, GBIC and Laplace approximation. One has $p = 50$ and $r = 2$ for these calculations and the multivariate Cauchy prior was used.

Substantial new results appear in Liang et al. (2005). They propose a mixture of Zellner's (Zellner (1986)) popular g-prior. In Zellner's form, the prior looks like $\boldsymbol{\mu}|M_1 \sim N(\mathbf{0}, \frac{g}{\sigma^2}(\mathbf{Z}'\mathbf{Z})^{-1})$ where $\mathbf{Z}$ is the design matrix (in our problem only composed of 0's and 1's). This g is usually elicited through an empirical Bayes method. The above authors consider a family of mixtures of g-priors (under which the Zellner-Siow Cauchy prior is a special case) and use those for model selection. They propose Laplace approximations to the

Table 9.1. Comparison of the Performance of GBIC and Laplace Approximation with BIC

c_p	True Log Bayes Factor	ΔBIC	$\Delta GBIC$	ΔLaplace Approx
0.1	-8.5348	-110.129	-1.956	-8.5776
0.5	-3.8251	-90.129	-1.956	-3.9083
1.0	6.0388	-65.129	5.715	5.9236
1.5	20.8203	-40.129	20.579	20.7564
2.0	38.4814	-15.129	38.387	38.4408
10.0	397.369	384.871	398.151	397.369

marginal likelihood under these general priors and show that the models thus selected are generally correct asymptotically if the complex model is true. Under the null model, this type of consistency still holds under the Zellner-Siow prior.

Further generalizations to non-normal problems appear in Berger (2005) and Chakrabarti and Ghosh (2005a). Both papers provide generalizations of BIC when the observations come from an exponential family of distributions in high-dimensional problems. In Table 9.2, using simulation results reported in Chakrabarti and Ghosh (2005a), the performance of GBIC and the Laplace approximation ($\log \hat{m}_2$) with BIC are compared in approximating the integrated likelihood under the more complex model (denoted by m_2) when the more complex model is actually true and observations come from Bernoulli, exponential, and Poisson distributions. In this case one has p groups of observations, each group having a (potentially) different parameter value and each group has r observations. Under the simpler model, these different groups are assumed to have the same (specified) parameter value, while for the more complex model the parameter vector is assumed to belong to $\mathcal{R}^p$. See the paper for details on the priors used.

In principle, the same methods apply to any two nested models

$$M_0 : \mu_i = 0, 1 \leq i \leq p_1, p_1 < p \text{ versus } M_1 : \boldsymbol{\mu} \in \mathcal{R}^p.$$

Table 9.2. Approximation to Integrated Likelihood in the Exponential Family

Distribution	p	r	$\log m_2$	$\log \hat{m}_2$	BIC	$GBIC$
Bernoulli	50	10	-327.45	-327.684	-349.577	-327.863
Bernoulli	50	200	-4018.026	-4018.072	-4052.757	-4018.587
Exponential	50	10	-662.526	-661.979	-640.320	-660.384
Exponential	50	200	-22186.199	-22186.100	-22178.759	-22186.117
Poisson	50	10	-671.504	-670.775	-683.383	-671.374
Poisson	50	200	-15704.585	-15704.618	-15713.139	-15705.010

9.8 High-dimensional Estimation and Prediction Based on Model Selection or Model Averaging[2]

Given a set of data from an experiment or observational study done on a given population, a statistician is asked the following three questions quite frequently. First, which among a given set of possible statistical models seems to be the correct model describing the underlying mechanism producing the data? Second, what will be the predicted value of a future observation, if the experimental conditions are kept at predetermined levels? Third, what is the estimate of a single parameter or a vector (may be infinite dimensional) of parameters? We will focus in this section on some Bayesian approaches to answer the last two types of questions. But before going into the details, we will explain briefly in the next paragraph how one would pose the above three questions from a decision theoretic point of view and what is the basic difference in the Bayesian approaches in tackling such questions.

Bayesian approaches to such questions are basically dictated by the goal of obtaining decision theoretic optimality, and hence the solutions are also heavily dependent upon the type of loss functions being used. The loss function, on the other hand, is mostly determined by the goal of the statistician or practitioner. The goal of the statistician in the first problem above is to select the correct model (which is assumed to be one in the list of models considered). The loss function often used in this problem is the 0-1 loss function. In the Bayesian approach to model selection, the statistician would put prior probabilities on the set of candidate models and a simple argument shows that for this loss, the optimum Bayesian model would be the posterior mode, i.e., the model that has the maximum posterior probability. As explained in the earlier section, BIC and GBIC can be used to select a model using the Bayesian paradigm with 0-1 loss if the sample size is large, in appropriate situations, as they approximate the integrated likelihood and hence can be used to find the model with highest posterior probability. On the other hand, if one is interested in answering the second or third question above (i.e., if one is interested in prediction or estimation of a parameter), the problem can be approached in two different ways. First, one might be interested in finding a particular model that does the best job of prediction (in some appropriate sense). Secondly, one might only want a predicted value, not a particular model for repeated future use in prediction. In either case, the most popular loss function is the squared prediction error loss, i.e., the square of the difference between the predicted/estimated value and the value being predicted/estimated. The best predictor/estimator turns out to be the Bayesian model averaging estimate (to be explained later) and the best predictive model is the one which minimizes the expected posterior predictive loss.

We now consider the problem of optimal prediction from a Bayesian approach. We use the ideas, notations, and results of Barbieri and Berger (2004)

[2] Section 9.8 may be omitted at first reading.

for this part. Consider the canonical model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{9.54}$$

where $\mathbf{y}$ is an $n \times 1$ vector of observations, $\mathbf{X}$ is the $n \times k$ full rank design matrix, $\boldsymbol{\beta}$ is the unknown $k \times 1$ vector of regression coefficients and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of random errors, which are i.i.d. $N(0, \sigma^2)$, σ^2 being known or unknown. Our goal is to predict a future observation y^*, given by

$$y^* = \mathbf{x}^*\boldsymbol{\beta} + \epsilon, \tag{9.55}$$

where $\mathbf{x^*} = (x_1^*, \ldots, x_k^*)$ is the value of the covariate vector for which the prediction is to be made. We consider the loss in predicting y^* by $\hat{y}^*$ as

$$L(\hat{y}^*, y^*) = (\hat{y}^* - y^*)^2; \tag{9.56}$$

i.e., the squared error prediction loss. Assume that we have submodels

$$M_{\mathbf{l}} : \mathbf{y} = \mathbf{X}_{\mathbf{l}}\boldsymbol{\beta}_{\mathbf{l}} + \boldsymbol{\epsilon}, \tag{9.57}$$

where $\mathbf{l} = (l_1, \ldots, l_k)$ with $l_i = 1$ or 0 according as the ith covariate is in the model $M_{\mathbf{l}}$ or not, $\mathbf{X}_{\mathbf{l}}$ is a matrix containing columns of $\mathbf{X}$ corresponding with the nonzero coordinates of $\mathbf{l}$ and $\boldsymbol{\beta}_{\mathbf{l}}$ is the corresponding vector of regression coefficients. Let $k_{\mathbf{l}}$ denote the number of covariates included in the model; then $\mathbf{X}_{\mathbf{l}}$ is of dimension $(n \times k_{\mathbf{l}})$ and $\boldsymbol{\beta}_{\mathbf{l}}$ is a $(k_{\mathbf{l}} \times 1)$ vector.

We put prior probabilities $P(M_{\mathbf{l}})$ to each model $M_{\mathbf{l}}$ included in the model space such that $\sum_{\mathbf{l}} P(M_{\mathbf{l}}) = 1$, and given model $M_{\mathbf{l}}$, a prior $\pi_{\mathbf{l}}(\boldsymbol{\beta}_{\mathbf{l}}, \sigma)$ is assumed on the parameters $(\boldsymbol{\beta}_{\mathbf{l}}, \sigma)$ included in model $M_{\mathbf{l}}$. Using standard posterior calculations, one obtains the quantities (a) $p_{\mathbf{l}} = P(M_{\mathbf{l}}|\mathbf{y})$, the posterior probability of model $M_{\mathbf{l}}$ and (b) $\pi_{\mathbf{l}}(\boldsymbol{\beta}_{\mathbf{l}}, \sigma|\mathbf{y})$, the posterior distribution of the unknown parameters in $M_{\mathbf{l}}$. With this setup in mind, we shall now discuss two optimal prediction strategies, as described below.

First note that the best predictor of y^* for a given value of $\mathbf{x}^*$ comes out as $\bar{y}^* = E(y^*|\mathbf{y})$, where the expectation is taken with respect to the posterior/predictive distribution of y^* given $\mathbf{y}$. This follows by noting that

$$E[(y^* - \hat{y}^*)^2] = E^{\mathbf{y}} E[(y^* - \hat{y}^*)^2|\mathbf{y}], \tag{9.58}$$

where the expectation inside is taken with respect to the posterior distribution of y^* given $\mathbf{y}$. But note that

$$\bar{y^*} = E(y^*|\mathbf{y}) = \sum_{\mathbf{l}} p_{\mathbf{l}} E(y^*|\mathbf{y}, M_{\mathbf{l}}) = \mathbf{x}^* \sum_{\mathbf{l}} p_{\mathbf{l}} H_{\mathbf{l}} \tilde{\boldsymbol{\beta}}_{\mathbf{l}}, \tag{9.59}$$

where $H_{\mathbf{l}}$ is a $(k \times k_{\mathbf{l}})$ matrix such that $\mathbf{x}^* H_{\mathbf{l}}$ is the subvector of $\mathbf{x}^*$ corresponding to the nonzero coordinates of $\mathbf{l}$ and $\tilde{\boldsymbol{\beta}}_{\mathbf{l}}$ is the posterior mean of $\boldsymbol{\beta}_{\mathbf{l}}$ with respect to $\pi_{\mathbf{l}}(\boldsymbol{\beta}_{\mathbf{l}}, \sigma|\mathbf{y})$. Noting that if we knew that $M_{\mathbf{l}}$ were the true model, then the optimal predictor of y^* for $\mathbf{x}$ fixed at $\mathbf{x}^*$ would be given by

$$\hat{y}_{\mathbf{l}}^* = \mathbf{x}^* H_{\mathbf{l}} \tilde{\boldsymbol{\beta}}_{\mathbf{l}}, \text{ we have} \tag{9.60}$$

$$\bar{y^*} = E(y^*|\mathbf{y}) = \mathbf{x}^* \bar{\boldsymbol{\beta}} \equiv \mathbf{x}^* \sum_{\mathbf{l}} p_{\mathbf{l}} H_{\mathbf{l}} \tilde{\boldsymbol{\beta}}_{\mathbf{l}} = \sum_{\mathbf{l}} p_{\mathbf{l}} \hat{y}_{\mathbf{l}}^*. \tag{9.61}$$

$\bar{y^*}$ is called the Bayesian model averaging estimate, in that it is a weighted average of the optimal Bayesian predictors under each individual model, the weights being the posterior probabilities of each model. Many authors have argued the use of the model averaging estimate as an appropriate predictive estimate. They justify this by saying that in using model selection to choose the best model and then making inference based on the assumption that the selected model is true, does not take into account the fact that there is uncertainty about the model itself. As a result, one might underestimate the uncertainty about the quantity of interest. See, for example, Madigan and Raftery (1994), Raftery, Madigan, and Hoeting (1997), Hoeting, Madigan, Raftery, and Volinsky (1999), and Clyde (1999); just to name a few, for detailed discussion on this point of view. However if the number of models in the model space is very large (e.g., in case all subsets of parameters are allowed in the model space, as will happen in high or even moderately high dimensions), the task of computing the Bayesian model averaging estimate exactly might be virtually impossible. Moreover, it is not prudent to keep in the model average those models that have small posterior probability indicating relative incompatibility with observed data. There are some proposals to get around this difficulty, as discussed in the literature cited above. Two of them are based on the 'Occam's window' method of Madigan and Raftery (1994) and the Markov chain Monte Carlo approach of Madigan and York (1995).

In the first approach, the averaging is done over a small set of appropriately selected models, which are parsimonious and supported by data. In the second approach, one constructs a Markov chain with state space same as the model space and equilibrium distribution $\{P(M_{\mathbf{l}}|\mathbf{y})\}$ where $M_{\mathbf{l}}$ varies over the model space. Upon simulation from this chain, the Bayesian model averaging estimator is approximated by taking average value of the posterior expectations under each model visited in the chain. But it must be commented that Bayesian model averaging (BMA) has its limitations in high-dimensional problems. Each approach addresses both issues but it is unclear how well.

