Bayesian Statistics

5. Hierarchical Models
for Combining Information

David Draper

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ams.ucsc.edu

http://www.ams.ucsc.edu/~draper

© 2004 David Draper (all rights reserved)
Contents

5.1 Formulating hierarchical models for quantitative outcomes from scientific context

5.2 Approximate fitting of Gaussian hierarchical models: maximum likelihood and empirical Bayes

5.3 Incorporating study-level covariates
  5.3.1 Shrinkage estimation

5.4 References
5. Hierarchical Models for Combining Information

5.1 Formulating hierarchical models for quantitative outcomes from scientific context

Case Study: Meta-analysis of effects of aspirin on heart attacks. Table 5.1 (Draper et al., 1993a) gives the number of patients and mortality rate from all causes, for six randomized controlled experiments comparing the use of aspirin and placebo by patients following a heart attack.

Table 5.1. Aspirin meta-analysis data.

<table>
<thead>
<tr>
<th>Study (i)</th>
<th>Aspirin</th>
<th>Placebo</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># of Patients</td>
<td>Mortality Rate (%)</td>
</tr>
<tr>
<td>UK-1</td>
<td>615</td>
<td>7.97</td>
</tr>
<tr>
<td>CDPA</td>
<td>758</td>
<td>5.80</td>
</tr>
<tr>
<td>GAMS</td>
<td>317</td>
<td>8.52</td>
</tr>
<tr>
<td>UK-2</td>
<td>832</td>
<td>12.26</td>
</tr>
<tr>
<td>PARIS</td>
<td>810</td>
<td>10.49</td>
</tr>
<tr>
<td>AMIS</td>
<td>2267</td>
<td>10.85</td>
</tr>
<tr>
<td>Total</td>
<td>5599</td>
<td>9.88</td>
</tr>
</tbody>
</table>

\[ y_i = \frac{\text{Diff}}{} \quad \sqrt{V_i} = \text{SE} \]

<table>
<thead>
<tr>
<th>Study (i)</th>
<th>Comparison</th>
<th>( Z_i^{\dagger} )</th>
<th>( p_i^{\S} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK-1</td>
<td>2.77</td>
<td>1.65</td>
<td>1.68</td>
</tr>
<tr>
<td>CDPA</td>
<td>2.50</td>
<td>1.31</td>
<td>1.91</td>
</tr>
<tr>
<td>GAMS</td>
<td>1.84</td>
<td>2.34</td>
<td>0.79</td>
</tr>
<tr>
<td>UK-2</td>
<td>2.56</td>
<td>1.67</td>
<td>1.54</td>
</tr>
<tr>
<td>PARIS</td>
<td>2.31</td>
<td>1.98</td>
<td>1.17</td>
</tr>
<tr>
<td>AMIS</td>
<td>-1.15</td>
<td>0.90</td>
<td>-1.27</td>
</tr>
<tr>
<td>Total</td>
<td>0.86</td>
<td>0.59</td>
<td>1.47</td>
</tr>
</tbody>
</table>

\( Z_i \) denotes the ratio of the difference in mortality rates over its standard error, assuming a binomial distribution. \( p_i \) is the one-sided \( p \) value associated with \( Z_i \), using the normal approximation.
Meta-Analysis

The first five trials are reasonably consistent in showing a (weighted average) mortality decline for aspirin patients of 2.3 percentage points, a 20% drop from the (weighted average) placebo mortality of 11.5% (this difference is highly clinically significant).

However, the sixth and largest trial, AMIS, went the other way: an increase of 1.2 percentage points in aspirin mortality (a 12% rise from the placebo baseline of 9.7%).

Some relevant questions (Draper, 1995):

Q₁ Why did AMIS get such different results?

Q₂ What should be done next to reduce the uncertainty about Q₁?

Q₃ If you were a doctor treating a patient like those eligible for the trials in Table 5.1, what therapy should you employ while answers to Q₁ and Q₂ are sought?

One possible paraphrase of Q₃: Q₄ How should the information from these six experiments be combined to produce a more informative summary than those obtained from each experiment by itself?

The discipline of meta-analysis is devoted to answering questions like Q₄.

One leading school of frequentist meta-analysis (e.g., Hedges and Olkin, 1985) looks for methods for combining the Z and p values in Table 5.1, an approach that often leads only to an overall p value.
A Gaussian HM

A more satisfying form of meta-analysis (which has both frequentist and Bayesian versions) builds a hierarchical model (HM) that indicates how to combine information from the mortality differences in the table.

