Overthinking: Jacobians and transforms. When you compile a map2stan model that contains varying slopes, and therefore a multivariate prior, you will see a warning like this:

```
Warning (non-fatal): Left-hand side of sampling statement (~) contains a
non-linear transform of a parameter or local variable.
 You must call increment_log_prob() with the log absolute determinant of the
 Jacobian of the transform.
  Sampling Statement left-hand-side expression:
    v_a_blockbp_blockbpc_block ~ multi_normal_log(...)
```

This is quite confusing. What is going on here is that Stan is being cautious and warning the user about a potential problem with the model definition. The multivariate prior for the varying effects transforms the parameters, because it constructs a vector parameter with the multivariate prior. Then the individually named varying effect parameters in the map2stan model are transformed into that common vector. In general, parameter transformations require taking account of any change in the geometry of the posterior distribution that results from the transform. This is done by calculating the rates of change of each dimension (parameter), and the matrix that holds these rates of change is called the *Jacobian*, a matrix of partial derivatives. Multiplying the posterior probability by the absolute value of the determinate of that matrix provides the right adjustment to account for the transform.

Stan detects the presence of a transformation in the varying slopes model, and so it warns you to be sure you've taken account of any change in geometry. In this case, everything is fine, because there is no change in geometry: Each parameter is one-to-one inserted in a fixed position in the vector parameter. So the absolute Jacobian is exactly 1—no adjustment required. But it's nice that Stan is cautious—there's no way for it to verify that the transform has been properly accounted for. So instead of being alarmed, feel the warm glow that comes from knowing that Stan is looking out for you.

13.2. Example: Admission decisions and gender

Now let's return to a previous data example and incorporate varying slopes. This will allow you to appreciate how variation in slopes arises in a natural context, as well as how the correlation between intercepts and slopes can provide hints about process.

Recall the UCBadmit data from Chapter 10. In those data, failing to model the varying means across departments led to exactly the opposite inference of the truth. But we left some information on the floor, so to speak, by not using varying effects to pool information across departments. As a consequence, we probably overfit the smaller departments. We also ignored variation across departments in how they treated male and female applicants. Varying slopes will provide a direct way to model such variation.

Here's the data again, also constructing the dummy variable for male applications and the index variable for departments:

R code
13.17

```
library(rethinking)
data(UCBadmit)
d <- UCBadmit
d$male <- ifelse( d$applicant.gender=="male" , 1 , 0 )
d$dept_id <- coerce_index( d$dept )
```

Now we're ready to fit some models.

13.2.1. Varying intercepts. We'll begin slowly, by presenting just the varying intercept model for these data. Here's the model, with the varying intercept components in blue:

$$
\begin{aligned}
A_i &\sim \text{Binomial}(n_i, p_i) && \text{[likelihood]}\\
\text{logit}(p_i) &= \alpha_{\text{DEPT}[i]} + \beta m_i && \text{[linear model]}\\
\alpha_{\text{DEPT}} &\sim \text{Normal}(\alpha, \sigma) && \text{[prior for varying intercepts]}\\
\alpha &\sim \text{Normal}(0, 10) && \text{[prior for } \alpha\text{]}\\
\beta &\sim \text{Normal}(0, 1) && \text{[prior for } \beta\text{]}\\
\sigma &\sim \text{HalfCauchy}(0, 2) && \text{[prior for } \sigma\text{]}
\end{aligned}
$$

The outcome variable A_i is the number of admit decisions, `admit`, and the sample size in each case is n_i, `applications`. Notice that I have placed the average intercept, α, in the linear model rather than inside the varying intercepts prior. This form is perfectly equivalent, recall, to placing α inside the prior. But it means the α_{DEPT} parameters are now displacements from the average department.

Here's the code to fit the varying intercepts model:

R code
13.18

```
m13.2 <- map2stan(
    alist(
        admit ~ dbinom( applications , p ),
        logit(p) <- a_dept[dept_id] + bm*male,
        a_dept[dept_id] ~ dnorm( a , sigma_dept ),
        a ~ dnorm(0,10),
        bm ~ dnorm(0,1),
        sigma_dept ~ dcauchy(0,2)
    ) ,
    data=d , warmup=500 , iter=4500 , chains=3 )
precis( m13.2 , depth=2 ) # depth=2 to display vector parameters
```

```
            Mean StdDev lower 0.89 upper 0.89 n_eff Rhat
a_dept[1]   0.67   0.10       0.51       0.83  4456    1
a_dept[2]   0.63   0.12       0.44       0.81  4758    1
a_dept[3]  -0.59   0.08      -0.70      -0.46  6831    1
a_dept[4]  -0.62   0.09      -0.76      -0.48  5258    1
a_dept[5]  -1.06   0.10      -1.22      -0.90  9368    1
a_dept[6]  -2.61   0.16      -2.87      -2.36  6969    1
a          -0.59   0.64      -1.61       0.36  5115    1
bm         -0.09   0.08      -0.22       0.04  3480    1
sigma_dept  1.48   0.60       0.71       2.16  4633    1
```

