CHAPTER 5

Hierarchical models

Many statistical applications involve multiple parameters that can be regarded as related or connected in some way by the structure of the problem, implying that a joint probability model for these parameters should reflect the dependence among them. For example, in a study of the effectiveness of cardiac treatments, with the patients in hospital j having survival probability θ_j, it might be reasonable to expect that estimates of the θ_j's, which represent a sample of hospitals, should be related to each other. We shall see that this is achieved in a natural way if we use a prior distribution in which the θ_j's are viewed as a sample from a common *population distribution*. A key feature of such applications is that the observed data, y_{ij}, with units indexed by i within groups indexed by j, can be used to estimate aspects of the population distribution of the θ_j's even though the values of θ_j are not themselves observed. It is natural to model such a problem hierarchically, with observable outcomes modeled conditionally on certain parameters, which themselves are given a probabilistic specification in terms of further parameters, known as *hyperparameters*. Such hierarchical thinking helps in understanding multiparameter problems and also plays an important role in developing computational strategies.

Perhaps even more important in practice is that nonhierarchical models are usually inappropriate for hierarchical data: with few parameters, they generally cannot fit large datasets accurately, whereas with many parameters, they tend to 'overfit' such data in the sense of producing models that fit the existing data well but lead to inferior predictions for new data. In contrast, hierarchical models can have enough parameters to fit the data well, while using a population distribution to structure some dependence into the parameters, thereby avoiding problems of overfitting. As we show in the examples in this chapter, it is often sensible to fit hierarchical models with more parameters than there are data points.

In Section 5.1, we consider the problem of constructing a prior distribution using hierarchical principles but without fitting a formal probability model for the hierarchical structure. We first consider the analysis of a single experiment, using historical data to create a prior distribution, and then we consider a plausible prior distribution for the parameters of a set of experiments. The treatment in Section 5.1 is not fully Bayesian, because, for the purpose of simplicity in exposition, we work with a point estimate, rather than a complete joint posterior distribution, for the parameters of the population distribution (the hyperparameters). In Section 5.2, we discuss how to construct a hierar-

chical prior distribution in the context of a fully Bayesian analysis. Sections 5.3–5.4 present a general approach to computation with hierarchical models in conjugate families by combining analytical and numerical methods. We defer details of the most general computational methods to Part III in order to explore immediately the important practical and conceptual advantages of hierarchical Bayesian models. The chapter concludes with two extended examples: a hierarchical model for an educational testing experiment and a Bayesian treatment of the method of 'meta-analysis' as used in medical research to combine the results of separate studies relating to the same research question.

5.1 Constructing a parameterized prior distribution

Analyzing a single experiment in the context of historical data

To begin our description of hierarchical models, we consider the problem of estimating a parameter θ using data from a small experiment and a prior distribution constructed from similar previous (or historical) experiments. Mathematically, we will consider the current and historical experiments to be a random sample from a common population.

Example. Estimating the risk of tumor in a group of rats
In the evaluation of drugs for possible clinical application, studies are routinely performed on rodents. For a particular study drawn from the statistical literature, suppose the immediate aim is to estimate θ, the probability of tumor in a population of female laboratory rats of type 'F344' that receive a zero dose of the drug (a control group). The data show that 4 out of 14 rats developed endometrial stromal polyps (a kind of tumor). It is natural to assume a binomial model for the number of tumors, given θ. For convenience, we select a prior distribution for θ from the conjugate family, $\theta \sim \text{Beta}(\alpha, \beta)$.

Analysis with a fixed prior distribution. From historical data, suppose we knew that the tumor probabilities θ among groups of female lab rats of type F344 follow an approximate beta distribution, with known mean and standard deviation. The tumor probabilities θ vary because of differences in rats and experimental conditions among the experiments. Referring to the expressions for the mean and variance of the beta distribution (see Appendix A), we could find values for α, β that correspond to the given values for the mean and standard deviation. Then, assuming a $\text{Beta}(\alpha, \beta)$ prior distribution for θ yields a $\text{Beta}(\alpha + 4, \beta + 10)$ posterior distribution for θ.