A Gaussian meta-analysis model for the aspirin data, for example (Draper et al., 1993a), might look like

\[
(\theta, \sigma^2) \sim p(\theta, \sigma^2) \quad \text{(prior)}
\]

\[
(\theta_i|\theta, \sigma^2) \overset{\text{iid}}{\sim} N(\theta, \sigma^2) \quad \text{(underlying effects)} \quad (1)
\]

\[
(y_i|\theta_i) \overset{\text{indep}}{\sim} N(\theta_i, V_i) \quad \text{(data)}.
\]

The bottom level of (1), the data level of the HM, says that—because of relevant differences in patient cohorts and treatment protocols—each study has its own underlying treatment effect \(\theta_i\), and the observed mortality differences \(y_i\) are like random draws from a normal distribution with mean \(\theta_i\) and variance \(V_i\) (the normality is reasonable because of the Central Limit Theorem, given the large numbers of patients).

In meta-analyses of data like those in Table 5.1 the \(V_i\) are typically taken to be known (again because the patient sample sizes are so big), \(V_i = SE_i^2\), where \(SE_i\) is the standard error of the mortality difference for study \(i\) in Table 5.1.

The middle level of the HM is where you would bring in the study-level covariates, if you have any, to try to explain why the studies differ in their underlying effects.

Here there are no study-level covariates, so the middle level of (1) is equivalent to a Gaussian regression with no predictor variables.
A Gaussian HM (continued)

Why the “error” distribution should be Gaussian at this level of the HM is not clear—it’s a conventional option, not a choice that’s automatically scientifically reasonable (in fact I’ll challenge it later).

$\sigma^2$ in this model represents study-level heterogeneity.

The top level of (1) is where the prior distribution on the regression parameters from the middle level is specified.

Here, with only an intercept term in the regression model, a popular conventional choice is the normal/scaled-inverse-$\chi^2$ prior we looked at earlier in our first Gaussian case study.

**Fixed effects and random effects.** If $\sigma^2$ were known somehow to be 0, all of the $\theta_i$ would have to be equal deterministically to a common value $\theta$, yielding a simpler model: $(y_i|\theta) \overset{\text{indep}}{\sim} N(\theta, V_i), \theta \sim p(\theta)$.

Meta-analysts call this a fixed-effects model, and refer to model (1) as a random-effects model.

When $\sigma^2$ is not assumed to be 0, with this terminology the $\theta_i$ are called random effects (this parallels the use of this term in the random-effects Poisson regression case study).
5.2 Approximate Fitting of Gaussian Hierarchical Models: Maximum Likelihood and Empirical Bayes

Fitting HM (1). Some algebra based on model (1) yields that the conditional distributions of the study-level effects \( \theta_i \) given the data and the parameters \( (\theta, \sigma^2) \), have a simple and revealing form:

\[
(\theta_i|y_i, \theta, \sigma^2) \overset{\text{indep}}{\sim} N[\theta_i^*, V_i(1 - B_i)],
\]

with \( \theta_i^* = (1 - B_i) y_i + B_i \theta \) and \( B_i = \frac{V_i}{V_i + \sigma^2} \).

In other words, the conditional mean of the effect for study \( i \) given \( y_i, \theta, \) and \( \sigma^2 \) is a weighted average of the sample mean for that study, \( y_i \), and the overall mean \( \theta \).

The weights are given by the so-called shrinkage factors \( B_i \) (e.g., Draper et al., 1993a), which in turn depend on how the variability \( V_i \) within study \( i \) compares to the between-study variability \( \sigma^2 \): the more accurately \( y_i \) estimates \( \theta_i \), the more weight the “local” estimate \( y_i \) gets in the weighted average.

The term shrinkage refers to the fact that, with this approach, unusually high or low individual studies are drawn back or “shrunken” toward the overall mean \( \theta \) when making the calculation \( (1 - B_i) y_i + B_i \theta \).

Note that \( \theta_i^* \) uses data from all the studies to estimate the effect for study \( i \)—this is referred to as borrowing strength in the estimation process.

