10 Memory retention

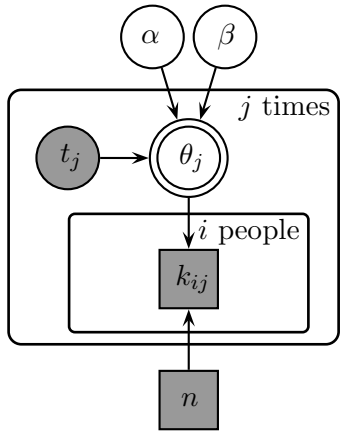
Finding a lawful relationship between memory retention and time is one of the oldest cognitive modeling question, going back to Ebbinghaus in the 1880s. The usual experiment involves giving people many items of information on a list, and then testing their ability to remember items from the list after different periods of time have elapsed. Various mathematical functions, usually with psychological interpretations, have been proposed as describing the relationship between time and the level of retention (Rubin & Wenzel, 1996; Rubin, Hinton, & Wenzel, 1999).

We consider a simplified version of the exponential decay model. The model assumes that the probability that an item will be remembered after a period of time t has elapsed is $\theta_t = \exp(-\alpha t) + \beta$, with the restriction $0 < \theta_t < 1$. The α parameter corresponds to the rate of decay of information. The β parameter corresponds to a baseline level of remembering that is assumed to remain even after very long time periods. Our analyses using this model are based on fictitious data from a potential memory retention study, to help illustrate key modeling points.

Table 10.1 Memory retention data for 4 subjects and 10 time intervals.

	Time Interval									
Subject	1	2	4	7	12	21	35	59	99	200
1	18	18	16	13	9	6	4	4	4	?
2	17	13	9	6	4	4	4	4	4	?
3	14	10	6	4	4	4	4	4	4	?
4	?	?	?	?	?	?	?	?	?	?

These data are given in Table 10.1, and relate to 4 subjects tested on 18 items at 10 time intervals: 1, 2, 4, 7, 12, 21, 35, 59, 99, and 200. The number of items tested and the first 9 time intervals are those used by Rubin et al. (1999). Each datum in Table 10.1 counts the number of correct memory recalls for each subject at each time interval. Included in Table 10.1 are missing data, shown by "?" symbols, so that the prediction and generalization properties of models can be tested. All of the subjects have missing data for the final time period of 200, which tests the ability of models to generalize to new measurements. For Subject 4, there are no data at all, which tests the ability of models to generalize to new subjects.



$$\alpha \sim \text{Beta}(1,1)$$
$$\beta \sim \text{Beta}(1,1)$$
$$\theta_j \leftarrow \min\left(1, \exp\left(-\alpha t_j\right) + \beta\right)$$
$$k_{ij} \sim \text{Binomial}(\theta_j, n)$$

Fig. 10.1 Graphical model for the exponential decay model of memory retention, assuming no individual differences.

10.1 No individual differences

The graphical model for our first attempt to account for the data is shown in Figure 10.1. The model assumes that every subject has the same retention curve, and so there is one true value for the α and β parameters. The outer plate corresponds to the different time periods with values given by the observed t_j variable. Together with the α and β parameters, the time period defines the probability θ_j that the jth item will be remembered.

The inner plate corresponds to the subjects. Each has the same probability of recall at any given time period, but their experimental data, given by the success counts k_{ij}, vary, and are binomially distributed according to the success rate and number of trials.