Approximate estimate of the population distribution using the historical data. Typically, the mean and standard deviation of underlying tumor risks are not available. Rather, historical *data* are available on previous experiments on similar groups of rats. In the rat tumor example, the historical data were in fact a set of observations of tumor incidence in 70 groups of rats (Table 5.1). In the jth historical experiment, let the number of rats with tumors be y_j and the total number of rats be n_j. We model the y_j's as independent binomial data, given sample sizes n_j and study-specific means θ_j. Assuming that the beta prior distribution with parameters (α, β) is a good description of the population distribution

fully Bayesian analysis. Sections
utation with hierarchical models
l and numerical methods. We de-
al methods to Part III in order
tical and conceptual advantages
er concludes with two extended
itional testing experiment and a
-analysis' as used in medical re-
lies relating to the same research

r distribution

of historical data

lels, we consider the problem of
a small experiment and a prior
ous (or historical) experiments.
it and historical experiments to
tion.

in a group of rats

l application, studies are routinely
y drawn from the statistical litera-
ite θ, the probability of tumor in a
F344' that receive a zero dose of the
1 out of 14 rats developed endome-
atural to assume a binomial model
lience, we select a prior distribution
3).

From historical data, suppose we
groups of female lab rats of type
n, with known mean and standard
use of differences in rats and exper-
Referring to the expressions for the
e Appendix A), we could find values
r the mean and standard deviation.
n for θ yields a Beta$(\alpha+4, \beta+10)$

distribution using the historical

eviation of underlying tumor risks
available on previous experiments
xample, the historical data were in
in 70 groups of rats (Table 5.1). In
of rats with tumors be y_j and the
as independent binomial data, given
Assuming that the beta prior distri-
ption of the population distribution

Previous experiments:

0/20	0/20	0/20	0/20	0/20	0/20	0/20	0/19	0/19	0/19
0/19	0/18	0/18	0/17	1/20	1/20	1/20	1/20	1/19	1/19
1/18	1/18	2/25	2/24	2/23	2/20	2/20	2/20	2/20	2/20
2/20	1/10	5/49	2/19	5/46	3/27	2/17	7/49	7/47	3/20
3/20	2/13	9/48	10/50	4/20	4/20	4/20	4/20	4/20	4/20
4/20	10/48	4/19	4/19	4/19	5/22	11/46	12/49	5/20	5/20
6/23	5/19	6/22	6/20	6/20	6/20	16/52	15/47	15/46	9/24

Current experiment:

4/14

Table 5.1 *Tumor incidence in historical control groups and current group of rats, from Tarone (1982). The table displays the values of y_j/n_j: (number of rats with tumors)/(total number of rats).*

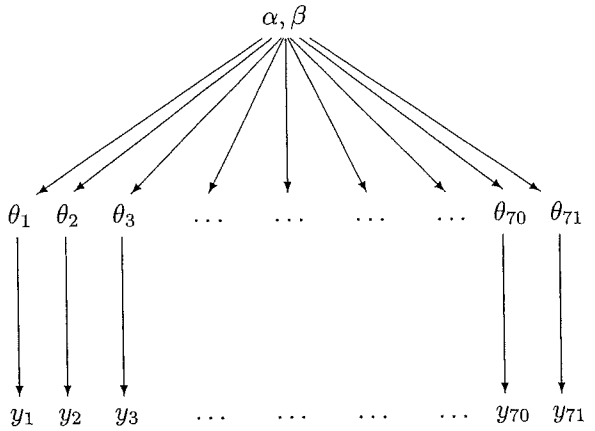

Figure 5.1 *Structure of the hierarchical model for the rat tumor example.*

of the θ_j's in the historical experiments, we can display the hierarchical model schematically as in Figure 5.1, with θ_{71} and y_{71} corresponding to the current experiment.

The observed sample mean and standard deviation of the 70 values y_j/n_j are 0.136 and 0.103. If we set the mean and standard deviation of the population distribution to these values, we can solve for α and β—see (A.3) on page 582 in Appendix A. The resulting estimate for (α, β) is $(1.4, 8.6)$. This is *not* a Bayesian calculation because it is not based on any specified full probability model. We present a better, fully Bayesian approach to estimating (α, β) for this example in Section 5.3. The estimate $(1.4, 8.6)$ is simply a starting point from which we can explore the idea of estimating the parameters of the population distribution.

Using the simple estimate of the historical population distribution as a prior distribution for the current experiment yields a Beta$(5.4, 18.6)$ posterior distribution for θ_{71}: the posterior mean is 0.223, and the standard deviation is 0.083. The prior

information has resulted in a posterior mean substantially lower than the crude proportion, $4/14 = 0.286$, because the weight of experience indicates that the number of tumors in the current experiment is unusually high.

