Bayesian Hierarchical Response Modeling

In the first chapter, an introduction to Bayesian item response modeling was given. The Bayesian methodology requires careful specification of priors since item response models contain many parameters, often of the same type. A hierarchical modeling approach is introduced that supports the pooling of information to improve the precision of the parameter estimates. The Bayesian approach for handling response modeling issues is given, and specific Bayesian elements related to response modeling problems will be emphasized. It will be shown that the Bayesian paradigm engenders new ways of dealing with measurement error, limited information about many individuals, clustered response data, and different sources of information.

2.1 Pooling Strength

In test situations, interest is focused on the within-individual and between-individual response heterogeneities. The within-individual response heterogeneity provides information about test characteristics, individual response behavior, and the individual level of ability, among other things. The between-individual response heterogeneity provides information about test characteristics, the relationship between ability and individual background information, and clustering effects of respondents, among other things. The within-individual response differences are informative about the average characteristic levels across items, whereas the between-individual response differences are informative about each single item characteristic.

It is important to score respondents on a common scale, and a response model is needed that is capable of generating individual-level parameter estimates and associated uncertainties. At the same time, a response model is needed that is capable of producing item-level estimates and a characterization of the corresponding uncertainties. The focus on within-individual and between-individual differences corresponds with the interest in making inferences at disaggregate and aggregate levels. Bayesian hierarchical response
models will prove to be very useful for making inferences at different hierarchical levels and make it possible to construct posterior distributions of individual-level parameters given that the data are sparse at the individual level.

Although response data are traditionally collected at the individual level, there is a tendency, specifically in large-scale survey studies, to collect (more) background information at various aggregated levels. Obtaining data at different levels allows inferences to be made at different levels. For example, response data from students across schools make it possible to compare students and schools with respect to the performances of the students. Background information can be used for making correct comparisons and also for explaining differences in performance at the student and school levels. The fact that more data become available at different levels creates modeling challenges and provides opportunities for making inferences that exploit the heterogeneity. The main challenge is to obtain accurate individual-level parameter estimates given very limited individual-level information but a large amount of survey respondents.

A common assumption is to regard the individual-level data as independent conditional on individual-level parameters. In that case, the marginal posterior of $N$ respondents' abilities in Equation (1.11) can be written as

$$p(\theta_1, \ldots, \theta_N | \mathbf{y}) \propto \int \prod_i p(y_i | \theta_i, \xi) p(\theta_i, \ldots, \theta_N | \theta_P) p(\xi) \, d\xi \quad (2.1)$$

where the proportionality sign is used since the normalizing constant is left out of the equation.

The joint prior on the abilities is simplified by assuming that the abilities are independent conditional on the hyperparameters $\theta_P$ and identically distributed. The distinction between a parameter and a hyperparameter is that the sampling distribution of the (response) data is expressed directly conditional on the former (Lindley and Smith, 1972). The parameter’s prior density is parameterized by the hyperparameters.

The first term on the right-hand side is the conditional likelihood. It follows that inferences for each respondent’s ability $\theta_i$ can be made independently of all other respondents. Furthermore, the individual-level inferences are based on the sample and prior information. Therefore, the posterior mean is constructed from a combination of the prior mean and a likelihood-based estimate. For example, when the common population parameters $\theta_P$ provide detailed information about the value of $\theta_i$, the posterior mean will be shrunk more towards the prior mean. Note that such a shift towards the prior mean (of item difficulty locations) was shown in the example of Section 1.4.

The amount of shrinkage is determined by the form of the prior and the values of the hyperparameters. When the amount of within-individual infor-
mation is relatively small, the posterior for the $\theta_i$ will reflect a high level of shrinkage induced by the prior. The opposite is also true. The posterior for the $\theta_i$ will reflect a low level of shrinkage induced by the prior when the amount of within-individual information is high. The level of shrinkage induced by the prior can be inferred from the data by constructing a second-stage prior on the hyperparameters $\theta_P$. This prior on the hyperparameters is sometimes called a hyperprior (Berger, 1985). The sample information and the prior for the hyperparameters will be used to make inferences about the $\theta_P$. Then, the level of shrinkage is determined by the information in the data (Exercise 3.3).