Although BMA is the optimal predictive estimation procedure, often a single model is desired for prediction. For example, choice of a single model will require observing only the covariates included in the model. Also, as noted earlier, in high dimensions, BMA has its problems. We will assume now that the future predictions will be made for covariates $\mathbf{x}^*$ such that

$$Q = E(\mathbf{x}^{*\prime}\mathbf{x}^*)$$

exists and is positive definite. A frequent choice of Q is $Q = \mathbf{X}'\mathbf{X}$, i.e., the future covariates will be like the ones observed in the past. In general, the best single model will depend on $\mathbf{x}^*$, but we present here some general characterizations which give the optimal predictive model without this dependence. In general, the optimal predictive model is not the one with the highest posterior probability. However, there are interesting exceptions. If there are only two models, it is easy to show the posterior mode with shrinkage estimate is optimal for prediction (Berger (1997) and Mukhopadhyay (2000)). This also holds sometimes in the context of variable selection for linear models with orthogonal design matrix, as in Clyde and Parmigiani (1996). As Berger (1997) notes, it is easy to see that if one is considering only two models, say M_1 and M_2 with prior probabilities $\frac{1}{2}$ each and proper priors are assigned to the unknown parameters under each model, the best predictive model turns out to be M_1 or M_2 according as the Bayes factor of M_1 to M_2 is greater than one or not, and hence the best predictive model is the one with the highest posterior probability. The characterizations we will describe here are in terms of what is called the 'median probability model.' If it exists, the median probability model $M_{\mathbf{l}^*}$ is defined to be the model consisting of those variables only whose posterior inclusion probabilities are at least $\frac{1}{2}$. The posterior inclusion probability for variable i is

$$p_i = \sum_{\mathbf{l}:l_i=1} P(M_{\mathbf{l}}|\mathbf{y}). \tag{9.62}$$

So, $\mathbf{l}^*$ is defined coordinatewise as $l_i = 1$ if $p_i \geq \frac{1}{2}$ and $l_i = 0$ otherwise. It is possible that the median probability model does not exist, in that the variables included according to the definition of $\mathbf{l}^*$ do not correspond with any model under consideration. But in the variable selection problem, if we are allowed to include or exclude any variable in the possible models, i.e., all possible values of $\mathbf{l}$ are allowed, then the median probability model will obviously exist. Another important class of models is a class of models with 'graphical model structure' for which the median probability model will always exist (this fact follows directly from the definition below).

Definition 9.4. *Suppose that for each variable index i, there is a corresponding index set $I(i)$ of other variables. A subclass of linear models is said to have 'graphical model structure' if it consists of all models satisfying the condition 'for each i, if variable x_i is in the model, then variables x_j with $j \in I(i)$ are in the model.'*

The class of models with 'graphical model structure' includes the class of models with all possible subsets of variables and sequences of nested models, $M_{\mathbf{l}(j)}$, $j = 0, 1, \ldots, k$, where $\mathbf{l}(j) = (1, \ldots, 1, 0, \ldots, 0)$ with j ones and $k - j$ zeros. For the all subsets scenario, $I(i)$ is the null set while in the nested case $I(i) = \{j : 1 \leq j < i\}$ for $i \geq 2$ and $I(i)$ is the null set for $i = 0$ or 1. The latter are natural in many examples including polynomial regression models, where j refers to the degree of polynomial used. Another example of nested models is provided by nonparametric regression (vide Chapter 10, Sections 10.2, 10.3). The unknown function is approximated by partial sums of its Fourier expansion, with all coefficients after stage j assumed to be zero. Note that in this situation, the median probability model has a simple description; one calculates the cumulative sum of posterior model probabilities beginning from the smallest model, and the median probability model is the first model for which this sum equals or exceeds $\frac{1}{2}$. Mathematically, the median probability model is $M_{\mathbf{l}(j^*)}$, where

$$\sum_{i=0}^{j^*-1} P(M_{\mathbf{l}(i)}|\mathbf{y}) < \frac{1}{2} \text{ and } \sum_{i=0}^{j^*} P(M_{\mathbf{l}(i)}|\mathbf{y}) \geq \frac{1}{2}. \tag{9.63}$$

We present some results on the optimality of the posterior median model in prediction. The best predictive model is found as follows. Once a model is selected, the best Bayesian predictor assuming that model is true is obtained. In the next stage, one finds the model such that the expected prediction loss (this expectation does not assume any particular model is true, but is an overall expectation) using this Bayesian predictor is minimized. The minimizer is the best predictive model. There are some situations where the median probability model and the highest posterior probability are the same. Obviously, if there is one model with posterior probability greater than $\frac{1}{2}$, this will be trivially true. Barbieri and Berger (2004) observe that when the highest posterior probability model has substantially larger probability than the other models, it will typically also be the median probability model. We describe another such situation later in the corollary to Theorem 9.8.

We state and prove two simple lemmas.

Lemma 9.5. *(Barbieri and Berger, 2004) Assume Q exists and is positive definite. The optimal model for predicting y^* under the squared error loss, is the unique model minimizing*

$$R(M_{\mathbf{l}}) \equiv (H_{\mathbf{l}}\tilde{\boldsymbol{\beta}}_{\mathbf{l}} - \bar{\boldsymbol{\beta}})'Q(H_{\mathbf{l}}\tilde{\boldsymbol{\beta}}_{\mathbf{l}} - \bar{\boldsymbol{\beta}}), \tag{9.64}$$

where $\bar{\boldsymbol{\beta}}$ is defined in (9.61).

Proof. As noted earlier, $\hat{y}^*_{\mathbf{l}}$ is the optimal Bayesian predictor assuming $M_{\mathbf{l}}$ is the true model. The optimal predictive model is found by minimizing with respect to $\mathbf{l}$, where $\mathbf{l}$ belongs to the space of models under consideration, the quantity $E(y^* - \hat{y}^*_{\mathbf{l}})^2$. Minimizing this is equivalent to minimizing for each $\mathbf{y}$ the quantity $E[(y^* - \hat{y}^*_{\mathbf{l}})^2|\mathbf{y}]$. It is easy to see that for a fixed $\mathbf{x}^*$,

$$E[(y^* - \hat{y}^*_{\mathbf{l}})^2|\mathbf{y}] = C + (\bar{y^*} - \hat{y}^*_{\mathbf{l}})^2, \tag{9.65}$$

where the symbols have been defined earlier and C is a quantity independent of $\mathbf{l}$. The expectation above is taken with respect to the predictive distribution of y^* given $\mathbf{y}$ and $\mathbf{x}^*$. So the optimal predictive model will be found by finding the minimizer of the expression obtained by taking a further expectation over $\mathbf{x}^*$ on the second quantity on the right hand side of (9.65). By plugging in the values of $\hat{y}_{\mathbf{l}}^*$ and $\bar{y}^*$, we immediately get

$$(\bar{y}^* - \hat{y}_{\mathbf{l}}^*)^2 = (H_{\mathbf{l}}\tilde{\boldsymbol{\beta}}_{\mathbf{l}} - \bar{\boldsymbol{\beta}})'\mathbf{x}^{*\prime}\mathbf{x}^*(H_{\mathbf{l}}\tilde{\boldsymbol{\beta}}_{\mathbf{l}} - \bar{\boldsymbol{\beta}}). \tag{9.66}$$

The lemma follows. The uniqueness follows from the fact that Q is positive definite. □

Lemma 9.6. *(Barbieri and Berger, 2004) If Q is diagonal with diagonal elements $q_i > 0$, and the posterior means $\tilde{\boldsymbol{\beta}}_{\mathbf{l}}$ satisfy $\tilde{\boldsymbol{\beta}}_{\mathbf{l}} = \mathbf{H}_{\mathbf{l}}'\tilde{\boldsymbol{\beta}}$ (where $\tilde{\boldsymbol{\beta}}$ is the posterior mean under the full model as in (9.54)) then*

$$R(M_{\mathbf{l}}) = \sum_{i=1}^{k} \tilde{\beta}_i^{\,2} q_i (l_i - p_i)^2. \tag{9.67}$$

Proof. From the fact $\tilde{\boldsymbol{\beta}}_{\mathbf{l}} = \mathbf{H}_{\mathbf{l}}'\tilde{\boldsymbol{\beta}}$, it follows that

$$\bar{\boldsymbol{\beta}} = \sum_{\mathbf{l}} p_{\mathbf{l}}\mathbf{H}_{\mathbf{l}}\tilde{\boldsymbol{\beta}}_{\mathbf{l}} = \sum_{\mathbf{l}} p_{\mathbf{l}}\mathbf{H}_{\mathbf{l}}\mathbf{H}_{\mathbf{l}}'\tilde{\boldsymbol{\beta}} = D(\mathbf{p})\tilde{\boldsymbol{\beta}}, \tag{9.68}$$

where $D(\mathbf{p})$ is the diagonal matrix with diagonal elements p_i, by noting that $H_{\mathbf{l}}(i,j) = 1$ if $l_i = 1$ and $j = \sum_{r=1}^{i} l_r$ and $H_{\mathbf{l}}(i,j) = 0$ otherwise. Similarly,

$$\begin{aligned} R(M_{\mathbf{l}}) &= (\mathbf{H}_{\mathbf{l}}\mathbf{H}_{\mathbf{l}}'\tilde{\boldsymbol{\beta}} - D(\mathbf{p})\tilde{\boldsymbol{\beta}})'Q(\mathbf{H}_{\mathbf{l}}\mathbf{H}_{\mathbf{l}}'\tilde{\boldsymbol{\beta}} - D(\mathbf{p})\tilde{\boldsymbol{\beta}}) \\ &= \tilde{\boldsymbol{\beta}}'(D(\mathbf{l}) - D(\mathbf{p}))Q(D(\mathbf{l}) - D(\mathbf{p}))\tilde{\boldsymbol{\beta}}, \end{aligned} \tag{9.69}$$

from where the result follows. □

Remark 9.7. The condition $\tilde{\boldsymbol{\beta}}_{\mathbf{l}} = \mathbf{H}_{\mathbf{l}}'\tilde{\boldsymbol{\beta}}$, simply means that the posterior mean of $\tilde{\boldsymbol{\beta}}_{\mathbf{l}}$ is found by taking the relevant coordinates of the posterior mean in the full model as in (9.54). As Barbieri and Berger (2004) comment, this will happen in two important cases. Assume $X'X$ is diagonal. In the first case, if one uses the reference prior $\pi_{\mathbf{l}}(\boldsymbol{\beta}_{\mathbf{l}}, \sigma) = 1/\sigma$ or a constant prior if σ is known, the LSE becomes same as the posterior means and the diagonality of $(\mathbf{X}'\mathbf{X})$ implies that the above condition will hold. Secondly, suppose in the full model $\pi(\boldsymbol{\beta}, \sigma) = N_k(\boldsymbol{\mu}, \sigma^2\Delta)$ where Δ is a known diagonal matrix, and for the submodels the natural corresponding prior $N_{k_{\mathbf{l}}}(\mathbf{H}_{\mathbf{l}}'\boldsymbol{\mu}, \sigma^2\mathbf{H}_{\mathbf{l}}'\Delta\mathbf{H}_{\mathbf{l}})$. Then it is easy to see that for any prior on σ^2 or if σ^2 is known, the above will hold.

We now state the first theorem.

Theorem 9.8. *(Barbieri and Berger, 2004) If Q is diagonal with $q_i > 0$ and $\tilde{\boldsymbol{\beta}}_{\mathbf{l}} = \mathbf{H}_{\mathbf{l}}'\tilde{\boldsymbol{\beta}}$, and the models have graphical model structure, then the median probability model is the best predictive model.*

Proof. Because $q_i > 0$, $\tilde{\beta}_i^2 \geq 0$ for each i and p_i (defined in (9.62)) does not depend on $\mathbf{l}$, to minimize $R(M_\mathbf{l})$ among all possible models, it suffices to minimize $(l_i - p_i)^2$ for each individual i and that is achieved by choosing $l_i = 1$ if $p_i \geq \frac{1}{2}$ and $l_i = 0$ if $p_i < \frac{1}{2}$, whence $\mathbf{l}$ as defined will be the median probability model. The graphical model structure ensures that this model is among the class of models under consideration. □

Remark 9.9. The above theorem obviously holds if we consider all submodels, this class having graphical model structure; provided the conditions of the theorem hold. By the same token, the result will hold under the situation where the models under consideration are nested.

Corollary 9.10. *(Barbieri and Berger, 2004) If the conditions of the above theorem hold, all submodels of the full model are allowed, σ^2 is known, $\mathbf{X'X}$ is diagonal and β_i's have $N(\mu_i, \lambda_i\sigma^2)$ distributions and*

$$P(M_\mathbf{l}) = \prod_{i=1}^{k} (p_i^0)^{l_i}(1 - p_i^0)^{(1-l_i)}, \tag{9.70}$$

where p_i^0 is the prior probability that variable x_i is in the model, then the optimal predictive model is the model with highest posterior probability which is also the median probability model.