These analyses require that the current tumor risk, θ_{71}, and the 70 historical tumor risks, $\theta_1, \ldots, \theta_{70}$, be considered a random sample from a common distribution, an assumption that would be invalidated, for example, if it were known that the historical experiments were all done in laboratory A but the current data were gathered in laboratory B, or if time trends were relevant. In practice, a simple, although arbitrary, way of accounting for differences between the current and historical data is to inflate the historical variance. For the beta model, inflating the historical variance means decreasing $(\alpha+\beta)$ while holding α/β constant. Other systematic differences, such as a time trend in tumor risks, can be incorporated in a more extensive model.

Having used the 70 historical experiments to form a prior distribution for θ_{71}, we might now like also to use this same prior distribution to obtain Bayesian inferences for the tumor probabilities in the first 70 experiments, $\theta_1, \ldots, \theta_{70}$. There are several logical and practical problems with the approach of directly estimating a prior distribution from existing data:

- If we wanted to use the estimated prior distribution for inference about the first 70 experiments, then the data would be used twice: first, all the results together are used to estimate the prior distribution, and then each experiment's results are used to estimate its θ. This would seem to cause us to overestimate our precision.
- The point estimate for α and β seems arbitrary, and using any point estimate for α and β necessarily ignores some posterior uncertainty.
- We can also make the opposite point: does it make sense to 'estimate' α and β at all? They are part of the 'prior' distribution: should they be known before the data are gathered, according to the logic of Bayesian inference?

Logic of combining information

Despite these problems, it clearly makes more sense to try to estimate the population distribution from all the data, and thereby to help estimate each θ_j, than to estimate all 71 values θ_j separately. Consider the following thought experiment about inference on two of the parameters, θ_{26} and θ_{27}, each corresponding to experiments with 2 observed tumors out of 20 rats. Suppose our prior distribution for both θ_{26} and θ_{27} is centered around 0.15; now suppose that you were told after completing the data analysis that $\theta_{26} = 0.1$ exactly. This should influence your estimate of θ_{27}; in fact, it would probably make you think that θ_{27} is lower than you previously believed, since the data for the two parameters are identical, and the postulated value of 0.1 is lower than you previously expected for θ_{26} from the prior distribution. Thus, θ_{26} and θ_{27} should be dependent in the posterior distribution, and they should not be analyzed separately.

We retain the advantages of using the data to estimate prior parameters

and eliminate all of the disadvantages just mentioned by putting a probability model on the entire set of parameters and experiments and then performing a Bayesian analysis on the joint distribution of all the model parameters. A complete Bayesian analysis is described in Section 5.3. The analysis using the data to estimate the prior parameters, which is sometimes called *empirical Bayes*, can be viewed as an approximation to the complete hierarchical Bayesian analysis. We prefer to avoid the term 'empirical Bayes' because it misleadingly suggests that the full Bayesian method, which we discuss here and use for the rest of the book, is not 'empirical.'

5.2 Exchangeability and setting up hierarchical models

Generalizing from the example of the previous section, consider a set of experiments $j = 1, \ldots, J$, in which experiment j has data (vector) y_j and parameter (vector) θ_j, with likelihood $p(y_j|\theta_j)$. (Throughout this chapter we use the word 'experiment' for convenience, but the methods can apply equally well to nonexperimental data.) Some of the parameters in different experiments may overlap; for example, each data vector y_j may be a sample of observations from a normal distribution with mean μ_j and common variance σ^2, in which case $\theta_j = (\mu_j, \sigma^2)$. In order to create a joint probability model for all the parameters θ, we use the crucial idea of exchangeability introduced in Chapter 1 and used repeatedly since then.

Exchangeability

If no information—other than the data y—is available to distinguish any of the θ_j's from any of the others, and no ordering or grouping of the parameters can be made, one must assume symmetry among the parameters in their prior distribution. This symmetry is represented probabilistically by exchangeability; the parameters $(\theta_1, \ldots, \theta_J)$ are *exchangeable* in their joint distribution if $p(\theta_1, \ldots, \theta_J)$ is invariant to permutations of the indexes $(1, \ldots, J)$. For example, in the rat tumor problem, suppose we have no information to distinguish the 71 experiments, other than the sample sizes n_j, which presumably are not related to the values of θ_j; we therefore use an exchangeable model for the θ_j's.