The script `Retention_1.txt` implements the graphical model in WinBUGS. Note that the code calculates the success rate for each subject at each interval separately, and so is more elaborate than it needs to be:

```
# Retention With No Individual Differences
model{
  # Observed and Predicted Data
  for (i in 1:ns){
    for (j in 1:nt){
      k[i,j] ~ dbin(theta[i,j],n)
      predk[i,j] ~ dbin(theta[i,j],n)
    }
  }
  # Retention Rate At Each Lag For Each Subject Decays Exponentially
  for (i in 1:ns){
    for (j in 1:nt){
      theta[i,j] <- min(1,exp(-alpha*t[j])+beta)
    }
```

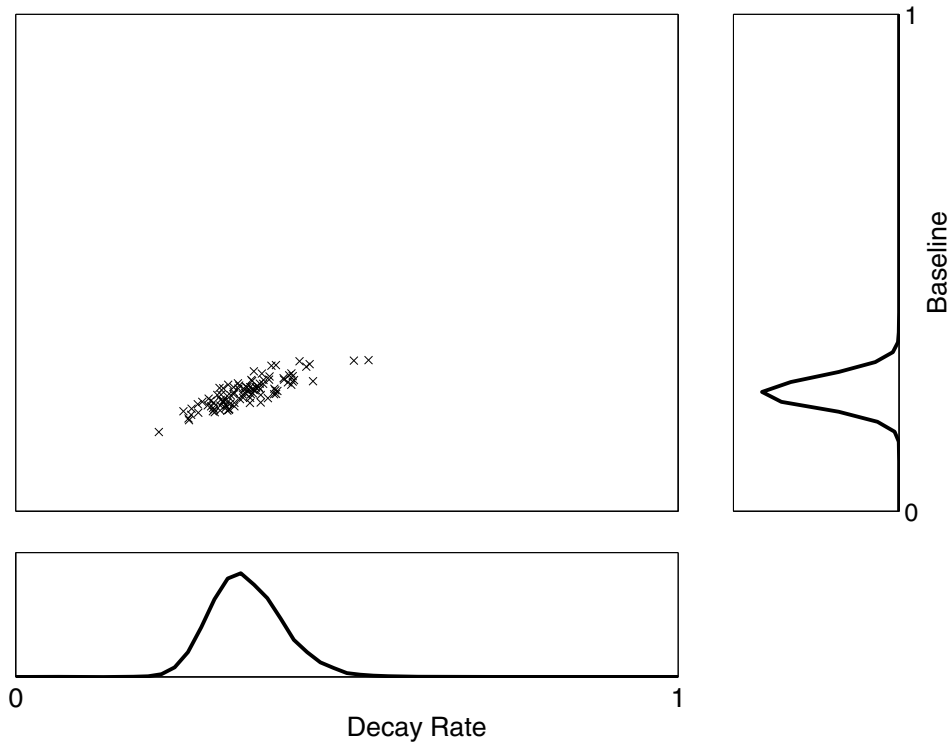


Fig. 10.2 The joint and marginal posterior distributions over the decay and baseline parameters, for the memory retention model that assumes no individual differences.

```
  }
  # Priors
  alpha ~ dbeta(1,1)
  beta ~ dbeta(1,1)
}
```

The code `Retention_1.m` or `Retention_1.R` applies the model to the data, and produces analysis of the posterior and posterior predictive distributions.

The joint posterior distribution of α and β is shown in the main panel of Figure 10.2, as a two-dimensional scatter-plot. Each of the points in the scatter-plot corresponds to a posterior sample selected at random from those available. The marginal distributions of both α and β are shown below and to the right, and are based on all samples. The marginals show the distribution of each parameter, conditioned on the data, considered independently of the other parameter (i.e., averaged across the other parameter).

It is clear from Figure 10.2 that the joint posterior carries more information than the two marginal distributions. If the joint posterior were independent, it would be just the product of the two marginals. But the joint posterior shows a mild relationship, with larger values of α generally corresponding to larger values of β.