In a strictly hierarchical modeling approach, it is assumed that the responses are nested in individuals, individuals nested in groups, and so on. Typically, response data are clustered in a cross-classified way where respondent effects and item effects are present. To separate the effects of individuals and items, two cross-cutting hierarchies need to be specified. Therefore, assume that the item-level response data are independent conditional on the item-level parameters $\xi_k$. This independence assumption leads to a within-item and between-item structure and defines the second hierarchy of the response data. Analogous to the specification of the joint posterior of abilities, the joint posterior of the item parameters can be written as

$$p(\xi_1, \ldots, \xi_K \mid y) \propto \int \prod_k p(y_k \mid \theta, \xi_k) p(\xi_1, \ldots, \xi_K \mid \xi_P) p(\theta) \, d\theta$$

where the prior for the item parameters is structured by assuming independence conditional on the hyperparameters $\xi_P$. Subsequently, prior and sample information are used to estimate the item effects, where shrinkage effects can be inferred from the data by adding a second layer for the hyperparameters $\xi_P$ (see the example in Section 1.4).

### 2.2 From Beliefs to Prior Distributions

The Bayesian approach to response modeling starts with the specification of prior distributions. In general, there are two types of prior information. First, when there is prior information about the values of the model parameters from, for example, related datasets, this information can be used to construct a case-specific prior. This type will be discussed in subsequent chapters when additional data are available and case-specific information can be used to construct priors.

Second, the prior information can come from mathematical properties. Prior distributions can be classified into conjugate or nonconjugate priors. A conjugate prior has the property that the posterior has the same algebraic form as the prior. This has the advantage that the posterior has a known
analytical form, which simplifies the statistical analysis. A nonconjugate prior leads to a posterior that often has a complicated functional form, which makes the statistical analysis mathematically more challenging. A conjugate prior can easily reflect a likely range of parameter values, but remember that the shape of the prior also has an impact on the posterior analysis. Lord (1986) stressed the practical advantages of priors since, for example, the priors restrict the parameter values to a plausible range. Although a range of priors can be considered for item response models, only a limited number based on well-known distributions are used in practice.

The hierarchical or multistage prior will prove to be very useful (e.g., Berger, 1985, Section 3.6; Lindley and Smith, 1972). The hierarchical prior consists of a first-stage prior for many parameters of the same type that are assumed to be independent given hyperparameters. The hyperparameters have their own prior at a second stage. The first-stage cluster-specific parameters (e.g., item and person parameters) can be related with an aggregated data source at the cluster level. The similarity between the first-stage parameters and their possible relationship with aggregated prior information motivates the hierarchical prior based on combining information to improve the precision of each first-stage parameter estimate. The hierarchical prior improves the estimation of the first-stage parameters by pooling information (borrowing strength) over clusters and by accounting for uncertainty in the hyperparameter estimates. Typically, there are many individuals but relatively little response data on each individual. Pooling information over individuals exploits the assumed similarity between the individual parameters to improve the individual parameter estimates. The hierarchical prior can be extended to more than two stages, and such extensions will be discussed in subsequent chapters.

A Hierarchical Prior for Item Parameters

In most cases, there is not much information about the values of the item parameters, and the response data are the source of information to distinguish the item parameters from each other. Without a priori knowledge to distinguish the item parameters, it is reasonable to assume a common distribution for them. In that case, the item parameters have a common population distribution and it is not possible to order or group the parameters. Typically, the parameters are said to be exchangeable in their joint distribution, which means that the joint probability of the item parameters is invariant with respect to permutations of the indices. That is, it is assumed that the prior information about the item parameters is exchangeable.

The exchangeability assumption is closely related to the concept of independently and identically distributed but not the same. Independently and identically distributed implies exchangeability, and exchangeable item parameters do have identical marginal distributions, but they are not necessarily independent. The assumption of exchangeable item parameters is equivalent
to the assumption of conditionally independent and identically distributed item parameters given hyperparameters and a prior density on the hyperparameters.