We have already encountered the concept of exchangeability in constructing iid models for unit- or individual-level data. In practice, ignorance implies exchangeability. Generally, the less we know about a problem, the more confidently we can make claims of exchangeability. (This is not, we hasten to add, a good reason to limit our knowledge of a problem before embarking on statistical analysis!) Consider the analogy to a roll of a die: we should initially assign equal probabilities to all six outcomes, but if we study the measurements of the die and weigh the die carefully, we might eventually notice imperfections, which might make us favor one outcome over the others and thus eliminate the symmetry among the six outcomes.

The simplest form of an exchangeable distribution has each of the parameters θ_j as an independent sample from a prior (or population) distribution governed by some unknown parameter vector ϕ; thus,

$$p(\theta|\phi) = \prod_{j=1}^{J} p(\theta_j|\phi). \tag{5.1}$$

In general, ϕ is unknown, so our distribution for θ must average over our uncertainty in ϕ:

$$p(\theta) = \int \left[\prod_{j=1}^{J} p(\theta_j|\phi)\right] p(\phi)d\phi, \tag{5.2}$$

This form, the mixture of iid distributions, is usually all that we need to capture exchangeability in practice.

A related theoretical result, *de Finetti's theorem*, to which we alluded in Section 1.2, states that in the limit as $J \to \infty$, any suitably well-behaved exchangeable distribution on $(\theta_1, \ldots, \theta_J)$ can be written in the iid mixture form (5.2). Formally, de Finetti's theorem does not hold when J is finite (see Exercise 5.2). Statistically, the iid mixture model characterizes parameters θ as drawn from a common 'superpopulation' that is determined by the unknown hyperparameters, ϕ. We are already familiar with exchangeable models for *data*, $y_1, \ldots, y_n$, in the form of 'iid' likelihoods, in which the n observations are independent and identically distributed, given some parameter vector θ.

Example. Exchangeability and sampling

The following thought experiment illustrates the role of exchangeability in inference from random sampling. For simplicity, we use a nonhierarchical example with exchangeability at the level of y rather than θ.

We, the authors, have selected eight states out of the United States and recorded the divorce rate per 1000 population in each state in 1981. Call these $y_1, \ldots, y_8$. What can you, the reader, say about y_8, the divorce rate in the eighth state?

Since you have no information to distinguish any of the eight states from the others, you must model them exchangeably. You might use a beta distribution for the eight y_j's, a logit normal, or some other prior distribution restricted to the range $[0, 1]$. Unless you are familiar with divorce statistics in the United States, your distribution on $(y_1, \ldots, y_8)$ should be fairly vague.

We now randomly sample seven states from these eight and tell you their divorce rates: 5.8, 6.6, 7.8, 5.6, 7.0, 7.1, 5.4, each in numbers of divorces per 1000 population (per year). Based primarily on the data, a reasonable posterior (predictive) distribution for the remaining value, y_8, would probably be centered around 6.5 and have most of its mass between 5.0 and 8.0.

Suppose initially we had given you the further prior information that the eight states are Mountain states: Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, and Wyoming, but selected in a random order; you still are not told which observed rate corresponds to which state. Now, before the seven data points were observed, the eight divorce rates should still be modeled exchangeably. However, your prior distribution (that is, *before* seeing the data), for the

bution has each of the param-
or (or population) distribution
ϕ; thus,

$b)$. (5.1)

ı for θ must average over our

$p(\phi)d\phi$, (5.2)

s usually all that we need to

orem, to which we alluded in
∞, any suitably well-behaved
be written in the iid mixture
not hold when J is finite (see
l characterizes parameters θ as
is determined by the unknown
with exchangeable models for
, in which the n observations
ven some parameter vector θ.

he role of exchangeability in in-
ve use a nonhierarchical example
ın θ.

f the United States and recorded
ıte in 1981. Call these $y_1, \ldots, y_8$.
orce rate in the eighth state?