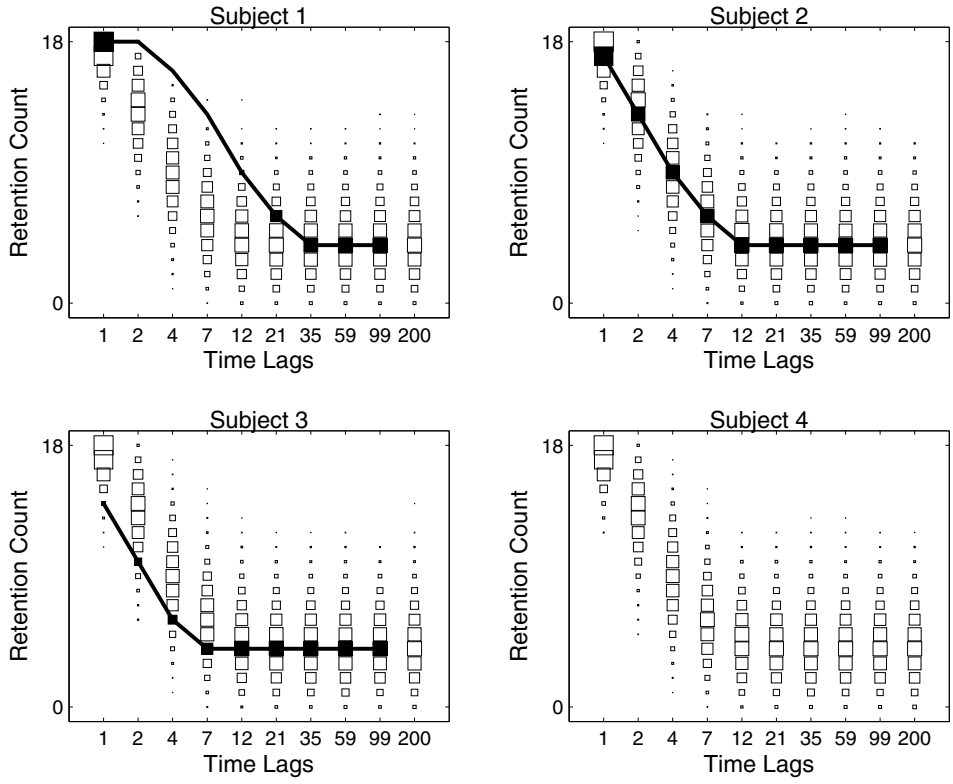
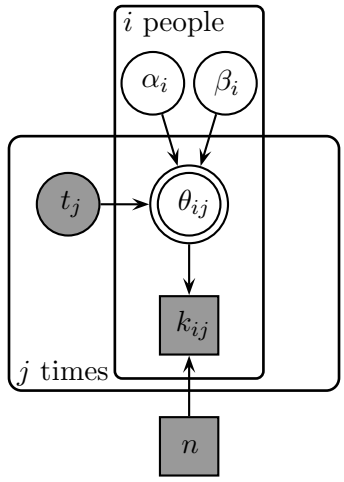


Fig. 10.3 The posterior predictive distribution for the model that assumes no individual differences.

This can be interpreted psychologically as meaning that it is uncertain whether the parameters are a relatively higher baseline coupled with a relatively higher decay rate, or a relatively lower baseline coupled with a relatively lower decay rate.

Figure 10.3 shows the posterior predictive distribution over the number of successful retentions at each time interval. For each subject, at each interval, the squares show the posterior mass given to each possible number of items recalled. These correspond to the model's predictions about observed behavior in the retention experiment, based on what the model has learned from the data. Also shown, by the black squares and connecting lines, are the actual observed data for each subject, where available.

The obvious feature of Figure 10.3 is that the current model does not meet a basic requirement of descriptive adequacy. For both Subjects 1 and 3 the model gives little posterior mass to the observed data at many time periods. It describes a steeper rate of decay than shown by the data of Subject 1, and a shallower rate of decay than shown by the data of Subject 3. After evaluating the model using the posterior predictive analysis, we can conclude that the modeling assumption of no individual differences is inappropriate. It is important to understand that



$$\begin{aligned}
\alpha_i &\sim \text{Beta}(1,1) \\
\beta_i &\sim \text{Beta}(1,1) \\
\theta_{ij} &\leftarrow \min\left(1, \exp\left(-\alpha_i t_j\right) + \beta_i\right) \\
k_{ij} &\sim \text{Binomial}(\theta_{ij}, n)
\end{aligned}$$

Fig. 10.4 Graphical model for the exponential decay model of memory retention, assuming full individual differences.

this conclusion neuters the usefulness of the posterior distribution over parameters shown in Figure 10.2. The posterior distribution is conditioned on the assumption that the model is appropriate, and is not relevant when the model is fundamentally deficient.