An intuitive assumption of an item characteristic curve is that the higher a respondent’s ability level the more likely it is that the respondent scores well on the item. This so-called monotonicity assumption implies that \( P(Y_{ik} = 1 | \theta_i) \) is nondecreasing in \( \theta_i \), which is satisfied when the discrimination parameter is restricted to be positive. For example, in Bilog-MG (Zimowski et al., 1996), a lognormal prior distribution can be specified to restrict a discrimination parameter to be positive. A normal prior is often used for each difficulty parameter (e.g., Albert, 1992; Patz and Junker, 1999a). Exchangeability is usually assumed such that the priors have the same hyperparameters in addition to the common form.

The assessment of the hyperparameters can be challenging since they can have a substantial effect on the item parameter estimates. The greater the discrepancy between the sample-based information and the prior information, the larger the amount of shrinkage towards the prior mean when keeping other factors constant. A small prior variance leads to an informative prior and almost no shrinkage. A large prior variance implies a noninformative prior and a small amount of shrinkage. The effective use of a prior for item parameters depends on the hyperparameter specifications, and that requires insight into the test characteristics and respondents. Mislevy (1986) remarked that incorrectly specifying the prior mean can result in biased item difficulty estimates. Below, the appropriateness of the mean and variance of the prior is arranged using an estimation procedure.

Albert (1992) and Patz and Junker (1999a), among others, suggested independent prior distributions for the item parameters with fixed hyperparameter values: a (log)normal prior for the discrimination parameters and a normal prior for the difficulty parameters. Johnson and Albert (1999) defined a hierarchical prior for the discrimination parameters: a normal density prior at stage 1, and at stage 2 an inverse gamma prior for the variance (with hyperparameters equal to one) and a uniform prior for the mean. Bradlow, Wainer and Wang (1999) and Kim, Cohen, Baker, Subkoviak and Leonard (1994) also defined independent hierarchical priors for the item parameters. Tsutakawa and Lin (1986) proposed a bivariate prior for the item parameters induced by the probability of correct responses to the items at different ability levels. This requires specifying the degree of belief about the probability of a correct response to each item at two ability levels. Although their prior provides a way to specify dependence among parameters within an item, it is often difficult to specify prior beliefs about success probabilities objectively.

A more straightforward realistic multivariate normal prior allows for within-item characteristic dependencies; that is,
\[(a_k, b_k)^t \sim \mathcal{N}(\mu, \Sigma) I_{A_k}(a_k), \quad (2.3)\]

where \(A_k = \{a_k \in \mathcal{R}, a_k > 0\}\). The truncated multivariate normal density in Equation (2.3) is an exchangeable prior for item parameters \(\xi_k (k = 1, \ldots, K)\). The hyperparameters are modeled at a higher level since there is usually little known about the mean item discrimination, mean item difficulty, and variances. An inverse Wishart density, denoted as \(IW\), with scale matrix \(\Sigma_0\) and degrees of freedom \(\nu \geq 2\), is commonly used to specify \(\Sigma_\xi\). This is a conjugated prior for the covariance matrix. Then, a normal prior is used to specify the prior mean given the prior variance matrix \(\Sigma_\xi\). The joint prior density for \((\mu, \Sigma)\) equals

\[
\Sigma_\xi \sim IW(\nu, \Sigma_0), \quad (2.4) \\
\mu | \Sigma_\xi \sim \mathcal{N}(\mu_0, \Sigma_\xi / K_0). \quad (2.5)
\]

This joint prior density is known as a normal inverse Wishart density, where \(K_0\) denotes the number of prior measurements (e.g., Gelman, Carlin, Stern and Rubin, 1995). A proper noninformative prior is specified with \(\mu_0 = (1, 0)^t\), \(\nu = 2\), a scale matrix \(\Sigma_0\) that is a minimally informative prior guess of \(\Sigma_\xi\), and \(K_0\) a small number.

The hierarchical prior accounts for within-item dependencies and for uncertainty of the prior’s parameters. The mean and variance of the normal prior for the item parameters are modeled at a higher prior level and need to be estimated from the data. As a result, the hierarchical prior gives rise to shrinkage estimates of the item parameters, where the amount of shrinkage is inferred from the data. An extreme and infinite item parameter estimate caused by the fact that the item is answered correctly or incorrectly by all respondents is avoided due to shrinkage towards the mean of the prior.