ıny of the eight states from the
ou might use a beta distribution
rior distribution restricted to the
e statistics in the United States,
vague.

ie eight and tell you their divorce
ers of divorces per 1000 popula-
reasonable posterior (predictive)
probably be centered around 6.5

prior information that the eight
Idaho, Montana, Nevada, New
random order; you still are not
tate. Now, before the seven data
ould still be modeled exchange-
before seeing the data), for the

eight numbers should change: it seems reasonable to assume that Utah, with its large Mormon population, has a much lower divorce rate, and Nevada, with its liberal divorce laws, has a much higher divorce rate, than the remaining six states. Perhaps, given your expectation of outliers in the distribution, your prior distribution should have wide tails. Given this extra information (the names of the eight states), when you see the seven observed values and note that the numbers are so close together, it might seem a reasonable guess that the missing eighth state is Nevada or Utah. Therefore its value might be expected to be much lower or much higher than the seven values observed. This might lead to a bimodal or trimodal posterior distribution to account for the two plausible scenarios. The prior distribution on the eight values y_j is still exchangeable, however, because you have no information telling which state corresponds to which index number. (See Exercise 5.4.)

Finally, we tell you that the state not sampled (corresponding to y_8) was Nevada. Now, even before seeing the seven observed values, you cannot assign an exchangeable prior distribution to the set of eight divorce rates, since you have information that distinguishes y_8 from the other seven numbers, here suspecting it is larger than any of the others. Once $y_1, \ldots, y_7$ have been observed, a reasonable posterior distribution for y_8 plausibly should have most of its mass above the largest observed rate.

Incidentally, Nevada's divorce rate in 1981 was 13.9 per 1000 population.

Exchangeability when additional information is available on the units

In the previous example, if we knew x_j, the divorce rate in state j *last* year, for $j = 1, \ldots, 8$, but not which index corresponded to which state, then we would certainly be able to distinguish the eight values of y_j, but the joint prior distribution $p(x_j, y_j)$ would be the same for each state. In general, the usual way to model exchangeability with covariates is through conditional independence: $p(\theta_1, \ldots, \theta_J | x_1, \ldots, x_J) = \int [\prod_{j=1}^{J} p(\theta_j | \phi, x_j)] p(\phi | x) d\phi$, with $x = (x_1, \ldots, x_J)$.

In this way, exchangeable models become almost universally applicable, because any information available to distinguish different units should be encoded in the x and y variables. For example, consider the probabilities of a given die landing on each of its six faces, *after* we have carefully measured the die and noted its physical imperfections. If we include the imperfections (such as the area of each face, the bevels of the corners, and so forth) as explanatory variables x in a realistic physical model, the probabilities $\theta_1, \ldots, \theta_6$ should become exchangeable, conditional on x. In this example, the six parameters θ_j are constrained to sum to 1 and so *cannot* be modeled with a mixture of iid distributions; nonetheless, they can be modeled exchangeably.

In the rat tumor example, we have already noted that the sample sizes n_j are the only available information to distinguish the different experiments. It does not seem likely that n_j would be a useful variable for modeling tumor rates, but if one were interested, one could create an exchangeable model for the J pairs $(n, y)_j$. A natural first step would be to plot y_j/n_j vs. n_j to see any obvious relation that could be modeled. For example, perhaps some studies

j had larger sample sizes n_j because the investigators correctly suspected rarer events; that is, smaller θ_j and thus smaller expected values of y_j/n_j. In fact, the plot of y_j/n_j versus n_j, not shown here, shows no apparent relation between the two variables.

Objections to exchangeable models

In virtually any statistical application, it is natural to object to exchangeability on the grounds that the units actually differ. For example, the 71 rat tumor experiments were performed at different times, on different rats, and presumably in different laboratories. Such information does *not*, however, invalidate exchangeability. That the experiments differ implies that the θ_j's differ, but it might be perfectly acceptable to consider them as if drawn from a common distribution. In fact, with no information available to distinguish them, we have no logical choice but to model the θ_j's exchangeably. Objecting to exchangeability for modeling ignorance is no more reasonable than objecting to an iid model for samples from a common population, objecting to regression models in general, or, for that matter, objecting to displaying points in a scatterplot without individual labels. As with regression, the valid concern is not about exchangeability, but about encoding relevant knowledge as explanatory variables where possible.