Exercise

Exercise 10.1.1 Why is the posterior predictive distribution for all four subjects the same? Are there any (real or fabricated) data that could make the model predict different patterns of retention for different subjects? What if there were massive qualitative differences, such as one subject remembering everything, and the other two remembering nothing?

10.2 Full individual differences

A revised graphical model that does accommodate individual differences is shown in Figure 10.4. The change from the previous model is that the ith subject now has their own α_i and β_i parameters, and that the probability of retention for an item θ_{ij} now changes for both subjects and retention intervals.

The script `Retention_2.txt` implements the graphical model in WinBUGS:

```
# Retention With Full Individual Differences
model{
  # Observed and Predicted Data
  for (i in 1:ns){
```

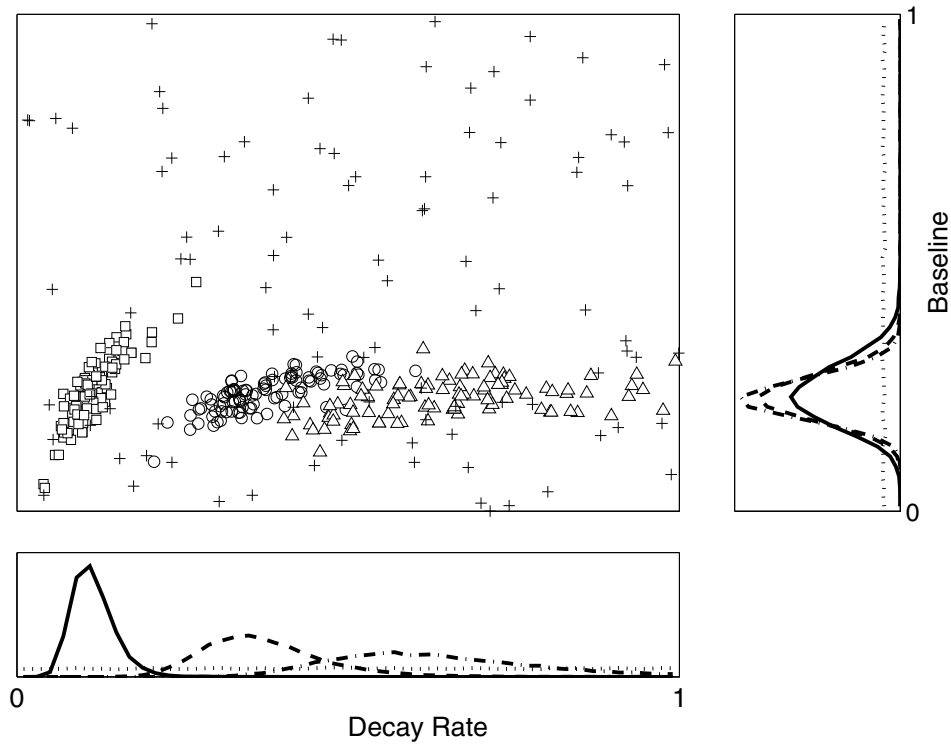


Fig. 10.5 The joint and marginal posterior distributions over the decay and baseline parameters, for the memory retention model that assumes full individual differences.

```
    for (j in 1:nt){
      k[i,j] ~ dbin(theta[i,j],n)
      predk[i,j] ~ dbin(theta[i,j],n)
    }
  }
  # Retention Rate At Each Lag For Each Subject Decays Exponentially
  for (i in 1:ns){
    for (j in 1:nt){
      theta[i,j] <- min(1,exp(-alpha[i]*t[j])+beta[i])
    }
  }
  # Priors For Each Subject
  for (i in 1:ns){
    alpha[i] ~ dbeta(1,1)
    beta[i] ~ dbeta(1,1)
  }
}
```

The code `Retention_2.m` or `Retention_2.R` applies the model to the same data, and again produces analysis of the posterior and posterior predictive distributions.