When observing responses on a continuous scale that are assumed to be normally distributed, the hierarchical prior presented is a conjugate prior (e.g., Mellenbergh, 1994b). In Section 4.2, for different reasons, the discrete observed response data will be augmented with normally distributed continuous data. The hierarchical normal prior will be shown to be a conjugate prior with respect to the normal likelihood of the augmented data, which will simplify the procedure for making posterior inferences. A lognormal version of the normal prior presented in Equation (2.3) will be discussed in Exercise 2.3 (see also Exercise 4.3).

The following theoretical motivation provides more arguments for modeling a covariance structure between item parameters. Let \(\theta_l\) and \(\theta_u\) be the average ability levels of respondents sampled from low- and high-ability groups with corresponding success probabilities \(P_k(\theta_l)\) and \(P_k(\theta_u)\) for item \(k\), respectively. Consider the difference between success probabilities, \(P_k(\theta_u) - P_k(\theta_l)\).

---

1 A \(q \times q\) random positive-definite symmetric matrix \(\Sigma_\xi\) is distributed according to an \(IW\) distribution with \(\nu\) degrees of freedom and scale matrix \(\Sigma_0\) if its probability density function is proportional to \(|\Sigma_\xi|^{-(\nu+q+1)/2} \exp\left(-\text{tr} \left(\Sigma_0 \Sigma_\xi^{-1}\right)/2\right)\).
as an estimate of discrimination of item $k$. The higher this difference, the better the item discriminates between respondents with low and high abilities. Consider the mean success probability $(P_k(\theta_u) + P_k(\theta_l))/2$ as an estimate of the item difficulty. A linear relationship between the item parameter estimates is defined as

\[
(P_k(\theta_u) + P_k(\theta_l))/2 = \rho (P_k(\theta_u) - P_k(\theta_l))
\]

\[
\iff P_k(\theta_u) + P_k(\theta_l) = 2\rho (P_k(\theta_u) - P_k(\theta_l))
\]

\[
\iff \frac{P_k(\theta_u)}{P_k(\theta_l)} = \frac{\rho + 1/2}{\rho - 1/2}
\]

for $k = 1, \ldots, K$, and for any constant $\rho > 1/2$. A consistency in the ratio of the mean success probabilities across the $K$ items induces a covariance structure between the item parameters. In Equation (2.6), items of decreasing difficulty will discriminate better between the low- and high-ability respondents for $\rho > 1/2$. If the success probabilities on the left-hand side of Equation (2.6) are replaced by the failure probabilities, items of increasing difficulty will discriminate better between the low- and high-ability respondents for $\rho > 1/2$. The relationships show that a consistency in the ratio of group-specific response probabilities across items can correspond with a common within-item dependency structure.

In the three-parameter model, the guessing parameter, $c_k$, is bounded above by one and below by zero since it represents the probability that a respondent correctly guessed the answer to item $k$. A convenient prior is the beta density with parameters $\alpha$ and $\beta$ that reflects $\alpha - 1$ prior successes and $\beta - 1$ prior failures (Swaminathan and Gifford, 1986; Zimowski et al., 1996). Most often the guessing parameters are assumed to be a priori independently and identically beta distributed (Exercise 2.4),

\[
p(c_k | \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \Gamma(\beta)} c_k^{\alpha - 1} (1 - c_k)^{\beta - 1}.
\]  

(2.7)

The normalizing constant contains the gamma function, which is defined as $\Gamma(\alpha) = \int_0^{\infty} x^{\alpha - 1} e^{-x} dx$ for $\alpha > 0$ (Exercise 3.8). A hierarchical prior for the guessing parameters is constructed by modeling the parameters of the beta prior. Usually a uniform hyperprior is specified for the beta parameters.

Both priors in Equations (2.3) and (2.7) are nonconjugate for the observed-data likelihood.\(^2\) In Chapter 4, it will be shown that a nonconjugate Bayesian analysis is more complex and requires a different method for estimating the item parameters. An exchangeable hierarchical normal prior for the three item parameters $(a_k, b_k, c_k)$ is possible via a reparameterization of the guessing parameter. A straightforward way is to use an inverse normal transformation

\(^2\) In Section 4.2, it is shown that, for a specific data augmentation scheme, the posterior of $c_k$ given augmented data is also a beta density, which makes the beta prior in Equation (2.7) a conjugate prior.
function, $\tilde{c}_k = \Phi^{-1}(c_k)$, which ensures that $\tilde{c}_k$ is defined on the whole real line (see Exercise 4.9).