Proof. Let $\hat{\beta}_i$ be the least squares estimate of β_i under the full model. Because $X'X$ is diagonal, $\hat{\beta}_i$'s are independent and the likelihood under M_l factors as

$$L(M_\mathbf{l}) \propto \prod_{i=1}^{k} (\lambda_i^0)^{l_i}(\lambda_i')^{1-l_i}$$

where λ_i^0 depends only on $\hat{\beta}_i$ and β_i, λ_i' depends only on $\hat{\beta}_i$ and the constant of proportionality here and below depend an $\boldsymbol{Y}$ and $\hat{\beta}_i$'s.

Also, the conditional prior distribution of β_i's given M_l has a factorization

$$\pi(\boldsymbol{\beta}|M_\mathbf{l}) = \prod_{i=1}^{k} [N(\mu_i, \lambda_i\sigma^2)]^{l_i}[\delta\{0\}]^{1-l_i}$$

where $\delta\{0\}$ = degenerate distribution with all mass at zero.

It follows from (9.70) and the above two factorizations that the posterior probability of M_l has a factorization

$$P(M_\mathbf{l}|\boldsymbol{Y}) \alpha \prod_{i=1}^{k} \{p_i^0 \int_{-\infty}^{\infty} \lambda_i^0 N(\mu_i, \lambda_i\sigma^2) d\beta\}^{l_i} \{(1 - p_i^0)\lambda_i'\delta\{0\}\}^{1-l_i}$$

which in turn implies that the marginal posterior of including or not including ith variable is proportional to the two terms respectively in the ith factor. This completes the proof, vide Problem 21. (The integral can be evaluated as in Chapter 2.) □

We have noted before that if the conditions in Theorem 9.8 are satisfied and the models are nested, then the best predictive model is the median probability model. Interestingly even if Q is not necessarily diagonal, the best predictive model turns out to be the median probability model under some mild assumptions, in the nested model scenario. Consider
Assumption 1: $Q = \gamma \mathbf{X}'\mathbf{X}$ for some $\gamma > 0$, i.e., the prediction will be made at covariates that are similar to the ones already observed in the past.
Assumption 2: $\tilde{\boldsymbol{\beta}}_{\mathbf{l}} = b\hat{\boldsymbol{\beta}}_{\mathbf{l}}$, where $b > 0$, i.e, the posterior means are proportional to the least squares estimates.

Remark 9.11. Barbieri and Berger (2004) list two situations when the second assumption will be satisfied. First, if one uses the reference prior $\pi_{\mathbf{l}}(\boldsymbol{\beta}_{\mathbf{l}}, \sigma) = 1/\sigma$, whereby the posterior means will be the LSE's. It will also be satisfied with $b = c/(1+c)$, if one uses g-type normal priors of Zellner (1986), where $\pi_{\mathbf{l}}(\boldsymbol{\beta}_{\mathbf{l}}|\sigma) \sim N_{k_{\mathbf{l}}}(\mathbf{0}, c\sigma^2(\mathbf{X}_{\mathbf{l}}'\mathbf{X}_{\mathbf{l}})^{-1})$ and the prior on σ is arbitrary.

Theorem 9.12. *For a sequence of nested models for which the above two conditions hold, the best predictive model is the median probability model.*

Proof. See Barbieri and Berger (2004). □

Barbieri and Berger(2004, Section 5) present a geometric formulation for identification of the optimal predictive model. They also establish conditions under which the median probability model and the maximum posterior probability model coincides; and that it is typically not enough to know only the posterior probabilities of each model to determine the optimal predictive model.

Till now we have concentrated on some Bayesian approaches to the prediction problem. It turns out that model selection based on the classical Akaike information criterion (AIC) also plays an important role in Bayesian prediction and estimation for linear models and function estimation. Optimality results for AIC in classical statistics are due to Shibata (1981, 1983), Li (1987), and Shao (1997).

The first Bayesian result about AIC is taken from Mukhopadhyay (2000). Here one has observations $\{y_{ij} : i = 1, \ldots, p,\ j = 1, \ldots, r,\ n = pr\}$ given by

$$y_{ij} = \mu_i + \epsilon_{ij}, \tag{9.71}$$

where ϵ_{ij} are i.i.d. $N(0, \sigma^2)$ with σ^2 known. The models are $M_1 : \mu_i = 0$ for all i and $M_2 : \eta^2 = \lim_{p->\infty} \frac{1}{p} \sum_{i=1}^{p} \mu_i^2 > 0$. Under M_2, we assume a $N(0, \tau^2 I_p)$ prior on μ where τ^2 is to be estimated from data using an empirical Bayes method. It is further assumed that $p \to \infty$ as $n \to \infty$. The goal is to predict a future set of observations $\{z_{ij}\}$ independent of $\{y_{ij}\}$ using the usual prediction error loss, with the 'constraint' that once a model is selected, least squares estimates have to be used to make the predictions. Theorem 9.13 shows that the constrained empirical Bayes rule is equivalent to AIC asymptotically. A weaker result is given as Problem 17.

Theorem 9.13. *(Mukhopadhyay, 2000) Suppose M_2 is true, then asymptotically the constrained empirical Bayes rule and AIC select the same model. Under M_1, AIC and the constrained empirical Bayes rule choose M_1 with probability tending to 1. Also under M_1, the constrained empirical Bayes rule chooses M_1 whenever AIC does so.*

The result is extended to general nested problems in Mukhopadhyay and Ghosh (2004a). It is however also shown in the above reference that if one uses Bayes estimates instead of least squares estimates, then the unconstrained Bayes rule does better than AIC asymptotically. The performance of AIC in the PEB setup of George and Foster (2000) is studied in Mukhopadhyay and Ghosh (2004a).

As one would expect from this, AIC also performs well in nonparametric regression which can be formulated as an infinite dimensional linear problem. It is shown in Chakrabarti and Ghosh (2005b) that AIC attains the optimal rate of convergence in an asymptotically equivalent problem and is also adaptive in the sense that it makes no assumption about the degree of smoothness. Because this result is somewhat technical, we only present some numerical results for the problem of nonparametric regression.

In the nonparametric regression problem

$$Y_i = f(\frac{i}{n}) + \epsilon_i,\ i = 1, \ldots, n, \tag{9.72}$$

one has to estimate the unknown smooth function f. In Table 9.3, we consider $n = 100$ and $f(x) = (\sin(2\pi x))^3$, $(\cos(\pi x))^4$, $7+\cos(2\pi x)$, and $e^{\sin(2\pi x)}$, the loss function $L(f, \hat{f}) \equiv \int_0^1 (f(x) - \hat{f}(x))^2 dx$, and report the average loss of modified James-Stein estimator of Cai et al. (2000), AIC, and the kernel method with Epanechnikov kernel in 50 simulations. To use the first two methods, we express f in its (partial sum) Fourier expansion with respect to the usual sine-cosine Fourier basis of $[0, 1]$ and then estimate the Fourier coefficients by the regression coefficients. Some simple but basic insight about the AIC may be obtained from Problems 15–17. It is also worth remembering that AIC was expected by Akaike to perform well in high-dimensional estimation or prediction problem when the true model is too complex to be in the model space.

9.9 Discussion

Bayesian model selection is passing through a stage of rapid growth, especially in the context of bioinformatics and variable selection. The two previous sections provide an overview of some of the literature. See also the review by Ghosh and Samanta (2001). For a very clear and systematic approach to different aspects of model selection, see Bernardo and Smith (1994).

Model selection based on AIC is used in many real-life problems by Burnham and Anderson (2002). However, its use for testing problems with 0-1

Table 9.3. Comparison of Simulation Performance of Various Estimation Methods in Nonparametric Regression

Function	Modified James-Stein	*AIC*	Kernel Method
$[Sin(2\pi x)]^3$	0.2165	0.0793	0.0691
$[Cos(\pi x)]^4$	0.2235	0.078	0.091
$7 + Cos(2\pi x)$	0.2576	0.0529	0.5380
$e^{Sin(2\pi x)}$	0.2618	0.0850	0.082

loss is questionable vide Problem 16. A very promising new model selection criterion due to Spiegelhalter et al. (2002) may also be interpreted as a generalization of AIC, see, e.g., Chakrabarti and Ghosh (2005a). In the latter paper, GBIC is also interpreted from the information theoretic point of view of Rissanen (1987).

We believe the Bayesian approach provides a unified approach to model selection and helps us see classical rules like BIC and AIC as still important but by no means the last word in any sense. We end this section with two final comments.

One important application of model selection is to examine model fit. Gelfand and Ghosh (1998) (see also Gelfand and Dey (1994)) use leave-k-out cross-validation to compare each collection of k data points and their predictive distribution based on the remaining observations. Based on the predictive distributions, one may calculate predicted values and some measure of deviation from the k observations that are left out. An average of the deviation over all sets of k left out observations provides some idea of goodness of fit. Gelfand and Ghosh (1998) use these for model selection. Presumably, the average distance for a model can be used for model check also. An interesting work of this kind is Bhattacharya (2005).

Another important problem is computation of the Bayes factor. Gelfand and Dey (1994) and Chib (1995) show how one can use MCMC calculations by relating the marginal likelihood of data to the posterior via $P(y) = L(\theta|y)P(\theta)/P(\theta|y)$. Other relevant papers are Carlin and Chib (1995), Chib and Greenberg (1998), and Basu and Chib (2003). There are interesting suggestions also in Gelman et al (1995).

9.10 Exercises

1. Show that $\pi(\eta_2|\mathbf{X})$ is an improper density if we take $\pi(\eta_1, \eta_2) = 1/\eta_2$ in Example 9.3.
2. Justify (9.2) and (9.3).
3. Complete the details to implement Gibbs sampling and E-M algorithm in Example 9.3 when $\boldsymbol{\mu}$ and σ^2 are unknown. Take $\pi(\eta_1, \sigma^2, \eta_2) = 1/\sigma^2$.
4. Let X_i's be independent with density $f(x|\theta_i)$, $i = 1, 2, \ldots, p$, $\theta_i \in \mathcal{R}$. Consider the problem of estimating $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ with loss function

$$L(\boldsymbol{\theta}, \boldsymbol{a}) = \sum_{i=1}^{p} L(\theta_i, a_i) = \sum_{i=1}^{p} (\theta_i - a_i)^2, \quad \boldsymbol{\theta}, \boldsymbol{a} \in \mathcal{R}^p.$$

i.e., the total loss is the sum of the losses in estimating θ_i by a_i. An estimator for $\boldsymbol{\theta}$ is the vector $(T_1(\boldsymbol{X}), T_2(\boldsymbol{X}), \ldots, T_p(\boldsymbol{X}))$. We call this a compound decision problem with p components.

(a) Suppose $\sup_\delta f(x|\delta) = f(x|T(x))$, i.e., $T(x)$ is the MLE (of θ_j in $f(x|\theta_j)$). Show that $(T(X_1), T(X_2), \ldots T(X_p))$ is the MLE of $\boldsymbol{\theta}$.

(b) Suppose $T(X)$ (not necessarily the $T(X)$ of (a)) satisfies the sufficient condition for a minimax estimate given at the end of Section 1.5. Is $(T(X_1), T(X_2), \ldots, T(X_p))$ minimax for $\boldsymbol{\theta}$ in the compound decision problem?

(c) Suppose $T(X)$ is the Bayes estimate with respect to squared error loss for estimating θ of $f(x|\theta)$. Is $(T(X_1), \ldots, T(X_p))$ a Bayes estimate for $\boldsymbol{\theta}$?

(d) Suppose $\mathbf{T} = (T_1(X_1), \ldots, T_p(X_p))$ and $T_j(X_i)$ is admissible in the jth component decision problem. Is $\mathbf{T}$ admissible?

5. Verify the claim of the best unbiased predictor (9.17).
6. Given the hierarchical prior of Section 9.3 for Morris's regression setup, calculate the posterior and the Bayes estimate as explicitly as possible. Find the full conditionals of the posterior distribution in order to implement MCMC.
7. Prove the claims of superiority made in Section 9.4 for the James-Stein-Lindley estimate and the James-Stein positive part estimate using Stein's identity.
8. Under the setup of Section 9.3, show that the PEB risk of $\hat{\theta}_i$ is smaller than the PEB risk of Y_i.
9. Refer to Sections 9.3 and 9.4. Compare the PEB risk of $\hat{\theta}_i$ and Stein's frequentist risk of $\hat{\boldsymbol{\theta}}$ and show that the two risks are of the same form but one has $E(\hat{B})$ and the other $E_\theta(\hat{B})$. (Hint: See equations (1.17) and (1.18) of Morris (1983)).
10. Consider the setup of Section 9.3. Show that $\hat{B}$ is the best unbiased estimate of B.
11. (Disease mapping) (See Section 10.1 for more details on the setup.) Suppose that the area to be mapped is divided into N regions. Let O_i and E_i be respectively the observed and expected number of cases of a disease in the ith region, $i = 1, 2, \ldots, N$. The unknown parameters of interest are θ_i, the relative risk in the ith region, $i = 1, 2, \ldots, N$. The traditional model for O_i is the Poisson model, which states that given $(\theta_1, \ldots, \theta_N)$, O_i's are independent and
$$O_i|\theta_i \sim \text{ Poisson } (E_i\theta_i).$$
Let $\theta_1, \theta_2, \ldots, \theta_N$ be i.i.d. $\sim$ Gamma(a, b). Find the PEB estimates of $\theta_1, \theta_2, \ldots, \theta_N$. In Section 10.1, we will consider hierarchical Bayes analysis for this problem.