Closed-form expressions for \( p(\theta|y) \) and \( p(\theta_i|y) \) with \( y = (y_1, \ldots, y_k), k = 6 \) are not available even with a normal/scaled-inverse-\( \chi^2 \) prior for \( (\theta, \sigma^2) \); MCMC is needed (see below).
Maximal Likelihood and Empirical Bayes

In the meantime, maximum likelihood calculations provide some idea of what to expect: the likelihood function based on model (1) is

\[
l(\theta, \sigma^2 | y) = c \prod_{i=1}^{k} \frac{1}{\sqrt{V_i + \sigma^2}} \exp \left[ -\frac{1}{2} \sum_{i=1}^{k} \frac{(y_i - \theta)^2}{V_i + \sigma^2} \right]. \tag{4}\]

The maximum likelihood estimates (MLEs) \((\hat{\theta}, \hat{\sigma}^2)\) then turn out to be the iterative solutions to the following equations:

\[
\hat{\theta} = \frac{\sum_{i=1}^{k} \hat{W}_i y_i}{\sum_{i=1}^{k} \hat{W}_i} \quad \text{and} \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{k} \hat{W}_i^2 [(y_i - \hat{\theta})^2 - V_i]}{\sum_{i=1}^{k} \hat{W}_i^2}, \tag{5}\]

where \(\hat{W}_i = \frac{1}{V_i + \hat{\sigma}^2}.\) \tag{6}

Start with \(\hat{\sigma}^2 = 0\) and iterate (5–6) to convergence (if \(\hat{\sigma}^2\) converges to a negative value, \(\hat{\sigma}^2 = 0\) is the MLE); the MLEs of the \(\theta_i\) are then given by

\[
\hat{\theta}_i = (1 - \hat{B}_i) y_i + \hat{B}_i \theta \quad \text{where} \quad \hat{B}_i = \frac{V_i}{V_i + \hat{\sigma}^2}. \tag{7}\]

These are called empirical Bayes (EB) estimates of the study-level effects, because it turns out that this analysis approximates a fully Bayesian solution by (in effect) using the data to estimate the prior specifications for \(\theta\) and \(\sigma^2\).

Large-sample (mainly meaning large \(k\)) approximations to the (frequentist) distributions of the MLEs are given by\[
\hat{\theta} \sim N \left( \theta, \left[ \sum_{i=1}^{k} \frac{1}{V_i + \hat{\sigma}^2} \right]^{-1} \right) \quad \text{and} \quad \hat{\theta}_i \sim N[\theta_i, V_i(1 - \hat{B}_i)]. \tag{8}\]
MLEB (continued)

NB The variances in (8) don’t account fully for the uncertainty in $\sigma^2$ and therefore underestimate the actual sampling variances for small $k$ (adjustments are available; see, e.g., Morris (1983, 1988)).

**MLEB estimation** can be implemented simply in about 15 lines of R code (Table 5.2).

Table 5.2. R program to perform MLEB calculations.

```r
mleb <- function( y, V, m ) {
  sigma2 <- 0.0
  for ( i in 1:m ) {
    W <- 1.0 / ( V + sigma2 )
    theta <- sum( W * y ) / sum( W )
    sigma2 <- sum( W^2 * (( y - theta )^2 - V ) ) / sum( W^2 )
    B <- V / ( V + sigma2 )
    effects <- ( 1 - B ) * y + B * theta
    se.theta <- 1.0 / sqrt( sum( 1.0 / ( V + sigma2 ) ) )
    se.effects <- sqrt( V * ( 1.0 - B ) )
    print( c( i, theta, se.theta, sigma2 ) )
    print( cbind( W, ( W / sum( W ) ), B, y, effects, se.effects ) )
  }
}
```

With the aspirin data it takes **18 iterations** (less than 0.1 second on a 400MHz UltraSPARC Unix machine) to get convergence to **4-digit accuracy**, leading to the summaries in Table 5.3 and the following estimates (standard errors in parentheses):

$$\hat{\theta} = 1.45 \ (0.809), \quad \hat{\sigma^2} = 1.53.$$  

Table 5.3. Maximum likelihood empirical Bayes meta-analysis of the aspirin data.

<table>
<thead>
<tr>
<th>study(i)</th>
<th>$\bar{W}_i$</th>
<th>normalized $\bar{W}_i$</th>
<th>$\bar{B}_i$</th>
<th>$y_i$</th>
<th>$\hat{\theta}_i$</th>
<th>$\hat{SE}(\hat{\theta}_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.235</td>
<td>0.154</td>
<td>0.640</td>
<td>2.77</td>
<td>1.92</td>
<td>0.990</td>
</tr>
<tr>
<td>2</td>
<td>0.308</td>
<td>0.202</td>
<td>0.529</td>
<td>2.50</td>
<td>1.94</td>
<td>0.899</td>
</tr>
<tr>
<td>3</td>
<td>0.143</td>
<td>0.0934</td>
<td>0.782</td>
<td>1.84</td>
<td>1.53</td>
<td>1.09</td>
</tr>
<tr>
<td>4</td>
<td>0.232</td>
<td>0.151</td>
<td>0.646</td>
<td>2.56</td>
<td>1.84</td>
<td>0.994</td>
</tr>
<tr>
<td>5</td>
<td>0.183</td>
<td>0.120</td>
<td>0.719</td>
<td>2.31</td>
<td>1.69</td>
<td>1.05</td>
</tr>
<tr>
<td>6</td>
<td>0.427</td>
<td>0.280</td>
<td>0.346</td>
<td>-1.15</td>
<td>-0.252</td>
<td>0.728</td>
</tr>
</tbody>
</table>
Aspirin Meta-Analysis: Conclusions