The joint posterior distributions for each subject are shown in the main panel of Figure 10.5. Each point in the scatter-plot corresponds to a posterior sample,

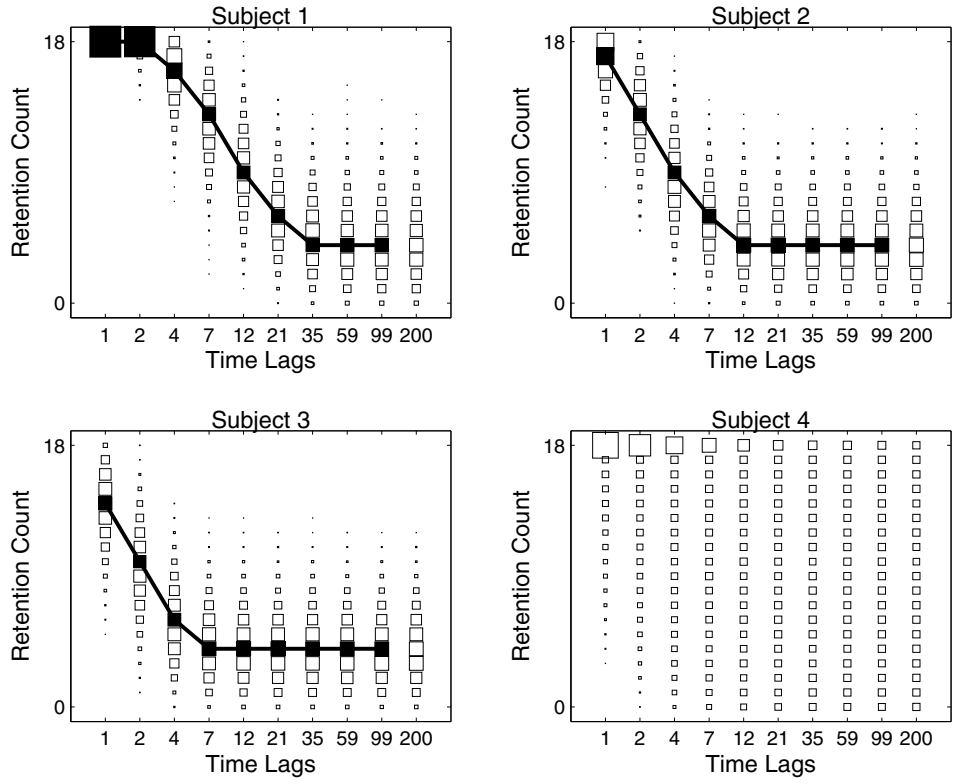


Fig. 10.6 The posterior predictive distribution for the model that assumes full individual differences.

with different markers representing different subjects. The first, second, third, and fourth subjects use square, circular, triangular, and cross markers, respectively. The marginal distributions are shown below and to the right, and use different line styles to represent the subjects.

Figure 10.6 shows the same analysis of the posterior predictive distributions over the number of successful retentions at each time interval, for each subject. It is clear that allowing for individual differences lets the model achieve a basic level of descriptive adequacy for Subjects 1, 2, and 3. The posteriors in Figure 10.5 show that different values for the α decay parameter are used for each of these subjects, corresponding to our intuitions from the earlier analysis.