12. Let Y_i be i.i.d $N(\theta_i, V)$, $i = 1, 2, \ldots, p$. Stein's heuristics (Section 9.4) shows $\|\boldsymbol{Y}\|^2$ is too large in a frequentist sense. Verify by a similar argument that if θ_i are i.i.d uniform on $\mathcal{R}$ then $\|\boldsymbol{Y}\|^2$ is too small in an improper Bayesian sense, i.e., there is extreme divergence between frequentist probability and naive objective Bayes probability in a high-dimensional case.
13. (Berger (1985a, p. 542)) Consider a multiparameter exponential family $f(\boldsymbol{x}|\boldsymbol{\theta}) = c(\boldsymbol{\theta}) \exp(\boldsymbol{\theta}' T(\boldsymbol{x})) h(\boldsymbol{x})$, where $\boldsymbol{x}$ and $\boldsymbol{\theta}$ are vectors of the same dimension. Assuming Stein's loss, show that (under suitable conditions) the Bayes estimate can be written as gradient$(\log m(\boldsymbol{x}))$ − gradient$(\log h(\boldsymbol{x}))$ where $m(\boldsymbol{x})$ is the marginal density of $\boldsymbol{x}$ obtained by integrating out $\boldsymbol{\theta}$.
14. Simulate data according to the model in Example 9.3, Section 9.1.
    (a) Examine how well the model can be checked from the data X_{ij}, $i = 1, 2, \ldots n$, $j = 1, 2, \ldots p$.
    (b) Suppose one uses the empirical distribution of $\bar{X}_j$'s as a surrogate prior for μ_j's. Compare critically the Bayes estimate of $\boldsymbol{\mu}$ for this prior with the PEB estimate.
15. (Stone's problem) Let
    $Y_{ij} = \alpha + \mu_i + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2), i = 1, 2, \ldots, p, j = 1, 2, \ldots, r, n = pr$ with σ^2 assumed known or estimated by $S^2 = \sum_{i=1}^{p} \sum_{j=1}^{r} (Y_{ij} - \bar{Y}_i)^2 / p(r-1)$. The two models are
    $$M_1 : \mu_i = 0 \forall i \text{ and } M_2 : \boldsymbol{\mu} \in \mathcal{R}^p.$$
    Suppose $n \to \infty, p \log n / n \to \infty$ and $\sum_{i=1}^{p} (\mu_i - \bar{\mu})^2 / (p-1) \to \tau^2 > 0$.
    (a) Show that even though M_2 is true, BIC will select M_1 with probability tending to 1. Also show that AIC will choose the right model M_2 with probability tending to one.
    (b) As a Bayesian how important do you think is this notion of consistency?
    (c) Explore the relation between AIC and selection of model based on estimation of residual sum of squares by leave-one-out cross validation.
16. Consider an extremely simple testing problem. $X \sim N(\mu, 1)$. You have to test $H_0 : \mu = 0$ versus $H_1 : \mu \neq 0$. Is AIC appropriate for this? Compare AIC, BIC, and the usual likelihood ratio test, keeping in mind the conflict between P-values and posterior probability of the sharp null hypothesis.
17. Consider two nested models and an empirical Bayes model selection rule with the evaluation based on the more complex model. Though you know the more complex model is true, you may be better off predicting with the simpler model.
    Let $Y_{ij} = \mu_i + \epsilon_{ij}$, ϵ_{ij} i.i.d $N(0, \sigma^2)$, $i = 1, 2, \ldots, p$, $j = 1, 2, \ldots, r$ with known σ^2. The models are
    $$M_1 : \boldsymbol{\mu} = 0$$
    $$M_2 : \boldsymbol{\mu} \in \mathcal{R}^p, \boldsymbol{\mu} \sim N_p(0, \tau^2 I_p), \tau^2 > 0.$$
    (a) Assume that in PEB evaluation under M_2 you estimate τ^2 by the moment estimate:

$$\hat{\tau}^2 = \left[\frac{1}{p} \sum_{i=1}^{p} \bar{Y}_i^2 - \frac{\sigma^2}{r} \right]^+ .$$

Show with PEB evaluation of risk under M_2 and $M_1, \bar{\boldsymbol{Y}}$ is preferred if and only if AIC selects M_2.

(b) Why is it desirable to have large p in this problem?

(c) How will you try to justify in an intuitive way occasional choice of the simple but false model?

(d) Use (a) to motivate how the penalty coefficient 2 arises in AIC.

(This problem is based on a result in Mukhopadhyay (2001)).

18. Burnham and Anderson (2002) generated data to mimic a real-life experiment of Stromberg et al. (1998). Select a suitable model from among the 9 models considered by Ghosh and Samanta (2001). The main issue is computation of the integrated likelihood under each model. You can try Laplace approximation, the method based on MCMC suggested at the end of Section 9.9, and importance sampling. All methods are difficult, but they give very close answers in this problem. The data and the models can be obtained from the Web page `http://www.isical.ac.in/~tapas/book`
19. Let $X_i \sim N(\mu, 1), i = 1, \ldots, n$ and $\mu \sim N(\eta_1, \eta_2)$. Find the PEB estimate of η_1 and η_2 and examine its implications for the inadequacy of the PEB approach in low-dimensional problems.
20. Consider NPEB multiple testing (Section 9.6.1) with known π_1 and an estimate $\hat{f}$ of $(1-\pi_1)f_0 + \pi_1 f_1$. Suppose for each i, you reject $H_{0i} : \mu_i = 0$ if

$$f_0(x_i) \le \hat{f}(x_i)\alpha, \text{ where } o < \alpha < 1.$$

Examine whether this test provides any control on the (frequentist) FDR. Define a Bayesian FDR and examine if, for small π_1, this is also controlled by the test. Suggest a test that would make the Bayesian FDR approximately equal to α. (The idea of controlling a Bayesian FDR is due to Storey (2003). The simple rules in this problem are due to Bogdan, Ghosh, and Tokdar (personal communication).)

21. For all subsets variable selection models show that the posterior median model and the posterior mode model are the same if

$$P(M_l|X) = \prod_{i=1}^{p} p_i^{l_i} (1 - p_i)^{1-l_i}$$

where $l_i = 1$ if the ith variable is included in M_l and $l_i = 0$ otherwise.

10

Some Applications

The popularity of Bayesian methods in recent times is mainly due to their successful applications to complex high-dimensional real-life problems in diverse areas such as epidemiology, microarrays, pattern recognition, signal processing, and survival analysis. This chapter presents a few such applications together with the required methodology. We describe the method without going into the details of the critical issues involved, for which references are given. This is followed by an application involving real or simulated data.

We begin with a hierarchical Bayesian modeling of spatial data in Section 10.1. This is in the context of disease mapping, an area of epidemiological interest. The next two sections, 10.2 and 10.3, present nonparametric estimation of regression function using wavelets and Dirichlet multinomial allocation. They may also be treated as applications involving Bayesian data smoothing. For several recent advances in Bayesian nonparametrics, see Dey et al. (1998) and Ghosh and Ramamoorthi (2003).

10.1 Disease Mapping

Our first application is from the area of epidemiology and involves hierarchical Bayesian spatial modeling. Disease mapping provides a geographical distribution of a disease displaying some index such as the relative risk of the disease in each subregion of the area to be mapped. Suppose that the area to be mapped is divided into N regions. Let O_i and E_i be respectively the observed and expected number of cases of a disease in the ith region, $i = 1, 2, \ldots, N$. The unknown parameters of interest are θ_i, the relative risk in the ith region, $i = 1, 2, \ldots, N$. Here E_i is a simple-minded expectation assuming all regions have the same disease rate (at least after adjustment for age), vide Banerjee et al. (2004, p. 158). The relative risk θ_i is the regional effect in a multiplicative model of expected number of cases: $E(O_i) = E_i\theta_i$. If $\theta_i = 1$, we have $E(O_i) = E_i$. The objective is to make inference about θ_i's across regions. Among other things, this helps epidemiologists and public health professionals to identify regions or cluster of regions having high relative risks and hence needing attention and also to identify covariates causing high relative risk. The traditional model for O_i is the Poisson model, which states that given $(\theta_1, \ldots, \theta_N)$, O_i's are independent and

$$O_i|\theta_i \sim \text{ Poisson } (E_i\theta_i). \tag{10.1}$$

Under this model E_i's are assumed fixed. The classical maximum likelihood estimate of θ_i is $\hat{\theta}_i = O_i/E_i$, known as the standardized mortality ratio (SMR) for region i and $\text{Var}(\hat{\theta}_i) = \theta_i/E_i$, which may be estimated as $\hat{\theta}_i/E_i$. However, it was noted in Chapter 9 that the classical estimates may not be appropriate here for simultaneous estimation of the parameters $\theta_1, \theta_2, \ldots, \theta_N$.

As mentioned in Chapter 9, because of the assumption of exchangeability of $\theta_1, \ldots, \theta_N$, there is a natural Bayesian solution to the problem. A Bayesian modeling involves specification of prior distribution of $(\theta_1, \ldots \theta_N)$. Clayton and Kaldor (1987) followed the empirical Bayes approach using a model that assumes

$$\theta_1, \theta_2, \ldots, \theta_N \text{ i.i.d. } \sim \text{ Gamma } (a, b) \tag{10.2}$$

and estimating the hyperparameters a and b from the marginal density of $\{O_i\}$ given a, b (see Section 9.2). Here we present a full Bayesian approach adopting a prior model that allows for spatial correlation among the θ_i's. A natural extension of (10.2) could be a multivariate Gamma distribution for $(\theta_1, \ldots, \theta_N)$. We, however, assume a multivariate normal distribution for the log-relative risks $\log \theta_i,\ i = 1, \ldots, N$. The model may also be extended to allow for explanatory covariates $\boldsymbol{x}_i$ which may affect the relative risk. Thus we consider the following hierarchical Bayesian model

$$\begin{aligned} O_i|\theta_i \text{ are independent } &\sim \text{ Poisson } (E_i\theta_i) \\ \text{where } \log \theta_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + \phi_i,\ i = 1, \ldots, N. \end{aligned} \tag{10.3}$$

The usual prior for $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)$ is given by the conditionally autoregressive (CAR) model (Besag, 1974), which is briefly described below. For details see, e.g., Besag (1974) and Banerjee et al. (2004, pp. 79–83, 163, 164). Suppose the full conditionals are specified as

$$\phi_i|\phi_j, j \neq i \sim N(\sum_{j \neq i} a_{ij}\phi_j, \sigma_i^2),\ i = 1, 2, \ldots, N. \tag{10.4}$$

These will lead to a joint distribution having density proportional to

$$\exp\left\{-\frac{1}{2}\boldsymbol{\phi}' D^{-1}(I - A)\boldsymbol{\phi}\right\} \tag{10.5}$$

where $D = \text{Diag}(\sigma_1^2, \ldots, \sigma_N^2)$ and $A = (a_{ij})_{N \times N}$. We look for a model that allows for spatial correlation and so consider a model where correlation depends on geographical proximity. A proximity matrix $W = (w_{ij})$ is an $N \times N$ matrix where w_{ij} spatially connects regions i and j in some manner. We consider here binary choices. We set $w_{ii} = 0$ for all i, and for $i \neq j$, $w_{ij} = 1$ if i is a neighbor of j, i.e., i and j share some common boundary and $w_{ij} = 0$ otherwise. Also, w_{ij}'s in each row may be standardized as $\widetilde{w}_{ij} = w_{ij}/w_{io}$ where $w_{io} = \sum_j w_{ij}$ is the number of neighbors of region i. Returning to our model (10.5), we now set $a_{ij} = \alpha w_{ij}/w_{io}$ and $\sigma_i^2 = \lambda/w_{io}$. Then (10.5) becomes

$$\exp\left\{-\frac{1}{2\lambda}\boldsymbol{\phi}'(D_w - \alpha W)\boldsymbol{\phi}\right\}$$

where $D_w = \mathrm{Diag}(w_{10}, w_{20}, \ldots, w_{N0})$. This also ensures that $D^{-1}(I - A) = \frac{1}{\lambda}(D_w - \alpha W)$ is symmetric.