The estimated effect of `male` is very similar to what we got in Chapter 10. But now we also have better estimates of the individual department average acceptance rates. You'll see that the departments are ordered from those with the highest proportions accepted to the lowest. Remember, the values above are the α_{DEPT} estimates, and so they are deviations from the global mean α, which in this case has posterior mean -0.58. So department A, "[1]" in the table, has the highest average admission rate. Department F, "[6]" in the table, has the lowest.

13.2.2. Varying effects of being male. Now let's consider the variation in gender bias among departments. Sure, overall there isn't much evidence of gender bias in the previous model. But what if we allow the effect of an applicant's being male to vary in the same way we already allowed the overall rate of admission to vary? This will constitute the varying slopes model in this context.

One extra feature of varying slopes that will arise here is that since there is substantial *imbalance* in sample size across departments and the numbers of male and female applications they received, pooling will be stronger for those cases with fewer applications. Department B, for example, received only 25 applications from females. So any estimate of how that department differently treats males and females will shrink towards the population average. In contrast, department F received hundreds of applications from both males and females. So pooling will do very little to the estimates for that department.

This is what the varying slopes model looks like, with the varying effects components in blue:

$$
\begin{aligned}
A_i &\sim \text{Binomial}(n_i, p_i) && \text{[likelihood]} \\
\text{logit}(p_i) &= \alpha_{\text{DEPT}[i]} + \beta_{\text{DEPT}[i]} m_i && \text{[linear model]} \\
\begin{bmatrix} \alpha_{\text{DEPT}} \\ \beta_{\text{DEPT}} \end{bmatrix} &\sim \text{MVNormal}\left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \mathbf{S} \right) && \text{[joint prior for varying effects]} \\
\mathbf{S} &= \begin{pmatrix} \sigma_\alpha & 0 \\ 0 & \sigma_\beta \end{pmatrix} \mathbf{R} \begin{pmatrix} \sigma_\alpha & 0 \\ 0 & \sigma_\beta \end{pmatrix} && \\
\alpha &\sim \text{Normal}(0, 10) && \text{[prior for } \alpha\text{]} \\
\beta &\sim \text{Normal}(0, 1) && \text{[prior for } \beta\text{]} \\
(\sigma_\alpha, \sigma_\beta) &\sim \text{HalfCauchy}(0, 2) && \text{[prior for each } \sigma\text{]} \\
\mathbf{R} &\sim \text{LKJcorr}(2) && \text{[prior for correlation matrix]}
\end{aligned}
$$

The symbol m_i indicates the value of `male` for the i-th row. It is multiplied by the sum $\beta + \beta_{\text{DEPT}[i]}$, which is a total slope defined by both a value common to all departments, β, and a value unique to the department for row i, $\beta_{\text{DEPT}[i]}$.

To fit this model:

R code 13.19

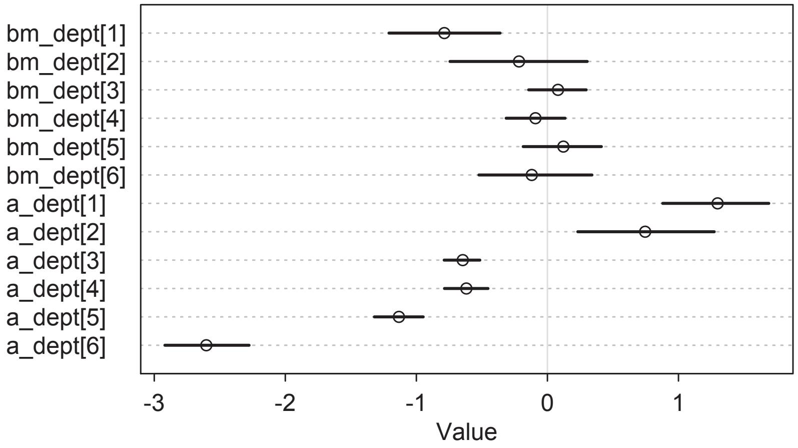
```
m13.3 <- map2stan(
    alist(
        admit ~ dbinom( applications , p ),
        logit(p) <- a_dept[dept_id] +
                    bm_dept[dept_id]*male,
        c(a_dept,bm_dept)[dept_id] ~ dmvnorm2( c(a,bm) , sigma_dept , Rho ),
        a ~ dnorm(0,10),
        bm ~ dnorm(0,1),
        sigma_dept ~ dcauchy(0,2),
        Rho ~ dlkjcorr(2)
    ) ,
    data=d , warmup=1000 , iter=5000 , chains=4 , cores=3 )
```