The weakness in the current model is evident in its predictions for Subject 4. Because each subject is assumed to have decay and baseline parameters that are different, the only information the model has about the new subject is the priors for the α and β parameters. The relationships between parameters for subjects that are visually evident in Figure 10.5 are not formally captured by the model. This means, as shown in Figure 10.5, the posteriors for Subject 4 are just the priors, and so the posterior predictive distribution for Subject 4, as shown in Figure 10.6, does not

have any useful structure. In this way, the model fails a basic test of generalizability, since it does not make sensible predictions for the behavior of subjects other than those for whom data are available. Intuitively, one might want to predict that Subject 4 will be likely to have model parameters consistent with regularities in the inferred parameters for Subjects 1, 2, and 3.

Exercise

Exercise 10.2.1 What are the relative strengths and weaknesses of the full individual differences model and the no individual differences model? Think about this, because the hierarchical approach we consider next could be argued to combine the best features of both of these approaches.

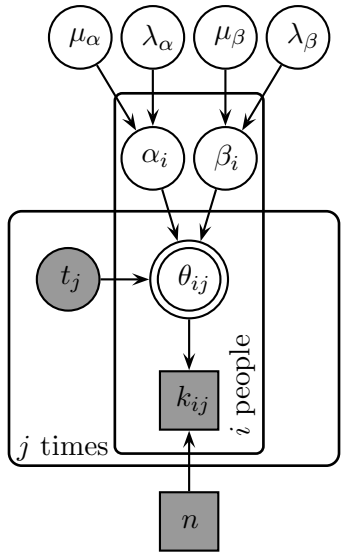
10.3 Structured individual differences

The relationship between the parameters of different subjects, visually evident in Figure 10.5, can be captured using a hierarchical model. A graphical model implementing this approach is shown in Figure 10.7. The key change is that now the α_i and β_i parameters for each subject are modeled as coming from Gaussian distributions. The over-arching Gaussian distribution models this group-level structure for each parameter. This group structure itself has parameters, in the form of means μ_α for the decay and μ_β for the baseline, and precisions λ_α for the decay and λ_β for the baseline. In this way, the individual differences between subjects are given structure.

Each α_i and β_i parameter is independently sampled, so they can be different, but they are sampled from the same distribution, so they have a relationship to one another. This means that inferences made for one subject influence predictions made for another. Since the means and precisions of the group-level distributions are common to all subjects, what is learned about them from one subject affects what is known about another. In addition, because they are sampled from over-arching distributions, the α_i and β_i parameters at the individual subject level no longer have priors explicitly specified, but inherit them from the priors on the means and precisions of the group-level Gaussian distributions.

The script `Retention_3.txt` implements the graphical model in WinBUGS:

```
# Retention With Structured Individual Differences
model{
  # Observed and Predicted Data
  for (i in 1:ns){
    for (j in 1:nt){
      k[i,j] ~ dbin(theta[i,j],n)
      predk[i,j] ~ dbin(theta[i,j],n)
    }
  }
  # Retention Rate At Each Lag For Each Subject Decays Exponentially
```



$$
\begin{aligned}
\mu_\alpha &\sim \text{Beta}(1,1) \\
\lambda_\alpha &\sim \text{Gamma}(.001,.001) \\
\mu_\beta &\sim \text{Beta}(1,1) \\
\lambda_\beta &\sim \text{Gamma}(.001,.001) \\
\alpha_i &\sim \text{Gaussian}(\mu_\alpha, \lambda_\alpha)_{\mathcal{I}(0,1)} \\
\beta_i &\sim \text{Gaussian}(\mu_\beta, \lambda_\beta)_{\mathcal{I}(0,1)} \\
\theta_{ij} &\leftarrow \min\left(1, \exp\left(-\alpha_i t_j\right) + \beta_i\right) \\
k_{ij} &\sim \text{Binomial}(\theta_{ij}, n)
\end{aligned}
$$