Thus the prior for $\boldsymbol{\phi}$ is multivariate normal

$$\boldsymbol{\phi} \sim N(\mathbf{0}, \Sigma) \text{ with } \Sigma = \lambda(D_w - \alpha W)^{-1}. \tag{10.6}$$

We take $0 < \alpha < 1$, which ensures propriety of the prior and positive spatial correlation; only the values of α close to 1 give enough spatial similarity. For $\alpha = 1$ we have the standard improper CAR model. One may use the improper CAR prior because it is known that the posterior will typically emerge as proper. For this and other relative issues, see Banerjee et al. (2004).

Having specified priors for all the unknown parameters including the spatial variance parameter λ and propriety parameter α $(0 < \alpha < 1)$, one can now do Bayesian analysis using MCMC techniques. We illustrate through an example.

Example 10.1. Table 10.1 presents data from Clayton and Kaldor (1987) on observed (O_i) and expected (E_i) cases of lip cancer during the period 1975–1980 for $N = 56$ counties of Scotland. Also available are x_i, values of a covariate, the percentage of the population engaged in agriculture, fishing, and forestry (AFF), for the 56 counties. The log-relative risk is modeled as

$$\log \theta_i = \beta_0 + \beta_1 x_i + \phi_i, \;\; i = 1, \ldots, N \tag{10.7}$$

where the prior for $(\phi_1, \ldots, \phi_N)$ is as specified in (10.6). We use vague priors for β_0 and β_1 and a prior having high concentration near 1 for the parameter α. The data may be analyzed using WinBUGS. A WinBUGS code for this example is put in the web page of Samanta. A part of the results – the Bayes estimates $\widehat{\theta}_i$ of the relative risks for the 56 counties – are presented in Table 10.1. The θ_i's are smoothed by pooling the neighboring values in an automatic adaptive way as suggested in Chapter 9. The estimates of β_0 and β_1 are obtained as $\widehat{\beta}_0 = -0.2923$ and $\widehat{\beta}_1 = 0.3748$ with estimates of posterior s.d. equal to 0.3426 and 0.1325, respectively.

Table 10.1. Lip Cancer Incidence in Scotland by County: Observed Numbers (O_i), Expected Numbers (E_i), Values of the Covariate AFF (x_i), and Bayes Estimates of the Relative Risk ($\widetilde{\theta}_i$).

County	O_i	E_i	x_i	$\widetilde{\theta}_i$	County	O_i	E_i	x_i	$\widetilde{\theta}_i$
1	9	1.4	16	4.705	29	16	14.4	10	1.222
2	39	8.7	16	4.347	30	11	10.2	10	0.895
3	11	3.0	10	3.287	31	5	4.8	7	0.860
4	9	2.5	24	2.981	32	3	2.9	24	1.476
5	15	4.3	10	3.145	33	7	7.0	10	0.966
6	8	2.4	24	3.775	34	8	8.5	7	0.770
7	26	8.1	10	2.917	35	11	12.3	7	0.852
8	7	2.3	7	2.793	36	9	10.1	0	0.762
9	6	2.0	7	2.143	37	11	12.7	10	0.886
10	20	6.6	16	2.902	38	8	9.4	1	0.601
11	13	4.4	7	2.779	39	6	7.2	16	1.008
12	5	1.8	16	3.265	40	4	5.3	0	0.569
13	3	1.1	10	2.563	41	10	18.8	1	0.532
14	8	3.3	24	2.049	42	8	15.8	16	0.747
15	17	7.8	7	1.809	43	2	4.3	16	0.928
16	9	4.6	16	2.070	44	6	14.6	0	0.467
17	2	1.1	10	1.997	45	19	50.7	1	0.431
18	7	4.2	7	1.178	46	3	8.2	7	0.587
19	9	5.5	7	1.912	47	2	5.6	1	0.470
20	7	4.4	10	1.395	48	3	9.3	1	0.433
21	16	10.5	7	1.377	49	28	88.7	0	0.357
22	31	22.7	16	1.442	50	6	19.6	1	0.507
23	11	8.8	10	1.185	51	1	3.4	1	0.481
24	7	5.6	7	0.837	52	1	3.6	0	0.447
25	19	15.5	1	1.188	53	1	5.7	1	0.399
26	15	12.5	1	1.007	54	1	7.0	1	0.406
27	7	6.0	7	0.946	55	0	4.2	16	0.865
28	10	9.0	7	1.047	56	0	1.8	10	0.773

10.2 Bayesian Nonparametric Regression Using Wavelets

Let us recall the nonparametric regression problem that was stated in Example 6.1. In this problem, it is of interest to fit a general regression function to a set of observations. It is assumed that the observations arise from a real-valued regression function defined on an interval on the real line. Specifically, we have

$$y_i = g(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \text{ and } x_i \in \mathcal{T}, \tag{10.8}$$

where ε_i are i.i.d. $N(0, \sigma^2)$ errors with unknown error variance σ^2, and g is a function defined on some interval $\mathcal{T} \subset \mathcal{R}^1$.

It can be immediately noted that a Bayesian solution to this problem involves specifying a prior distribution on a large class of regression functions. In general, this is a rather difficult task. A simple approach that has been successful is to decompose the regression function g into a linear combination of a set of basis functions and to specify a prior distribution on the regression coefficients. In our discussion here, we use the (orthonormal) wavelet basis. We provide a very brief non-technical overview of wavelets including multiresolution analysis (MRA) here, but for a complete and thorough discussion refer to Ogden (1997), Daubechies (1992), Hernández and Weiss (1996), Müller and Vidakovic (1999), and Vidakovic (1999).

10.2.1 A Brief Overview of Wavelets

Consider the function

$$\psi(x) = \begin{cases} 1 & 0 \le x < 1/2; \\ -1 & 1/2 \le x \le 1; \\ 0 & \text{otherwise.} \end{cases} \tag{10.9}$$

which is known as the Haar wavelet, simplest of the wavelets. Note that its dyadic dilations along with integer translations, namely,

$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k), \quad j, k \in \mathcal{Z}, \tag{10.10}$$

provide a complete orthonormal system for $\mathcal{L}^2(\mathcal{R})$. This says that any $f \in \mathcal{L}^2(\mathcal{R})$ can be approximated arbitrarily well using step functions that are simply linear combinations of wavelets $\psi_{j,k}(x)$. What is more interesting and important is how a finer approximation for f can be written as an orthogonal sum of a coarser approximation and a detail function. In other words, for $j \in \mathcal{Z}$, let

$$\begin{aligned} V_j = \Big\{ f \in \mathcal{L}^2(\mathcal{R}) : \ & f \text{ is piecewise constant on intervals} \\ & [k2^{-j}, (k+1)2^{-j}),\, k \in \mathcal{Z} \Big\}. \end{aligned} \tag{10.11}$$

Now suppose $P^j f$ is the projection of $f \in \mathcal{L}^2(\mathcal{R})$ onto V_j. Then note that

$$\begin{aligned} P^j f &= P^{j-1} f + g^{j-1} \\ &= P^{j-1} f + \sum_{k \in \mathcal{Z}} < f, \psi_{j-1,k} > \psi_{j-1,k}, \end{aligned} \tag{10.12}$$

with g^{j-1} being the detail function as shown, so that

$$V_j = V_{j-1} \oplus W_{j-1}, \tag{10.13}$$

where $W_j = \text{span}\,\{\psi_{j,k}, k \in \mathcal{Z}\}$. Also, corresponding with the 'mother' wavelet ψ (Haar wavelet in this case), there is a father wavelet or scaling function

$\phi = I_{[0,1]}$ such that $V_j = \text{span}\{\phi_{j,k}, k \in \mathcal{Z}\}$, where $\phi_{j,k}$ is the dilation and translation of ϕ similar to the definition (10.10), i.e.,

$$\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k), \quad j, k \in \mathcal{Z}, \tag{10.14}$$

In fact, the sequence of subspaces $\{V_j\}$ has the following properties:

1. $\cdots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots$.
2. $\cap_{j\in\mathcal{Z}} V_j = \{0\}, \overline{\cup_{j\in\mathcal{Z}} V_j} = \mathcal{L}^2(\mathcal{R})$.
3. $f \in V_j$ iff $f(2.) \in V_{j+1}$.
4. $f \in V_0$ implies $f(. - k) \in V_0$ for all $k \in \mathcal{Z}$.
5. There exists $\phi \in V_0$ such that $\quad \text{span}\{\phi_{0,k} = \phi(. - k), k \in \mathcal{Z}\} = V_0$.

Given this ϕ, the corresponding ψ can be easily derived (see Ogden (1997) or Vidakovic (1999)). What is interesting and useful to us is that there exist scaling functions ϕ with desirable features other than the Haar function. Especially important are Daubechies wavelets that are compactly supported and each having a different degree of smoothness.

Definition: Closed subspaces $\{V_j\}_{j\in\mathcal{Z}}$ satisfying properties 1–5 are said to form a multi-resolution analysis (MRA) of $\mathcal{L}^2(\mathcal{R})$. If $V_j = \text{span}\{\phi_{j,k}, k \in \mathcal{Z}\}$ form an MRA of $\mathcal{L}^2(\mathcal{R})$, then the corresponding ϕ is also said to generate this MRA.

In statistical inference, we deal with finite data sets, so wavelets with compact support are desirable. Further, the regression functions (or density functions) that we need to estimate are expected to have certain degree of smoothness. Therefore, the wavelets used here should have some smoothness also. The Haar wavelet does have compact support but is not very smooth. In the application discussed later, we use wavelets from the family of compactly supported smooth wavelets introduced by Daubechies (1992). These, however, cannot be expressed in closed form. A sketch of their construction is as follows.

Because, from property 5 above of MRA, $\phi \in V_0 \subset V_1$, we have

$$\phi(x) = \sum_{k\in\mathcal{Z}} h_k \phi_{1,k}(x), \tag{10.15}$$

where the ‘filter’ coefficients h_k are given by

$$h_k =< \phi, \phi_{1,k} >= \sqrt{2}\int \phi(x)\phi(2x - k)\,dx. \tag{10.16}$$

For compactly supported wavelets ϕ, only finitely many h_k’s will be non-zero. Define the 2π-periodic trigonometric polynomial

$$m_o(\omega) = \frac{1}{\sqrt{2}}\sum_{k\in\mathcal{Z}} h_k e^{-ik\omega} \tag{10.17}$$

associated with $\{h_k\}$. The Fourier transforms of ϕ and ψ can be shown to be of the form

$$\hat{\phi}(\omega) = \frac{1}{\sqrt{2}} \prod_{j=1}^{\infty} m_0(2^{-j}\omega), \tag{10.18}$$

$$\hat{\psi}(\omega) = -e^{-i\omega/2}\overline{m_0(\frac{\omega}{2} + \pi)}\hat{\phi}(\frac{\omega}{2}). \tag{10.19}$$

Depending on the number of non-zero elements in the filter $\{h_k\}$, wavelets of different degree of smoothness emerge.

It is natural to wonder what is special about MRA. Smoothing techniques such as linear regression, splines, and Fourier series all try to represent a signal in terms of component functions. At the same time, wavelet-based MRA studies the detail signals or differences in the approximations made at adjacent resolution levels. This way, local changes can be picked up much more easily than with other smoothing techniques.

With this short introduction to wavelets, we return to the nonparametric regression problem in (10.8). Much of the following discussion closely follows Angers and Delampady (2001). We begin with a compactly supported wavelet function $\psi \in \mathcal{C}^s$, the set of real-valued functions with continuous derivatives up to order s. We note that then g has the wavelet decomposition

$$g(x) = \sum_{|k| \le K_0} \alpha_k \phi_k(x) + \sum_{j \ge 0} \sum_{|k| \le K_j} \beta_{jk} \psi_{j,k}(x), \tag{10.20}$$

with

$$\phi_k(x) = \phi(x - k), \text{ and}$$
$$\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k),$$

where K_j is such that $\phi_k(x)$ and $\psi_{j,k}(x)$ vanish on $\mathcal{T}$ whenever $|k| > K_j$, and ϕ is the scaling function ('father wavelet') corresponding with the 'mother wavelet' ψ. Such K_j's exist (and are finite) because the wavelet function that we have chosen has compact support. For any specified resolution level J, we have

$$\begin{aligned} g(x) &= \sum_{|k| \le K_0} \alpha_k \phi_k(x) + \sum_{j=0}^{J} \sum_{|k| \le K_j} \beta_{jk} \psi_{j,k}(x) + \sum_{j=J+1}^{\infty} \sum_{|k| \le K_j} \beta_{jk} \psi_{j,k}(x) \\ &= g_J(x) + R_J(x), \end{aligned} \tag{10.21}$$

where

$$\begin{aligned} g_J(x) &= \sum_{|k| \le K_0} \alpha_k \phi_k(x) + \sum_{j \ge 0}^{J} \sum_{|k| \le K_j} \beta_{jk} \psi_{j,k}(x), \text{ and} \\ R_J(x) &= \sum_{j=J+1}^{\infty} \sum_{|k| \le K_j} \beta_{jk} \psi_{j,k}(x). \end{aligned} \tag{10.22}$$

In the representation (10.22), we note that the ϕ functions appearing in the first part detect the global features of g, and subsequently the ψ functions in the second part check for local details.