The full Bayesian treatment of the hierarchical model

Returning to the problem of inference, the key 'hierarchical' part of these models is that ϕ is not known and thus has its own prior distribution, $p(\phi)$. The appropriate Bayesian posterior distribution is of the vector (ϕ, θ). The joint prior distribution is

$$p(\phi, \theta) = p(\phi)p(\theta|\phi),$$

and the joint posterior distribution is

$$\begin{aligned} p(\phi, \theta|y) &\propto p(\phi, \theta)p(y|\phi, \theta) \\ &= p(\phi, \theta)p(y|\theta), \end{aligned} \tag{5.3}$$

with the latter simplification holding because the data distribution, $p(y|\phi, \theta)$, depends only on θ; the hyperparameters ϕ affect y only through θ. Previously, we assumed ϕ was known, which is unrealistic; now we include the uncertainty in ϕ in the model.

The hyperprior distribution

In order to create a joint probability distribution for (ϕ, θ), we must assign a prior distribution to ϕ. If little is known about ϕ, we can assign a diffuse prior distribution, but we must be careful when using an improper prior density to check that the resulting posterior distribution is proper, and we should

e investigators correctly suspected
maller expected values of y_j/n_j. In
n here, shows no apparent relation

s natural to object to exchangeabil-
liffer. For example, the 71 rat tumor
imes, on different rats, and presum-
ıation does *not*, however, invalidate
ffer implies that the θ_j's differ, but
er them as if drawn from a common
n available to distinguish them, we
θ_j's exchangeably. Objecting to ex-
o more reasonable than objecting to
population, objecting to regression
ecting to displaying points in a scat-
ı regression, the valid concern is not
g relevant knowledge as explanatory

chical model

the key 'hierarchical' part of these
has its own prior distribution, $p(\phi)$.
ribution is of the vector (ϕ, θ). The

$)p(\theta|\phi)$,

$\phi, \theta)p(y|\phi, \theta)$
$\phi, \theta)p(y|\theta)$, (5.3)

cause the data distribution, $p(y|\phi, \theta)$,
ϕ affect y only through θ. Previously,
listic; now we include the uncertainty

tribution for (ϕ, θ), we must assign a
about ϕ, we can assign a diffuse prior
hen using an improper prior density
istribution is proper, and we should

assess whether our conclusions are sensitive to this simplifying assumption. In most real problems, one should have enough substantive knowledge about the parameters in ϕ at least to constrain the hyperparameters into a finite region, if not to assign a substantive hyperprior distribution. As in nonhierarchical models, it is often practical to start with a simple, relatively noninformative, prior distribution on ϕ and seek to add more prior information if there remains too much variation in the posterior distribution.

In the rat tumor example, the hyperparameters are (α, β), which determine the beta distribution for θ. We illustrate one approach to constructing an appropriate hyperprior distribution in the continuation of that example in the next section.

Posterior predictive distributions

Hierarchical models are characterized both by hyperparameters, ϕ, in our notation, and parameters θ. There are two posterior predictive distributions that might be of interest to the data analyst: (1) the distribution of future observations $\tilde{y}$ corresponding to an existing θ_j, or (2) the distribution of observations $\tilde{y}$ corresponding to future θ_j's drawn from the same superpopulation. We label the future θ_j's as $\tilde{\theta}$. Both kinds of replications can be used to assess model adequacy, as we discuss in Chapter 6. In the rat tumor example, future observations can be (1) additional rats from an existing experiment, or (2) results from a future experiment. In the former case, the posterior predictive draws $\tilde{y}$ are based on the posterior draws of θ_j for the existing experiment. In the latter case, one must first draw $\tilde{\theta}$ for the new experiment from the population distribution, given the posterior draws of ϕ, and then draw $\tilde{y}$ given the simulated $\tilde{\theta}$.

5.3 Computation with hierarchical models

Our computational strategy for hierarchical models follows the general approach to multiparameter problems presented in Section 3.8 but is more difficult in practice because of the large number of parameters that commonly appear in a hierarchical model. In particular, we cannot generally plot the contours or display a scatterplot of the simulations from the joint posterior distribution of (θ, ϕ). With care, however, we can follow a similar approach as before, treating θ as the vector parameter of interest and ϕ as the vector of nuisance parameters (though we recognize that both ϕ and θ will be of interest in some problems).

In this section, we present an approach that combines analytical and numerical methods to obtain simulations from the joint posterior distribution, $p(\theta, \phi|y)$, for some simple but important hierarchical models in which the population distribution, $p(\theta|\phi)$, is conjugate to the likelihood, $p(y|\theta)$. For the many nonconjugate hierarchical models that arise in practice, more advanced computational methods, presented in Part III of this book, are necessary. Even