Fig. 10.7 Graphical model for the exponential decay model of memory retention, assuming structured individual differences.

```
  for (i in 1:ns){
    for (j in 1:nt){
      theta[i,j] <- min(1,exp(-alpha[i]*t[j])+beta[i])
    }
  }
  # Parameters For Each Subject Drawn From Gaussian Group Distributions
  for (i in 1:ns){
    alpha[i] ~ dnorm(alphamu,alphalambda)I(0,1)
    beta[i] ~ dnorm(betamu,betalambda)I(0,1)
  }
  # Priors For Group Distributions
  alphamu ~ dbeta(1,1)
  alphalambda ~ dgamma(.001,.001)I(.001,)
  alphasigma <- 1/sqrt(alphalambda)
  betamu ~ dbeta(1,1)
  betalambda ~ dgamma(.001,.001)I(.001,)
  betasigma <- 1/sqrt(betalambda)
}
```

The code `Retention_3.m` or `Retention_3.R` applies the model to the data, and again produces an analysis of the posterior and posterior predictive distributions.

The joint and marginal posterior distributions for this model are shown in Figure 10.8 using the same markers and lines as before. For Subjects 1, 2, and 3, these distributions are extremely similar to those found using the full individual differences model. The important difference is for Subject 4, who now has sensible posterior distributions for both parameters. For the decay parameter α there is still considerable uncertainty, consistent with the range of values seen for the first

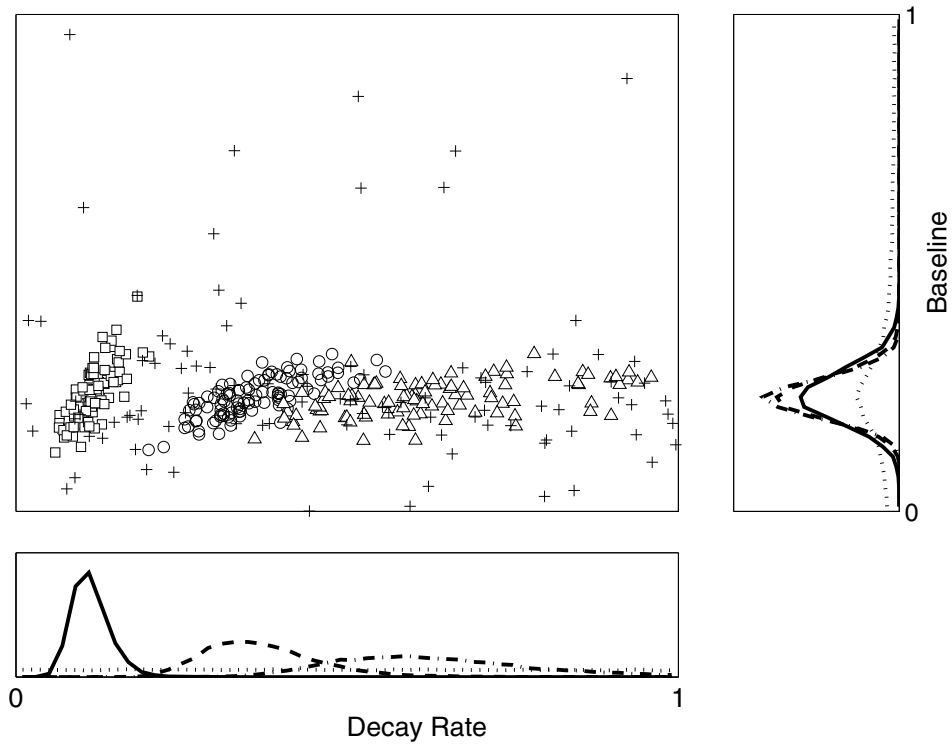


Fig. 10.8 The joint and marginal posterior distributions over the decay and baseline parameters, for the memory retention model that assumes structured individual differences.

three subjects, but for the baseline parameter β, Subject 4 now has a much more constrained posterior distribution.