Check for yourself that the chains mixed and converged excellently. You might get a warning about a few “divergent iterations.” We’ll focus on those in the next section.

We’re interested in what adding varying slopes has revealed. So let’s look at the marginal posterior distributions for the varying effects only:

```
plot( precis(m13.3,pars=c("a_dept","bm_dept"),depth=2) )
```

R code 13.20



Notice that the intercepts range all over the place, while the slopes all cling close to zero. This reflects the fact that departments varied a lot in overall admission rates, but they neither discriminated much between male and female applicants nor varied much in how much they discriminated.

But there are a few departments with slopes consistent with noticeable bias, departments 1 and 2 in particular. Department 1 has a slope estimate centered almost 1 log-odds below the mean, and it doesn't have any mass on the other side of zero. Department 2 has a highly uncertain slope, but that means it also includes plausibly large effects. Just because a marginal posterior overlaps zero does not mean we should think of it as zero.

Notice also that these two departments have the largest intercepts. So let's look at the estimated correlation between intercepts and slopes next, as well as the two-dimensional shrinkage it induces.

13.2.3. Shrinkage. The posterior correlation between intercepts and slopes is shown in the left-hand plot of FIGURE 13.6. The majority of the probability mass is below zero, indicating a negative correlation. This corresponds to the fact from just above: The departments with the highest admissions rates also have the smallest slopes.

The right-hand plot shows the shrinkage in both intercepts and slopes. This plot is analogous to the shrinkage plot from the previous section, FIGURE 13.5 (page 397). The blue points are again the raw empirical (unpooled) estimates. The open points are the varying effect (adaptively pooled) estimates. The lines connect points from the same departments, and the text labels correspond to the department labels. The gray contours show the inferred population of intercepts and slopes.

Again, shrinkage follows a negative correlation, although a weaker one in this case. The department with the most shrinkage is A, which has the most extreme intercept and slope. More interesting and instructive is the shrinkage for department F. That department had an average raw slope, but an unusually low intercept. As a result, the intercept moves a little towards the mean, as a result of pooling, and the slope moves down, as a result of the negative correlation in the population. That is to say that the model thinks that, since high intercepts

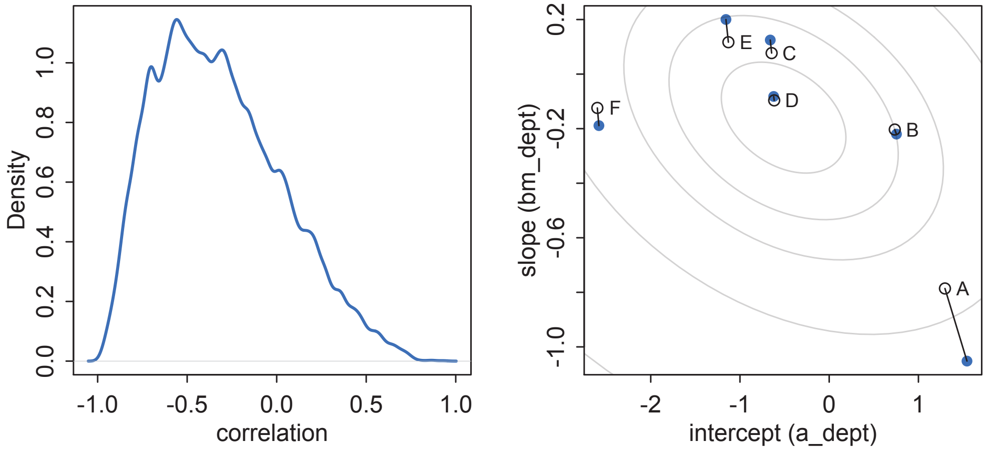


FIGURE 13.6. Left: Posterior distribution of the correlation between intercepts and slopes for the UCB admissions model, `m13.3`. Right: Two-dimensional shrinkage of unpooled (blue) and adaptively pooled (open) estimates.

and associated with low slopes overall, if department F's intercepts are too small, then its slope is probably also too big.