The prior distribution for the threshold parameters in, for example, the graded response model is usually chosen to be noninformative. Albert and Chib (1993) defined a uniform prior distribution for parameter $\kappa_{k,c}$ but truncated to the region $\{\kappa_{k,c} \in \mathcal{R}, \kappa_{k,c-1} < \kappa_{k,c} \leq \kappa_{k,c+1}\}$ to take account of the order constraints. Johnson and Albert (1999) argued that a uniform prior for the thresholds assigns equal weight to all grades, although some grades may be known to be rare. They proposed a more sophisticated prior where the lower threshold equals the upper threshold when they correspond to an unobserved grade. Other prior distributions for the threshold parameters will be discussed in subsequent chapters.

A Hierarchical Prior for Person Parameters

A prior for person parameters assumes that the respondents represent a sample from a known population. In the case of a simple random sample, a subset of individuals (a sample) are chosen from a larger set (a population) and each individual is chosen randomly, where each individual has the same probability of being chosen at any stage during the sampling process. The respondents are assumed to be sampled independently from a large population; that is, the person parameters are independently distributed from a normal population distribution,

$$\theta_i \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2),$$

for $i = 1, \ldots, N$, where the mean and variance parameters are unknown and are modeled at a higher level. A normal inverse gamma prior is the conjugate prior for the normal distribution with unknown mean and variance. Therefore, a joint hyperprior is specified as

$$\sigma_\theta^2 \sim \mathcal{IG}(g_1, g_2),$$

$$\mu_\theta | \sigma_\theta^2 \sim \mathcal{N}(\mu_0, \sigma_\theta^2/n_0),$$

where $g_1$ and $g_2$ are the parameters of the inverse gamma density denoted as $\mathcal{IG}$ and $n_0$ presents the number of prior measurements.³ Other prior population distributions for more complex sampling designs will be thoroughly discussed in later chapters.

2.2.1 Improper Priors

There have been attempts to construct noninformative priors that contain no (or minimal) information about the parameters. The noninformative prior

³ Parameter $\sigma$ is inverse gamma ($\mathcal{IG}$) distributed with shape parameter $g_1$ and scale parameter $g_2$ when its pdf is proportional to $\sigma^{-(g_1+1)} \exp \left(-g_2/\sigma\right)$. The inverse gamma distribution with $\nu$ degrees of freedom corresponds to an inverse chi-square distribution when $g_1 = \nu/2$ and $g_2 = 1/2$. 
does not favor possible values of the parameter over others. A prior for the difficulty parameter that is completely determined or dominated by the likelihood is one that does not change much in the region where the likelihood is appreciable and does not assume large values outside this region. A prior with these properties is referred to as a locally uniform prior (Box and Tiao, 1973, p. 23). An exchangeable locally uniform prior for the difficulty parameters leads to improper independent priors \( p(b_k) \) equal to a constant. When the difficulty parameters have a locally uniform prior, meaning that they bear no strong relationship to one another, they can be considered fixed effects parameters. In that case, interest is focused on the estimation of the difficulty parameters rather than the properties of their population distribution.

For example, a noninformative prior for the item difficulty parameter gives equal weight to all possible values. That is, \( p(b_k) = c \), where \( c \) is a constant greater than zero. As a result, \( \int p(b_k) \, db_k = \infty \), and this noninformative prior is an improper prior since it does not integrate to one. The value of \( c \) is unimportant, and typically the noninformative prior density is \( p(b_k) = 1 \).