To proceed further, many standard wavelet based procedures apply the 'discrete wavelet transform' to the data and work with the resulting wavelet coefficients (see Vidakovic (1999), Müller and Vidakovic (1999)). We, however, use the familiar hierarchical Bayesian approach to specify the prior model for g in (10.8). At the resolution level J, (10.8) can be expressed as

$$y_i = g_J(x_i) + \eta_i + \varepsilon_i, \tag{10.23}$$

where $\eta_i = R_J(x_i)$. Because the amount of information available in the likelihood function to estimate the infinitely many parameters β_{jk}, $j > J$, $|k| \le K_j$ (arising from the higher levels of resolution and appearing in η_i) is very limited, it is best to treat these η_i as nuisance parameters and eliminate them by integrating out with respect to the prior given in (10.24) while estimating g_J. Otherwise, one will need to elicit some very informative prior on these parameters, thus attracting prior robustness issues as well. One other important issue is how large J should be. Note that the number of unknown parameters in the model grows exponentially with J, so it cannot be large for practical reasons. Also, there is no need for large J because its purpose is to check for local details only.

10.2.2 Hierarchical Prior Structure and Posterior Computations

In the first-stage prior specification, α_k and β_{jl} are all assumed to be independent normal random variables with mean 0. A common prior variance of τ^2 is assigned for α_k, whereas to accommodate the decreasing effect of the 'detail' coefficients β_{jl}, their variance is assumed to be $2^{-2js}\tau^2$. Now a joint prior distribution on σ^2 and τ^2 completes the prior specification. Even though conditionally, given τ^2, α_k and β_{jl} are normally distributed, unconditionally they do have heavy tailed prior distributions possessing robustness properties.

Let us now introduce some notations to facilitate the derivation of posterior quantities. Let $\boldsymbol{\gamma} = (\boldsymbol{\alpha}', \boldsymbol{\beta}')'$, where $\boldsymbol{\alpha} = (\alpha_k)_{|k| \le K_0}$, and $\boldsymbol{\beta} = (\beta_{jk})_{0 \le j \le J, |k| \le K_j}$. Then the first stage prior specified above is

$$\boldsymbol{\gamma}|\tau^2 \sim N_{2K_0+1+M_\beta}(\mathbf{0}, \tau^2 \mathit{\Gamma}), \text{ where } \mathit{\Gamma} = \begin{pmatrix} I_{2K_0+1} & 0 \\ 0 & \mathit{\Delta}_{M_\beta} \end{pmatrix},$$

with $M_\beta = \sum_{j=0}^{J}(2K_j + 1)$ and the diagonal matrix $\mathit{\Delta}$ being the variance-covariance matrix of β. Also,

$$\boldsymbol{\eta} = (\eta_1, \dots \eta_n)'|\tau^2 \sim N_n(\mathbf{0}, \tau^2 Q_n), \tag{10.24}$$

where, to keep the covariance structure of η_i simple, we choose

$$(Q_n)_{ij} = \tau^2 2^{-2Js} \exp(-c|x_i - x_j|),$$

for some moderate value of c. Further, let $X = (\varPhi', S')$ with the ith row of $\varPhi'$ being $\{\phi_k(x_i)\}'_{|k|\le K_0}$ and the ith row of S' being $\{\psi_{jk}(x_i)\}'_{0\le j\le J, |k|\le K_j}$. Then, given $\boldsymbol{\gamma}$, σ^2 and τ^2, we have the following linear model for the observation vector $\mathbf{Y} = (y_1, \ldots, y_n)'$:

$$\mathbf{Y} = X\boldsymbol{\gamma} + \mathbf{u}, \tag{10.25}$$

where $\mathbf{u} = \boldsymbol{\eta} + \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \varSigma)$ with $\varSigma = \sigma^2 I_n + \tau^2 Q_n$. This follows from the fact that

$$\begin{aligned} \mathbf{Y}|\gamma, \boldsymbol{\eta}, \sigma^2, \tau^2 &\sim N_n(X\boldsymbol{\gamma} + \boldsymbol{\eta}, \sigma^2 I_n), \qquad (10.26) \\ \boldsymbol{\eta}|\tau^2 &\sim N_n(\mathbf{0}, \tau^2 Q_n). \end{aligned}$$

From (10.25) and using standard hierarchical Bayes techniques (*cf.* Lindley and Smith (1972)) and matrix identities (*cf.* Searle (1982)), it follows that

$$\begin{aligned} \mathbf{Y}|\sigma^2, \tau^2 &\sim N_n(\mathbf{0}, \sigma^2 I_n + \tau^2 \left(X\varGamma X' + Q_n\right)), \qquad (10.27) \\ \boldsymbol{\gamma}|\mathbf{Y}, \sigma^2, \tau^2 &\sim N(A\mathbf{Y}, B), \qquad (10.28) \end{aligned}$$

where

$$\begin{aligned} A &= \tau^2 \varGamma X' \left(\sigma^2 I_n + \tau^2 \left(X\varGamma X' + Q_n\right)\right)^{-1}, \\ B &= \tau^2 \varGamma - \tau^4 \varGamma X' \left(\sigma^2 I_n + \tau^2 \left(X\varGamma X' + Q_n\right)\right)^{-1} X\varGamma. \end{aligned}$$

To proceed to the second-stage calculations, some algebraic simplifications are needed (see Angers and Delampady (1992)). Spectral decomposition yields $X\varGamma X' + Q_n = HDH'$, where $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ is the matrix of eigenvalues and H is the orthogonal matrix of eigenvectors. Thus,

$$\begin{aligned} \sigma^2 I_n + \tau^2 \left(X\varGamma X' + Q_n\right) &= H\left(\sigma^2 I_n + \tau^2 D\right) H' \\ &= \tau^2 H \left(vI_n + D\right) H', \qquad (10.29) \end{aligned}$$

where $v = \sigma^2/\tau^2$. Using this spectral decomposition, the marginal density of $\mathbf{Y}$ given τ^2 and v can be written as

$$\begin{aligned} m(\mathbf{Y} \mid \tau^2, v) &= \frac{1}{(2\pi\tau^2)^{n/2}} \frac{1}{\det(vI_n + D)^{1/2}} \\ &\quad \times \exp\left\{-\frac{1}{2\tau^2}\mathbf{Y}' H (vI_n + D)^{-1} H' Y\right\} \\ &= \frac{1}{(2\pi\tau^2)^{n/2}} \frac{1}{\prod_{i=1}^n (v + d_i)^{1/2}} \exp\left\{-\frac{1}{2\tau^2} \sum_{i=1}^n \frac{t_i^2}{v + d_i}\right\}, \qquad (10.30) \end{aligned}$$

where $\mathbf{t} = (t_1, \ldots, t_n)' = H'\mathbf{Y}$.

To derive the wavelet smoother, all that we need to do now is to eliminate the hyper- and nuisance parameters from the first-stage posterior distribution by integrating out these variables with respect to the second-stage prior on them. This is what we will do now. Alternatively, one could employ an empirical Bayes approach and estimate σ^2 and τ^2 from equation (10.27) and replace σ^2 and τ^2 by their estimates in equation (10.28) to approximate $\widehat{\gamma}$. However, this will underestimate the variance of the wavelet estimator, $\widehat{\mathbf{Y}} = X\widehat{\boldsymbol{\gamma}}$. Suppose, then, $\pi_2(\tau^2, v)$ is the second stage prior. It is well known in the context of hierarchical Bayesian analysis (see Chapter 9, specially equation (9.7) and Berger, 1985a) that the sensitivity of the second and higher stage hyper-priors on the final Bayes estimator is somewhat limited. Therefore, for computational ease, we choose $\pi_2(\tau^2, v) = \pi_{22}(v)(\tau^2)^{-a}$ for some suitable choice of $a > 0$; π_{22} is the prior specified for v.

Once a and π_{22} are specified, using equation (10.28) along with (10.29) and taking the expectation with respect to τ^2, we have that

$$E\left(\boldsymbol{\gamma} \mid \mathbf{Y}\right) = \widehat{\boldsymbol{\gamma}} = \mathit{\Gamma} X' H E\left[(vI_n + D)^{-1} \mid \mathbf{Y}\right]\mathbf{t}, \tag{10.31}$$

where the expectation is taken with respect to $\pi_{22}(v \mid \mathbf{Y})$. Again using equations (10.28) and (10.29), the posterior covariance matrix of $\boldsymbol{\gamma}$ can be written as

$$\begin{aligned} Var(\boldsymbol{\gamma} \mid \mathbf{Y}) = {} & \frac{1}{n+2a} E\left[\sum_{i=1}^{n} \frac{t_i^2}{v+d_i} \mid \mathbf{Y}\right] \mathit{\Gamma} \\ & - \frac{1}{n+2a} \mathit{\Gamma} X' H E\left[\left(\sum_{i=1}^{n} \frac{t_i^2}{v+d_i}\right)(vI_n + D)^{-1} \mid \mathbf{Y}\right] H' X \mathit{\Gamma} \\ & + E\left[\widehat{\boldsymbol{\gamma}}(v)\widehat{\boldsymbol{\gamma}}(v)' \mid \mathbf{Y}\right], \end{aligned} \tag{10.32}$$

where $\widehat{\boldsymbol{\gamma}}(v) = \mathit{\Gamma} X' H (vI_n + D)^{-1}\mathbf{t}$.

To compute these expectations, one can use several techniques. Because they involve only single dimensional integrals, standard numerical integration methods will work quite well. Several versions of the standard Monte Carlo approach can be employed quite satisfactorily and efficiently also. An example illustrating the methodology follows.

Example 10.2. This is based on data provided by Prof. Abraham Verghese (F.R.E.S.) of the Indian Institute of Horticultural Research, Bangalore, India (personal communication), which have already been analyzed in Angers and Delampady (2001). The variable of interest y that we have chosen from the data set is the weekly average humidity level. The observations were made from June 1, 1995, to December 13, 1998. (For some reason, the observations were not recorded on the same day of the week every time.) We have chosen time (day of recording the observation) as the covariate x. (Any other available covariate can be used also because wavelet-based smoothing with respect to any arbitrary covariate (measured in some general way) can be handled with

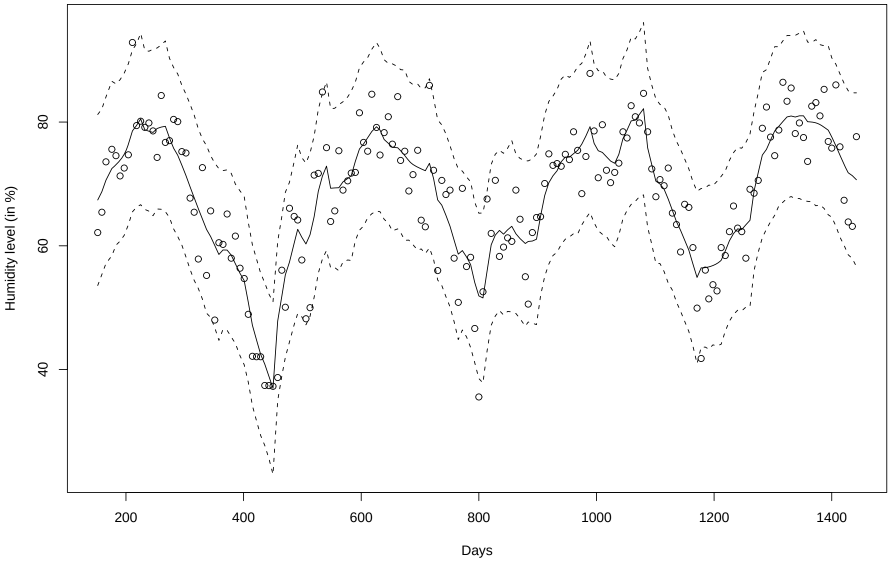

Fig. 10.1. Wavelet smoother and its error bands for the Humidity data.

our methodology.) For illustration purposes, we have chosen the model with $J = 6$; the hyperparameter a is 0.5 and the prior π_{22} corresponds with an F distribution with degrees of freedom 24 and 4. We have used compactly supported Daubechies wavelets for this analysis. As explained earlier, these cannot be expressed in closed form, but computations with these wavelets are possible using any of the several statistical and mathematical software packages. In Figure 10.1, we have plotted $\hat{g}_J$ (solid line) along with its error bands (dotted lines), $\pm 2\sqrt{Var(y \mid \mathbf{Y})}$, where

$$Var(y \mid \mathbf{Y}) = Var(g_J(x) + \eta + \varepsilon \mid \mathbf{Y}).$$

More details on this example as well as other studies can be found in Angers and Delampady (2001).