Note that (1) AMIS gets **much less weight** (normalized \( \hat{W}_i \)) than would have been expected given its small \( V_i \); (2) the **shrinkage factors** (\( \hat{\beta}_i \)) are considerable, with AMIS shrunk almost all the way into positive territory (see Figure 5.1); (3) there is **considerable study-level heterogeneity** (\( \hat{\sigma} \doteq 1.24 \) percentage points of mortality); and (4) the standard errors of the effects are by and large smaller than the \( \sqrt{V_i} \) (from the **borrowing of strength**) but are still considerable.

**Figure 5.1.** **Shrinkage plot** for the aspirin MLEB meta-analysis.

The **95% interval estimate** of \( \theta \), the overall underlying effect of aspirin on mortality, from this approach comes out

\[
\hat{\theta} \pm 1.96 \cdot \widehat{SE}(\hat{\theta}) = (-0.140, 3.03),
\]

which if **interpreted Bayesianly** gives

\[
P(\text{aspirin reduces mortality|data}) \doteq 1 - \Phi\left(\frac{0 - 1.45}{0.809}\right) = 0.96,
\]

where \( \Phi \) is the **standard normal CDF**.

Thus although the interval includes 0, so that in a frequentist sense the effect is not statistically significant, **in fact from a Bayesian point of view the evidence is running strongly in favor of aspirin’s usefulness.**
Educational Meta-Analysis

In many cases (as with this example) empirical Bayes methods have the advantage of yielding closed-form solutions, but I view them at best as approximations to fully Bayesian analyses—which can in any case be carried out with MCMC—so I won’t have any more to say about EB methods here (see Carlin and Louis, 1996, for more on this topic).

5.3 Incorporating Study-Level Covariates

Case Study: Meta-analysis of the effect of teacher expectancy on student IQ (Bryk and Raudenbush, 1992). Do teachers’ expectations influence students’ intellectual development, as measured by IQ scores?

Table 5.4. Results from 19 experiments estimating the effects of teacher expectancy on pupil IQ.

<table>
<thead>
<tr>
<th>Study (i)</th>
<th>Weeks of Prior Contact ($x_i$)</th>
<th>Estimated Effect Size ($y_i$)</th>
<th>Standard Error of $y_i = \sqrt{v_i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Rosenthal et al. (1974)</td>
<td>2</td>
<td>0.03</td>
<td>0.125</td>
</tr>
<tr>
<td>2. Conn et al. (1968)</td>
<td>3</td>
<td>0.12</td>
<td>0.147</td>
</tr>
<tr>
<td>3. Jose &amp; Cody (1971)</td>
<td>3</td>
<td>-0.14</td>
<td>0.167</td>
</tr>
<tr>
<td>4. Pellegrini &amp; Hicks (1972)</td>
<td>0</td>
<td>1.18</td>
<td>0.373</td>
</tr>
<tr>
<td>5. Pellegrini &amp; Hicks (1972)</td>
<td>0</td>
<td>0.26</td>
<td>0.369</td>
</tr>
<tr>
<td>6. Evans &amp; Rosenthal (1969)</td>
<td>3</td>
<td>-0.06</td>
<td>0.103</td>
</tr>
<tr>
<td>7. Fielder et al. (1971)</td>
<td>3</td>
<td>-0.02</td>
<td>0.103</td>
</tr>
<tr>
<td>8. Claiborn (1969)</td>
<td>3</td>
<td>-0.32</td>
<td>0.220</td>
</tr>
<tr>
<td>9. Kester &amp; Letchworth (1972)</td>
<td>0</td>
<td>0.27</td>
<td>0.164</td>
</tr>
<tr>
<td>10. Maxwell (1970)</td>
<td>1</td>
<td>0.80</td>
<td>0.251</td>
</tr>
<tr>
<td>11. Carter (1970)</td>
<td>0</td>
<td>0.54</td>
<td>0.302</td>
</tr>
<tr>
<td>12. Flowers (1966)</td>
<td>0</td>
<td>0.18</td>
<td>0.223</td>
</tr>
<tr>
<td>13. Keshock (1970)</td>
<td>1</td>
<td>-0.02</td>
<td>0.289</td>
</tr>
<tr>
<td>14. Henrickson (1970)</td>
<td>2</td>
<td>0.23</td>
<td>0.290</td>
</tr>
<tr>
<td>15. Fine (1972)</td>
<td>3</td>
<td>-0.18</td>
<td>0.159</td>
</tr>
<tr>
<td>16. Greiger (1970)</td>
<td>3</td>
<td>-0.06</td>
<td>0.167</td>
</tr>
<tr>
<td>17. Rosenthal &amp; Jacobson (1968)</td>
<td>1</td>
<td>0.30</td>
<td>0.139</td>
</tr>
<tr>
<td>18. Fleming &amp; Anttonen (1971)</td>
<td>2</td>
<td>0.07</td>
<td>0.094</td>
</tr>
<tr>
<td>19. Ginsburg (1970)</td>
<td>3</td>
<td>-0.07</td>
<td>0.174</td>
</tr>
</tbody>
</table>
Teacher Expectancy