Improper priors may lead to improper posteriors, which means that the posterior is not a valid distribution and posterior moments have no meaning (Exercise 6.8). Ghosh, Ghosh, Chen and Agresti (2000) showed that for the one-parameter model the specification of improper priors for person and item parameters leads to an improper joint posterior. Improper priors are often conjugate and may lead to a straightforward simulation-based estimation method. Such simulation-based methods make use of conditional distributions of the model parameters that are available in closed form, but the marginal posterior distributions are not. The conditional distributions can all be well defined and can be simulated from, yet the joint posterior can be improper (Robert and Casella, 1999, pp. 328–332). Demonstrating the property of the posterior is often impossible. The best way to avoid improper posteriors is to use proper priors. Noninformative proper priors can also be used to specify less or no prior information, and they avoid the risk of getting improper posteriors.

### 2.2.2 A Hierarchical Bayes Response Model

Summing up, the posterior density of interest is constructed from a response model for the observed data and a hierarchical prior density. In a hierarchical prior modeling approach, the prior parameters are modeled explicitly and have hyperprior densities at the second stage of the prior.

Suppose posterior inferences are to be made about the item and person parameters that require integration over the density functions of the hyperparameters. The hyperparameters are denoted by \( \theta_P = (\mu_\theta, \sigma^2_\theta) \) and \( \xi_P = (\mu_\xi, \Sigma_\xi) \). The posterior density of interest can be expressed as
\[ p(\xi, \theta \mid y) \propto \int \int p(y \mid \theta, \xi) p(\theta, \xi \mid \theta_P, \xi_P) p(\theta_P, \xi_P) \, d\xi_P d\theta_P \]
\[ \propto \int \int p(y \mid \theta, \xi) p(\theta \mid \theta_P) p(\xi \mid \xi_P) p(\theta_P \mid \theta_P) p(\xi_P) \, d\xi_P d\theta_P \]
\[ \propto \int \int \prod_{i,k} \left[ p(y_{ik} \mid \theta_i, \xi_k) p(\theta_i \mid \theta_P) p(\xi_k \mid \xi_P) \right] \cdot p(\theta_P) p(\xi_P) \, d\xi_P d\theta_P. \]  

(2.11)

The hierarchical prior model can be recognized in the first equation, where the (hyper)parameters of the prior also have a prior density. The item and person parameters are assumed to be independent from each other, and the corresponding hyperparameters are also assumed to be independent from each other. This leads to the factorization in the second equation.

Typically, the observations are assumed to be conditionally and independently distributed given item and person parameters. That is, the observations are assumed to be clustered in a cross-classified way. Furthermore, the person parameters as well as the item parameters are assumed to be independent from one another. As a result, the joint posterior of the parameters of interest can be expressed as a product of identically distributed observations given person and item parameters, where the person and the item parameters are identically distributed given common hyperparameters; see Equation (2.11).

The last factorization, Equation (2.11), illustrates the hierarchical modeling approach. The observations are modeled conditionally independent at the first stage given item and person parameters, \( p(y_{ik} \mid \theta_i, \xi_k) \). This is the likelihood part of the model which describes the distribution of the data given first-stage parameters. At the second stage, priors are specified for the first-stage parameters. The first-stage priors consist of a prior describing the between-individual heterogeneity, \( p(\theta_i \mid \theta_P) \), and a prior describing the between-item heterogeneity, \( p(\xi_k \mid \xi_P) \). At the third stage, hyperpriors are defined for the parameters of the first-stage priors.

Typically, the variability between individuals is modeled via a conditionally independent prior by conditioning on second-stage parameters. This allows making inferences independently of other respondents, and the conditional independence assumption simplifies the joint prior for the numerous person parameters. In the same way, the prior that describes the between-item variability in item characteristics assumes independent item characteristics given second-stage parameters. The second-stage parameters or hyperparameters control the priors for the lower-level parameters. Then, a second-stage prior for the hyperparameters, \( p(\theta_P) \) and \( p(\xi_P) \), is defined. The combination of a likelihood and a hierarchical prior that consists of a first-stage conditional independent prior and a second-stage prior for the hyperparameters is called a hierarchical Bayes model.

Inferences about the parameters of interest, the first-stage parameters, are based on information from the data and prior information. The contribution
of the first-stage prior to the posterior depends on the values of the prior parameters. For example, the posterior means of the first-stage parameters will show an amount of shrinkage towards the prior mean depending on the values of the hyperparameters. More shrinkage will occur when specifying a more informative first-stage prior. Assessment of the hyperparameters is difficult, and it is desirable to let the level of control of the first-stage prior be driven by information in the data. Therefore, the hyperparameters are modeled at the second stage of the prior. As a result, the priors’ contribution to the posterior is arranged by information in the data taking the hyperparameters’ uncertainty into account.