10.3 Estimation of Regression Function Using Dirichlet Multinomial Allocation

In Section 10.2, wavelets are used to represent the nonparametric regression function in (10.8) and a prior is put on the wavelet coefficients. Here we present an alternative approach based on the observation that the unknown regression function is locally linear and hence one may use a high-dimensional parametric family for modeling locally linear regression. Suppose we have a regression problem with a response variable Y and a regressor variable X. Let $(X_1, Y_1), \ldots (X_n, Y_n)$ be independent paired observations on (X, Y). Consider first the usual normal linear regression model where given values of the regressor variables x_i's, the Y_i's are independently normally distributed with common variance σ_Y^2 and mean $E(Y_i|x_i) = \beta_1 + \beta_2 x_i$, a linear function of x_i.

Let $Z_i = (X_i, Y_i)$ be independent, Z_i having the density

$$f(z|\phi_i) = f(x, y|\phi_i) = f_X(x|\mu_i, \sigma_i^2) f_Y(y|x, \beta_{1i}, \beta_{2i}, \sigma_Y^2)$$

where $f_X(x|\mu_i, \sigma_i^2)$ and $f_Y(y|x, \beta_{1i}, \beta_{2i}, \sigma_Y^2)$ denote respectively $N(\mu_i, \sigma_i^2)$ density for X_i and $N(\beta_{1i} + \beta_{2i}x, \sigma_Y^2)$ density for Y_i given x, $\phi_i = (\mu_i, \sigma_i^2, \beta_{1i}, \beta_{2i})$, $i = 1, \ldots, n$.

For simplicity we assume σ_Y^2 is known, say, equal to 1.

For the remaining parameters $\phi_i, i = 1, \ldots, n$, we have the Dirichlet multinomial allocation (DMA) prior, defined in the next paragraph.

(1) Let $k \sim p(k)$, a distribution on $\{1, 2, \ldots, n\}$.

(2) Given k, ϕ_i, $i = 1, \ldots, n$ have at most k distinct values $\theta_1, \ldots, \theta_k$, where θ_i's are i.i.d. $\sim G_0$ and G_0 is a distribution on the space of $(\mu, \sigma^2, \beta_1, \beta_2)$ (our choice of G_0 is mentioned below).

(3) Given k, the vector of weights $(w_1, \ldots, w_k) \sim$ Dirichlet $(\delta_1, \ldots, \delta_k)$.

(4) Allocation variables $a_1, \ldots, a_n$ are independent with

$$P(a_i = j) = w_j, \ j = 1, \ldots, k.$$

(5) Finally $\phi_i = \theta_{a_i}$, $i = 1, \ldots, n$.

For simplicity, we illustrate with a known k (which will be taken appropriately large). We refer to Richardson and Green (1997) for the treatment of the case with unknown k; see also the discussion of this paper by Gruet and Robert, and Green and Richardson (2001). Under this prior $\phi_i = (\mu_i, \sigma_i^2, \beta_{1i}, \beta_{2i})$, $i = 1, \ldots, n$ are exchangeable. This allows borrowing of strength, as in Chapter 9, from clusters of (x_i, y_i)'s with similar values. To see how this works, one has to calculate the Bayes estimate through MCMC.

We take G_0 to be the product of a normal distribution for μ, an inverse Gamma distribution for σ^2 and normal distributions for β_1, and β_2. The full conditionals needed for sampling from the posterior using Gibbs sampler can be easily obtained, see Robert and Casella (1999) in this context. For example, the conditional posterior distribution of $a_1, \ldots, a_n$ given other parameters are as follows:

$$a_i = j \text{ with probability } w_j f(Z_i|\theta_j) / \sum_{r=1}^{k} w_r f(Z_i|\theta_r).$$

$j = 1, \ldots, k$, $i = 1, \ldots, n$ and $a_1, \ldots, a_n$ are independent.

Due to conjugacy, the other full conditional distributions can be easily obtained. You are invited to calculate the conditional posteriors in Problem 4.

Note that given k, $\theta_1, \ldots, \theta_k$ and $w_1, \ldots, w_k$, we have a mixture with k components. Each mixture models a locally linear regression. Because θ_i and w_i are random, we have a rich family of locally linear regression models from which the posterior chooses different members and assigns to each member model a weight equal to its posterior probability density. The weight is a measure of how close is this member model to data. The Bayes estimate of the regression function is a weighted average of the (conditional) expectations of locally linear regressions.

We illustrate the use of this method with a set of data simulated form a model for which

$$E(Y|x) = \sin(2x) + \epsilon.$$

We generate 100 pairs of observations (X_i, Y_i) with normal errors ϵ_i. A scatter plot of the data points and a plot of the estimated regression at each X_i (using the Bayes estimates of β_{1i}, β_{2i}) together with the graph of $\sin(2x)$

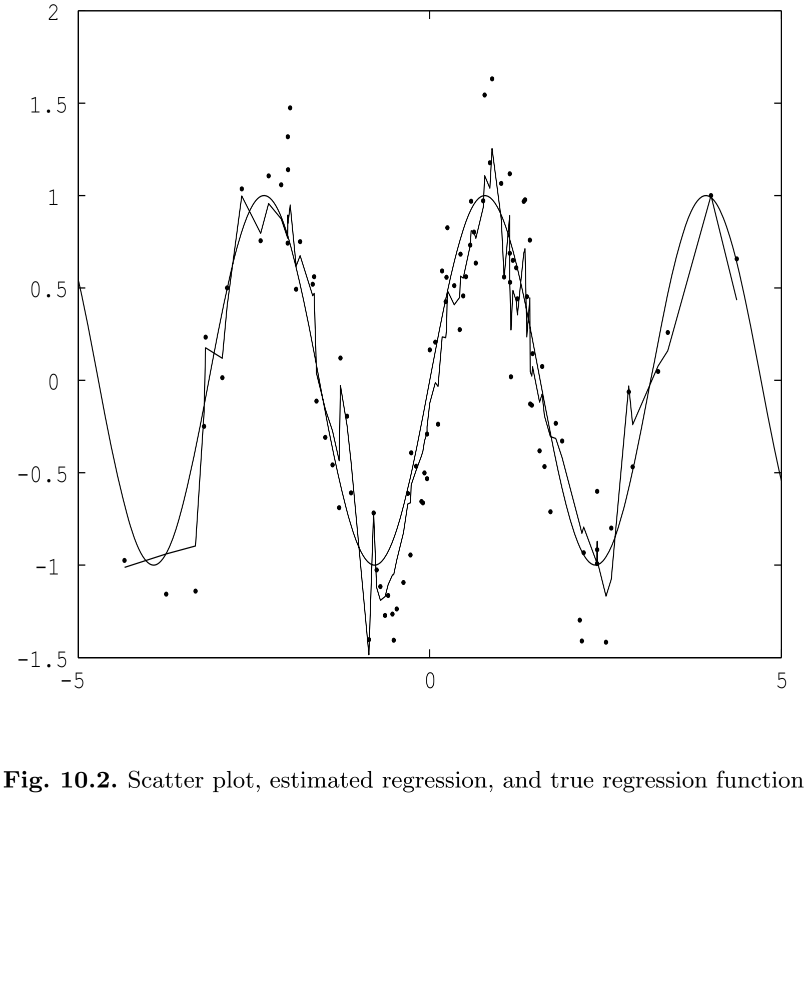

Fig. 10.2. Scatter plot, estimated regression, and true regression function.

are presented in Figure 10.2. In our calculation, we have chosen hyperparameters of the priors suitably to have priors with small information. Seo (2004) discusses the choice of hyperpriors and hyperparameters in examples of this kind.

Following Müller et al. (1996), Seo (2004) also uses a Dirichlet process prior instead of the DMA. The Dirichlet process prior is beyond the scope of our book. See Ghosh and Ramamoorthi (2003, Chapter 3) for details.

It is worth noting that the method developed works equally well if X is non-stochastic (as in Section 10.2) or has a known distribution. The trick is to ignore these facts and pretend that X is also random as above. See Müller et al. (1996) for further discussion of this point.

10.4 Exercises

1. Verify that Haar wavelets generate an MRA of $\mathcal{L}^2(\mathcal{R})$.
2. Indicate how Bayes factors can be used to obtain the optimal resolution level J in (10.21).
3. Derive an appropriate wavelet smoother for the data given in Table 5.1 and compare the results with those obtained using linear regression in Section 5.4.
4. For the problem in Section 10.3, explain how MCMC can be implemented, deriving explicitly all the full conditionals needed.
5. Choose any of the high-dimensional problems in Chapters 9 or 10 and suggest how hyperparameters may be chosen there. Discuss whether your findings will apply to all the higher levels of hierarchy.

A

Common Statistical Densities

For quick reference, listed below are some common statistical densities that are used in examples and exercise problems in the book. Only brief description including the name of the density, the notation (abbreviation) used in the book, the density itself, the range of the variable argument, and the parameter values and some useful moments are supplied.

A.1 Continuous Models

1. Univariate normal ($N(\mu, \sigma^2)$):
$$f(x|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left(-(x-\mu)^2/(2\sigma^2)\right),$$
$-\infty < x < \infty,\ -\infty < \mu < \infty, \sigma^2 > 0.$
Mean $= \mu$, variance $= \sigma^2$.
Special case: $N(0,1)$ is known as standard normal.
2. Multivariate normal ($N_p(\boldsymbol{\mu}, \Sigma)$):
$$f(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = (2\pi)^{-p/2}|\Sigma|^{-1/2} \exp\left(-(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right),$$
$\mathbf{x} \in \mathcal{R}^p, \boldsymbol{\mu} \in \mathcal{R}^p, \Sigma_{p\times p}$ positive definite.
Mean vector $= \boldsymbol{\mu}$, covariance or dispersion matrix $= \Sigma$.
3. Exponential ($Exp(\lambda)$):
$$f(x|\lambda) = \lambda \exp(-\lambda x), x > 0, \lambda > 0.$$
Mean $= 1/\lambda$, variance $= 1/\lambda^2$.
4. Double exponential or Laplace ($DExp(\mu, \sigma)$):
$$f(x|\mu, \sigma) = \frac{1}{2\sigma} \exp\left(-\frac{|x-\mu|}{\sigma}\right),$$
$-\infty < x < \infty, -\infty < \mu < \infty, \sigma > 0.$
Mean $= \mu$, variance $= 2\sigma^2$.

5. Gamma ($Gamma(\alpha, \lambda)$):

$$f(x|\alpha, \lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\lambda x), x > 0, \alpha > 0, \lambda > 0.$$

Mean $= \alpha/\lambda$, variance $= \alpha/\lambda^2$.
Special cases:
(i) $Exp(\lambda)$ is $Gamma(1, \lambda)$.
(ii) Chi-square with n degrees of freedom (χ^2_n) is $Gamma(n/2, 1/2)$.

6. Uniform ($U(a, b)$):

$$f(x|a, b) = \frac{1}{b-a} I_{(a,b)}(x), -\infty < a < b < \infty.$$

Mean $= (a + b)/2$, variance $= (b - a)^2/12$.

7. Beta ($Beta(\alpha, \beta)$):

$$f(x|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1} I_{(0,1)}(x), \alpha > 0, \beta > 0.$$

Mean $= \alpha/(\alpha + \beta)$, variance $= \alpha\beta/\{(\alpha + \beta)^2(\alpha + \beta + 1)\}$.
Special case: $U(0, 1)$ is $Beta(1, 1)$.

8. Cauchy ($Cauchy(\mu, \sigma^2)$):

$$f(x|\mu, \sigma^2) = \frac{1}{\pi\sigma} \left(1 + \frac{(x - \mu)^2}{\sigma^2}\right)^{-1}, -\infty < x < \infty,$$

$-\infty < \mu < \infty, \sigma^2 > 0$. Mean and variance do not exist.

9. t distribution ($t(\alpha, \mu, \sigma^2)$):

$$f(x|\alpha, \mu, \sigma^2) = \frac{\Gamma((\alpha + 1)/2)}{\sigma\sqrt{\alpha\pi}\,\Gamma(\alpha/2)} \left(1 + \frac{(x - \mu)^2}{\alpha\sigma^2}\right)^{-(\alpha+1)/2},$$

$-\infty < x < \infty, \alpha > 0, -\infty < \mu < \infty, \sigma^2 > 0$.
Mean $= \mu$ if $\alpha > 1$, variance $= \alpha\sigma^2/(\alpha - 2)$ if $\alpha > 2$.
Special cases:
(i) $Cauchy(\mu, \sigma^2)$ is $t(1, \mu, \sigma^2)$.
(ii) $t(k, 0, 1) = t_k$ is known as Student's t with k degrees of freedom.