Overall, there isn't much shrinkage going on in this model. This is because there are a lot of applications on each row. But even though the absolute scale of shrinkage is small here, the nature of the shrinkage again illustrates the properties of varying effects.

13.2.4. Model comparison. To make this more interesting, we can also fit the model that ignores gender. Then we can compare all of these models, using WAIC.

R code 13.21

```
m13.4 <- map2stan(
    alist(
        admit ~ dbinom( applications , p ),
        logit(p) <- a_dept[dept_id],
        a_dept[dept_id] ~ dnorm( a , sigma_dept ),
        a ~ dnorm(0,10),
        sigma_dept ~ dcauchy(0,2)
    ) ,
    data=d , warmup=500 , iter=4500 , chains=3 )

compare( m13.2 , m13.3 , m13.4 )
```

```
        WAIC pWAIC dWAIC weight    SE  dSE
m13.3 5191.4  11.3   0.0   0.99 57.30   NA
m13.4 5201.2   6.0   9.8   0.01 56.84 6.84
m13.2 5201.9   7.2  10.5   0.01 56.94 6.50
```

The model that ignores gender, `m13.4`, earns the same expected out-of-sample performance as the model that includes a *constant* effect of gender, `m13.2`. The varying slopes model, `m13.3`, dominates both. This is despite the fact that the *average* slope in `m13.3` is nearly zero.

The average isn't what matters, however. It is the individual slopes, one for each department, that matter. If we wish to generalize to new departments, the variation in slopes suggests that it'll be worth paying attention to gender, even if the average slope is nearly zero in the population.

13.2.5. More slopes. The varying slopes strategy generalizes to as many slopes as you like, within practical limits. All that happens is that each new predictor you want to construct varying slopes for adds one more dimension to the covariance matrix of the varying effects prior. So this means one more standard deviation parameter and one more dimension to the correlation matrix.

For example, suppose the UCB admissions data also recorded the test scores of each applicant. Then we could also include test score as a predictor. But we don't have another predictor for these data. So instead we'll turn to another data set, in the next section, to go deeper into varying slopes.

13.3. Example: Cross-classified chimpanzees with varying slopes

To see how to construct a model with more than two varying effects—varying intercepts plus more than one varying slope—as well as with more than one type of cluster, we'll return to the chimpanzee experiment data that was introduced in Chapter 10. In these data, there are two types of clusters: actors and blocks. We explored *cross-classification* with two kinds of varying intercepts back on page 370. We also modeled the experiment with two different slopes: one for the effect of the prosocial option (the side of the table with two pieces of food) and one for the interaction between the prosocial option and the presence of another chimpanzee. So now we'll model both types of clusters and place varying effects on the intercepts and both slopes.

I'll also use this example to emphasize the importance of **NON-CENTERED PARAMETERIZATION** for some multilevel models. For any given multilevel model, there are several different ways to write it down. These ways are called "parameterizations." Mathematically, these alternative parameterizations are equivalent, but inside the MCMC engine they are not. Remember, how you fit the model is part of the model. Choosing a better parameterization is an awesome way to improve sampling for your MCMC model fit, and the non-centered parameterization tends to help a lot with complex varying effect models like the one you'll work with in this section. I'll hide the details of the technique in the main text. But as usual, there is an Overthinking box at the end that provides some detail.

Okay, let's construct a cross-classified varying slopes model. To maintain some sanity with this complicated model, we'll use more than one linear model in the formulas. This will allow us to compartmentalize sub-models for the intercepts and each slope. Here's what the likelihood and its linear models look like:

$$
\begin{aligned}
L_i &\sim \text{Binomial}(1, p_i) \\
\text{logit}(p_i) &= \mathcal{A}_i + (\mathcal{B}_{P,i} + \mathcal{B}_{PC,i}C_i)P_i && \text{[linear model skeleton]} \\
\mathcal{A}_i &= \alpha + \alpha_{\text{ACTOR}[i]} + \alpha_{\text{BLOCK}[i]} && \text{[intercept model]} \\
\mathcal{B}_{P,i} &= \beta_P + \beta_{P,\text{ACTOR}[i]} + \beta_{P,\text{BLOCK}[i]} && \text{[}P\text{ slope model]} \\
\mathcal{B}_{PC,i} &= \beta_P + \beta_{PC,\text{ACTOR}[i]} + \beta_{PC,\text{BLOCK}[i]} && \text{[}P \times C\text{ interaction model]}
\end{aligned}
$$