The posterior predictive distributions for each subject under the hierarchical model are shown in Figure 10.9. The predictions remain useful for the first three subjects, and are now also appropriate for Subject 4. The structured prediction for Subject 4, from whom no data have yet been collected, comes directly from the nature of the hierarchical model. Based on the data from Subjects 1, 2, and 3, inferences are made about the means and precisions of the group distributions for the two parameters of the retention model. The new Subject 4 has values sampled from the Gaussians with these parameters, producing the sensible parameter distributions in Figure 10.8 that lead to the sensible predictive distributions in Figure 10.9.

Psychologically, hierarchical models are powerful because they are able to represent knowledge at different levels of abstraction in a cognitive process (Lee, 2011a). Just as the data have been assumed to be generated by the decay and baseline parameters combining in a memory process for individual subjects, the hierarchical model assumes that those parameters themselves are generated by more abstract latent parameters that describe group distributions across subjects. In other words,

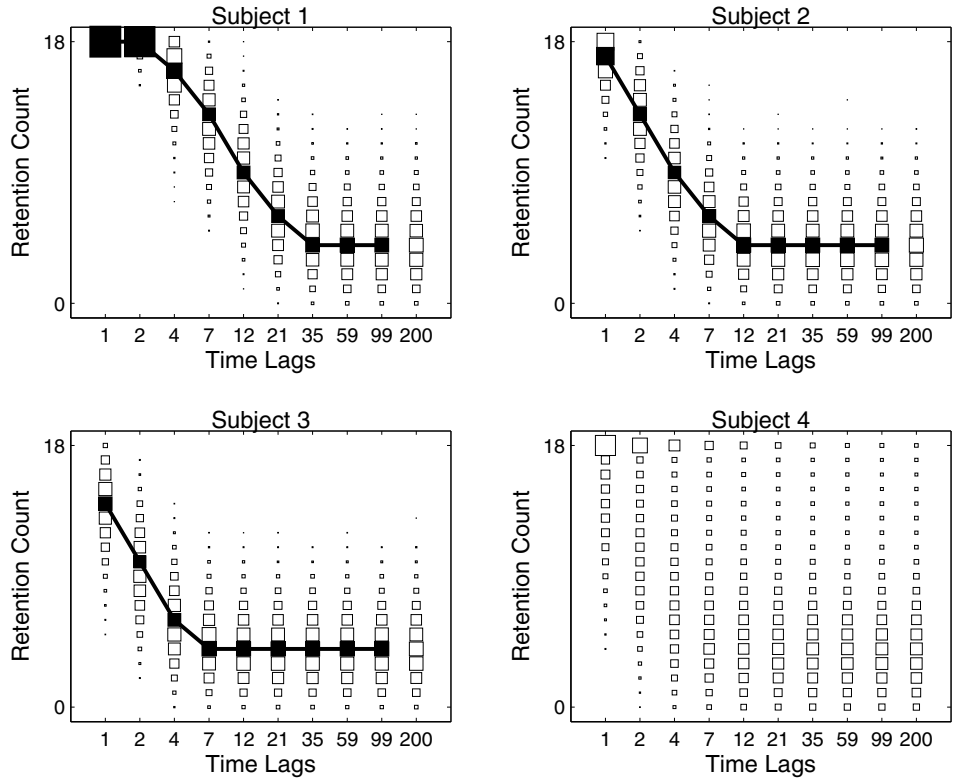


Fig. 10.9 The posterior predictive for the model that assumes structured individual differences.

a hierarchical model lets a theory of memory retention be combined with a theory of individual differences, to provide a more complete account of behavioral data from multiple subjects.

Exercises

Exercise 10.3.1 Think of a psychological model and data, in a different context from the current memory retention example, where the hierarchical approach might be useful.

Exercise 10.3.2 Develop a modified model that does not require you to truncate the rate scale when sampling the α_i decay rates and β_i baselines for each subject. The truncation is not only theoretically inelegant, but technically problematic as it is implemented in WinBUGS, which does not distinguish between censoring and truncation. Implement your modified model and see whether it leads to different conclusions than the ones presented here.