10. Multivariate t ($t_p(\alpha, \boldsymbol{\mu}, \Sigma)$):

$$f(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{\Gamma((\alpha + p)/2))}{(\alpha\pi)^{p/2}\Gamma(\alpha/2)} |\Sigma|^{-1/2} \left(1 + \frac{1}{\alpha}(\mathbf{x} - \boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)^{-(\alpha+p)/2},$$

$\mathbf{x} \in \mathcal{R}^p, \alpha > 0, \boldsymbol{\mu} \in \mathcal{R}^p, \Sigma_{p\times p}$ positive definite.
Mean vector $= \boldsymbol{\mu}$ if $\alpha > 1$, covariance or dispersion matrix $=$ $\alpha\Sigma/(\alpha - 2)$ if $\alpha > 2$.

11. F distribution with degrees of freedom α and β ($F(\alpha,\beta)$):

$$f(x|\alpha,\beta) = \frac{\Gamma((\alpha+\beta)/2)}{\Gamma(\alpha/2)\Gamma(\beta/2)}\left(\frac{\alpha}{\beta}\right)^{\alpha/2}\frac{x^{\alpha/2-1}}{\left(1+\frac{\alpha}{\beta}x\right)^{(\alpha+\beta)/2}}, x>0, \alpha>0, \beta>0.$$

Mean $= \beta/(\beta-2)$ if $\beta>2$, variance $= 2\beta^2(\alpha+\beta-2)/\{\alpha(\beta-4)(\beta-2)^2\}$ if $\beta>4$.
Special cases:
(i) If $X \sim t(\alpha,\mu,\sigma^2)$, $(X-\mu)^2/\sigma^2 \sim F(1,\alpha)$.
(ii) If $\mathbf{X} \sim t_p(\alpha,\boldsymbol{\mu},\Sigma)$, $\frac{1}{p}(\mathbf{X}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}) \sim F(p,\alpha)$.

12. Inverse Gamma (*inverse Gamma*(α,λ)):

$$f(x|\alpha,\lambda) = \frac{\lambda^\alpha}{\Gamma(\alpha)}x^{-(\alpha+1)}\exp(-\lambda/x), x>0, \alpha>0, \lambda>0.$$

Mean $= \lambda/(\alpha-1)$ if $\alpha>1$, variance $= \lambda^2/\{(\alpha-1)^2(\alpha-2)\}$ if $\alpha>2$.
If $X \sim$ *inverse Gamma*(α,λ), $1/X \sim$ *Gamma*(α,λ).

13. Dirichlet (finite dimensional) ($D(\boldsymbol{\alpha})$):

$$f(\mathbf{x}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)}\prod_{i=1}^k x_i^{\alpha_i-1},$$

$\mathbf{x} = (x_1,\ldots,x_k)'$ with $0 \le x_i \le 1$, for $1 \le i \le k$, $\sum_{i=1}^k x_i = 1$ and $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_k)'$ with $\alpha_i > 0$ for $1 \le i \le k$.
Mean vector $= \boldsymbol{\alpha}/(\sum_{i=1}^k \alpha_i)$, covariance or dispersion matrix $= C_{k\times k}$ where

$$C_{ij} = \begin{cases} \dfrac{\alpha_i \sum_{l\neq i}\alpha_l}{\left(\sum_{l=1}^k \alpha_l\right)^2\left(\sum_{l=1}^k \alpha_l+1\right)} & \text{if } i=j; \\ -\dfrac{\alpha_i\alpha_j}{\left(\sum_{l=1}^k \alpha_l\right)^2\left(\sum_{l=1}^k \alpha_l+1\right)} & \text{if } i\neq j. \end{cases}$$

14. Wishart ($W_p(n,\Sigma)$):

$$f(A|\Sigma) = \frac{1}{2^{np/2}\Gamma_p(n/2)}|\Sigma|^{-n/2}\exp\left(-trace\{\Sigma^{-1}A\}/2\right)|A|^{(n-p-1)/2},$$

$A_{p\times p}$ positive definite, $\Sigma_{p\times p}$ positive definite, $n \ge p$, p positive integer,

$$\Gamma_p(a) = \int_{A \text{ positive definite}} \exp\left(-trace\{A\}\right)|A|^{a-(p+1)/2}\,dA,$$

for $a > (p-1)/2$.
Mean $= n\Sigma$. For other moments, see Muirhead (1982).
Special case: χ_n^2 is $W_1(n,1)$.
If $W^{-1} \sim W_p(n,\Sigma)$ then W is said to follow inverse-Wishart distribution.

15. Logistic ($(Logistic(\mu, \sigma)$):

$$f(x|\mu,\sigma) = \frac{1}{\sigma} \frac{\exp(-\frac{x-\mu}{\sigma})}{\left(1+\exp(-\frac{x-\mu}{\sigma})\right)^2},$$

$-\infty < x < \infty$, $-\infty < \mu < \infty$, $\sigma > 0$.
Mean $= \mu$, variance $= \pi^2\sigma^2/3$.

A.2 Discrete Models

1. Binomial ($B(n, p)$):

$$f(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x},$$

$x = 0, 1, \ldots, n$, $0 \le p \le 1$, $n \ge 1$ integer.
Mean $= np$, variance $= np(1-p)$.
Special case: $Bernoulli(p)$ is $B(1, p)$.

2. Poisson ($\mathcal{P}(\lambda)$):

$$f(x|n,p) = \frac{\exp(-\lambda)\lambda^x}{x!},$$

$x = 0, 1, \ldots$, $\lambda > 0$.
Mean $= \lambda$, variance $= \lambda$.

3. Geometric ($Geometric(p)$):

$$f(x|p) = (1-p)^x p,$$

$x = 0, 1, \ldots$, $0 < p \le 1$.
Mean $= (1-p)/p$, variance $= (1-p)/p^2$.

4. Negative binomial ($Negative\ binomial(k, p)$):

$$f(x|k,p) = \binom{x+k-1}{k-1} (1-p)^x p^k,$$

$x = 0, 1, \ldots$, $0 < p \le 1$, $k \ge 1$ integer.
Mean $= k(1-p)/p$, variance $= k(1-p)/p^2$.
Special case: $Geometric(p)$ is $Negative\ binomial(1, p)$.

5. Multinomial ($Multinomial(n, \mathbf{p})$):

$$f(\mathbf{x}|n,\mathbf{p}) = \frac{n!}{\prod_{i=1}^k x_i!} \prod_{i=1}^k p_i^{x_i},$$

$\mathbf{x} = (x_1, \ldots, x_k)'$ with x_i an integer between 0 and n, for $1 \le i \le k$, $\sum_{i=1}^k x_i = n$ and $\mathbf{p} = (p_1, \ldots, p_k)'$ with $0 \le p_i \le 1$ for $1 \le i \le k$, $\sum_{i=1}^k p_i = 1$.
Mean vector $= n\mathbf{p}$, covariance or dispersion matrix $= C_{k\times k}$ where

$$C_{ij} = \begin{cases} np_i(1-p_i) & \text{if } i = j; \\ -np_ip_j & \text{if } i \ne j. \end{cases}$$

B

Birnbaum's Theorem on Likelihood Principle

The object of this appendix is to rewrite the usual proof of Birnbaum's theorem (e.g., as given in Basu (1988)) using only mathematical statements and carefully defining all symbols and the domain of discourse.

Let $\theta \in \Theta$ be the parameter of interest. A statistical experiment $\mathcal{E}$ is performed to generate a sample x. An experiment $\mathcal{E}$ is given by the triplet $(\mathcal{X}, \mathcal{A}, p)$ where $\mathcal{X}$ is the sample space, $\mathcal{A}$ is the class of all subsets of $\mathcal{X}$, and $p = \{p(\cdot|\theta), \theta \in \Theta\}$ is a family of probability functions on $(\mathcal{X}, \mathcal{A})$, indexed by the parameter space Θ. Below we consider experiments with a fixed parameter space Θ.

A (finite) mixture of experiments $\mathcal{E}_1, \ldots, \mathcal{E}_k$ with mixture probabilities $\pi_1, \ldots, \pi_k$ (non-negative numbers free of θ, summing to unity), which may be written as $\sum_{i=1}^{k} \pi_i \mathcal{E}_i$, is defined as a two stage experiment where one first selects $\mathcal{E}_i$ with probability π_i and then observes $x_i \in \mathcal{X}_i$ by performing the experiment $\mathcal{E}_i$.

Consider now a class of experiments closed under the formation of (finite) mixtures. Let $\mathcal{E} = (\mathcal{X}, \mathcal{A}, p)$ and $\mathcal{E}' = (\mathcal{X}', \mathcal{A}', p')$ be two experiments and $x \in \mathcal{X}, x' \in \mathcal{X}'$. By equivalence of the two points $(\mathcal{E}, x)$ and $(\mathcal{E}', x')$, we mean one makes the same inference on θ if one performs $\mathcal{E}$ and observes x or performs $\mathcal{E}'$ and observes x', and we denote this as

$$(\mathcal{E}, x) \sim (\mathcal{E}', x').$$

We now consider the following principles.

The likelihood principle (LP): We say that the equivalence relation "$\sim$" obeys the likelihood principle if $(\mathcal{E}, x) \sim (\mathcal{E}', x')$ whenever

$$p(x|\theta) = c\, p'(x'|\theta) \text{ for all } \theta \in \Theta \tag{B.1}$$

for some constant $c > 0$.

The weak conditionality principle (WCP): An equivalence relation "$\sim$" satisfies WCP if for a mixture of experiments $\mathcal{E} = \sum_{i=1}^{k} \pi_i \mathcal{E}_i$,

$$(\mathcal{E}, (i, x_i)) \sim (\mathcal{E}_i, x_i)$$

for any $i \in \{1, \dots, k\}$ and $x_i \in \mathcal{X}_i$.

The sufficiency principle (SP): An equivalence relation "$\sim$" satisfies SP if $(\mathcal{E}, x) \sim (\mathcal{E}, x')$ whenever $S(x) = S(x')$ for some sufficient statistic S for θ (or equivalently, $S(x) = S(x')$ for a minimal sufficient statistic S).

It is shown in Basu and Ghosh (1967) (see also Basu (1969)) that for discrete models a minimal sufficient statistic exists and is given by the likelihood partition, i.e., the partition induced by the equivalence relation (B.1) for two points x, x' from the same experiment. The difference between the likelihood principle and sufficiency principle is that in the former, x, x' may belong to possibly different experiments while in the sufficiency principle they belong to the same experiment.

The weak sufficiency principle (WSP): An equivalence relation "$\sim$" satisfies WSP if $(\mathcal{E}, x) \sim (\mathcal{E}, x')$ whenever $p(x|\theta) = p(x'|\theta)$ for all θ.

If follows that SP implies WSP, which can be seen by noting that

$$S(x) = \left\{ \frac{p(x|\theta)}{\sum_{\theta' \in \Theta} p(x|\theta')}, \ \theta \in \Theta \right\}$$

is a (minimal) sufficient statistic. We assume without loss of generality that

$$\sum_{\theta \in \Theta} p(x|\theta) > 0 \text{ for all } x \in \mathcal{X}.$$

We now state and prove Birnbaum's theorem on likelihood principle (Birnbaum (1962)).

Theorem B.1. *WCP and WSP together imply LP, i.e., if an equivalence relation satisfies WCP and WSP then it also satisfies LP.*

Proof. Suppose an equivalence relation "$\sim$" satisfies WCP and WSP. Consider two experiments $\mathcal{E}_1 = (\mathcal{X}_1, \mathcal{A}_1, p_1)$ and $\mathcal{E}_2 = (\mathcal{X}_2, \mathcal{A}_2, p_2)$ with same Θ and samples $x_i \in \mathcal{X}_i$, $i = 1, 2$, such that

$$p_1(x_1|\theta) = c p_2(x_2|\theta) \text{ for all } \theta \in \Theta \tag{B.2}$$

for some $c > 0$.

We are to show that $(\mathcal{E}_1, x_1) \sim (\mathcal{E}_2, x_2)$. Consider the mixture experiment $\mathcal{E}$ of $\mathcal{E}_1$ and $\mathcal{E}_2$ with mixture probabilities $1/(1+c)$ and $c/(1+c)$ respectively, i.e.,

$$\mathcal{E} = \frac{1}{1+c}\mathcal{E}_1 + \frac{c}{1+c}\mathcal{E}_2.$$

The points $(1, x_1)$ and $(2, x_2)$ in the sample space of $\mathcal{E}$ have probabilities $p_1(x_1|\theta)/(1+c)$ and $p_2(x_2|\theta)c/(1+c)$, respectively, which are the same by (B.2). WSP then implies that

$$(\mathcal{E}, (1, x_1)) \sim (\mathcal{E}, (2, x_2)). \tag{B.3}$$

Also, by WCP

$$(\mathcal{E}, (1, x_1)) \sim (\mathcal{E}_1, x_1) \text{ and } (\mathcal{E}, (2, x_2)) \sim (\mathcal{E}_2, x_2). \tag{B.4}$$

From (B.3) and (B.4), we have $(\mathcal{E}_1, x_1) \sim (\mathcal{E}_2, x_2)$. $\square$