**Posterior Computation**

The integration problem quickly expands when computing marginal posterior means for all parameters and when extending the prior and/or likelihood model. For example, the computation of the ability posterior density requires integration over the item parameter, item population, and ability population densities. This leads to the computation of

\[
p(\theta \mid y) \propto \int \int \int p(y \mid \theta, \xi) p(\theta \mid \theta_P) p(\theta_P) p(\xi \mid \xi_P) d\theta_P d\xi d\xi_P.
\]  

When assuming normal priors, the integrations might be performed via a Gauss Hermite. Although other quadrature (numerical integration) methods are available, a quadrature approach is limited to a certain range of integration problems.

Obtaining a satisfactory approximation of the integration problem in (2.12) via numerical integration is a complex task, and the available numerical integration methods cannot handle integrals in dimensions greater than four. It will be shown that in large-scale survey research, the computation of properties of posteriors of interest is often further complicated due to the use of complex parameter spaces (e.g., constrained parameter sets), complex sampling models with intractable likelihoods (e.g., due to the presence of missing data), an extremely large dataset, complex prior distributions, or a complex inferential procedure.

In the next chapter, simulation-based Bayesian estimation methods are discussed that are capable of estimating all model parameters simultaneously. This powerful estimation method supports the hierarchical modeling framework in a natural way. In subsequent chapters, it will be shown that all kinds of model extensions can also be handled by the computational methods discussed.
Hierarchical response modeling makes up part of the popular hierarchical modeling framework, which has a tremendous amount of literature. The hierarchical models are used in many areas of statistics. For a more complete overview with applications in the educational and social sciences, see, for example, Goldstein (2003), Longford (1993), Raudenbush and Bryk (2002), Skrondal and Rabe-Hesketh (2004), and Snijders and Bosker (1999).

Recently, the Bayesian formulation (Berger, 1985; Lindley and Smith, 1972) has received much attention. Congdon (2001), Gelman and Hill (2007), and Rossi et al. (2005) show various examples of Bayesian hierarchical modeling using WinBUGS and R. The Bayesian modeling approach accounts for uncertainty in the variance parameters, which is particularly important when the hierarchical variances are difficult to estimate or to distinguish from zero (see Carlin and Louis, 1996; Gelman et al., 1995). The Bayesian hierarchical modeling approach for response data will be thoroughly discussed in subsequent chapters.

The Bayesian hierarchical item response modeling framework has been advocated by Mislevy (1986), Novick, Lewis and Jackson (1973), Swaminathan and Gifford (1982, 1985), Tsutakawa and Lin (1986), and Tsutakawa and Soltys (1988). Swaminathan and Gifford (1982) defined a hierarchical Rasch model with, at the second stage, normal priors for the ability and difficulty parameters. At the third stage, parameters of the prior for the ability parameters were fixed to identify the model. The mean and variance parameters of the exchangeable prior for the difficulty parameters were assumed to be uniformly and inverse chi-square distributed, respectively. For the hierarchical two-parameter model, a chi-square prior was additionally specified for the discrimination parameters with fixed parameters (Swaminathan and Gifford, 1985, 1986). Tsutakawa and Lin (1986) described a hierarchical prior for the difficulty parameters and a standard normal prior for the ability parameters.

In the 1990s, the hierarchical item response modeling approach was picked up by Kim et al. (1994) and Bradlow et al. (1999). After the millennium, the hierarchical item response modeling approach became more popular. This approach will be further pursued in Chapter 4, where several examples are given to illustrate its usefulness.
2.4 Exercises

The exercises can be made using WinBUGS and are focused on hierarchical item response modeling. Different response models are specified in WinBUGS, and the output is used for making posterior inferences. The computational aspects will be discussed in Chapter 3. When it is required, run one chain of 10,000 MCMC iterations and use the last 5,000 iterations.

2.1. Johnson and Albert (1999) consider mathematics placement test data of freshmen at Bowling Green State University. The mathematics placement test is designed to assess the math skill levels of the entering students and recommend an appropriate first college math class. The test form B consists of 35 (dichotomously scored) multiple-choice items.

(a) Adjust the code in Listing 1.1 to fit a two-parameter logistic response model. Obtain item parameter estimates given the fixed hyperparameter values.

(b) Explain that a less informative prior is defined when increasing the prior variance.

(c) Estimate the item parameters for different item prior variances.

(d) Observe and explain the effects of shrinkage by comparing the item parameter estimates.

(e) Evaluate the estimated posterior standard deviations of the item parameters. Do they change due to different hyperparameter settings?

2.2. (continuation of Exercise 2.1) In this problem, attention is focused on assessing the parameters of the difficulty and discrimination prior.

(a) Explain the hierarchical item prior specified in Listing 2.1.

Listing 2.1. WinBUGS code: Independent hierarchical prior for item parameters.

```winbugs
for (k in 1:K) {
  a[k] ~ dnorm(mu[1], prec[1]) I(0,)
  b[k] ~ dnorm(mu[2], prec[2])
}
mu[1] ~ dnorm(1, 1.0E-02)
mu[2] ~ dnorm(0, 1.0E-02)
prec[1] ~ dgamma(1, 1)
prec[2] ~ dgamma(1, 1)
```

(b) Fit the two-parameter logistic response model with the independent hierarchical prior for the item parameters. Investigate the influence of the second-stage prior settings. (Use appropriate starting values for the hyperparameters.)

(c) Plot the estimated posterior means of the discrimination parameters against the differences between the estimated posterior means of Exercises
2.2(b) and 2.1(a), and explain the graph. Do the same for the difficulty parameters.

2.3. (continuation of Exercise 2.1) In this problem, the within-item characteristic dependencies are investigated.
(a) Given the results from Exercise 2.2(b), plot the posterior means of the difficulty parameters against the posterior means of the discrimination parameters and see whether there is a trend visible.
(b) A hierarchical prior for the item parameters is specified in Listing 2.2. Explain that the discrimination parameters are positively restricted by a log transformation.

Listing 2.2. WinBUGS code: Hierarchical prior for item parameters.

```winbugs
def (k in 1:K) {
  item[k, 1:2] ~ dnorm(mu[1:2], prec[1:2, 1:2])
  a[k] <- exp(item[k, 1])
  b[k] <- item[k, 2]
}
p préc [1:2, 1:2] ~ dwish(S[1:2, 1:2], 4)
Sigma[1:2, 1:2] <- inverse(prec[1:2, 1:2])
```

(c) Use the code in Listing 2.2 to define a two-parameter logistic response model with the hierarchical prior for the item parameters, and fit the model. (Use appropriate starting values for the hyperparameters.)
(d) Evaluate the estimated posterior means of the within-item covariance parameters, and explain the results.

2.4. (continuation of Exercise 2.1) In this problem, the three-parameter logistic response model is used to study guessing behavior of students taking the mathematics placement test.
(a) Adjust the code of Listing 1.1 to define a three-parameter response model (Equation (1.5)) with a beta prior for the guessing parameters (Equation (2.7)).
(b) The parameters $\alpha$ and $\beta$ of the beta prior specify the mean and variance of the density via

$$E(c_k) = \frac{\alpha}{\alpha + \beta},$$

$$\text{Var}(c_k) = \frac{\alpha \beta}{(\alpha + \beta)^2(\alpha + \beta + 1)},$$

and the beta density is unimodal when both parameters are greater than one. Specify the beta parameters such that a priori the mean guessing probability is .25 (random guessing) with a standard deviation greater than .1. Fit the three-parameter model, and evaluate the estimated posterior means of the guessing parameters. Do they vary around the prior mean?
(c) Assume, at the second prior stage, that the parameters of the beta prior are uniformly distributed on the interval $[2, 100]$, and fit the three-parameter model. Does the beta prior dominate the final results?
Bayesian Item Response Modeling
Theory and Applications
Fox, J.-P.
2010, XIV, 313 p., Hardcover
ISBN: 978-1-4419-0741-7