Raudenbush (1984) found \( k = 19 \) experiments, published between 1966 and 1974, estimating the effect of teacher expectancy on student IQ (Table 5.4).

In each case the experimental group was made up of children for whom teachers were (deceptively) encouraged to have high expectations (e.g., experimenters gave treatment teachers lists of students, actually chosen at random, who allegedly displayed dramatic potential for intellectual growth), and the controls were students about whom no particular expectations were encouraged.

The estimated effect sizes \( y_i = \frac{T_i - \bar{C}_i}{\text{SD}_{i, \text{pooled}}} \) (column 3 in Table 5.4) ranged from \(-0.32\) to \(+1.18\); why?

One good reason: the studies differed in how well the experimental teachers knew their students at the time they were given the deceptive information—this time period \( x_i \) (column 2 in Table 5.4) ranged from 0 to 3 weeks.

Figure 5.2 plots \( y_i \) against \( x_i \)—you can see that the studies with bigger \( x_i \) had smaller IQ effects on average.

\[ y_i = \frac{T_i - \bar{C}_i}{\text{SD}_{i, \text{pooled}}} \]

\[ x_i \]

\[ \text{Weeks of Prior Contact} \]

\[ \text{Estimated Effect Size} \]

\[ 0.0 \]

\[ 0.5 \]

\[ 1.0 \]

\[ 0 \]

\[ 1 \]

\[ 2 \]

\[ 3 \]

Figure 5.2. Scatterplot of estimated effect size against weeks of prior contact in the IQ meta-analysis. Radii of circles are proportional to \( w_i = V_i^{-1} \) (see column 4 in Table 5.4); fitted line is from weighted regression of \( y_i \) on \( x_i \) with weights \( w_i \).
Conditional Exchangeability

Evidently model (1) will not do here—it says that your predictive uncertainty about all the studies is exchangeable (similar, i.e., according to (1) the underlying study-level effects \( \theta_i \) are like IID draws from a normal distribution), whereas Figure 5.2 clearly shows that the \( x_i \) are useful in predicting the \( y_i \).

This is another way to say that your uncertainty about the studies is not unconditionally exchangeable but conditionally exchangeable given \( x \)

(Draper et al., 1993b).

In fact Figure 5.2 suggests that the \( y_i \) (and therefore the \( \theta_i \)) are related linearly to the \( x_i \).

Bryk and Raudenbush, working in the frequentist paradigm, fit the following HM to these data:

\[
(\theta_i|\alpha, \beta, \sigma^2_{\theta}) \stackrel{\text{indep}}{\sim} N(\alpha + \beta x_i, \sigma^2_{\theta}) \quad \text{(underlying effects)}
\]

\[
(y_i|\theta_i) \stackrel{\text{indep}}{\sim} N(\theta_i, V_i) \quad \text{(data)}. \quad (9)
\]

According to this model the estimated effect sizes \( y_i \) are like draws from a Gaussian with mean \( \theta_i \) and variance \( V_i \), the squared standard errors from column 4 of Table 5.4—here as in model (1) the \( V_i \) are taken to be known—and the \( \theta_i \) themselves are like draws from a Gaussian with mean \( \alpha + \beta x_i \) and variance \( \sigma^2_{\theta} \).

The top level of this HM in effect assumes, e.g., that the 5 studies with \( x = 0 \) are sampled representatively from \{all possible studies with \( x = 0 \}\}, and similarly for the other values of \( x \).

This (and the Gaussian choice on the top level) are conventional assumptions, not automatically scientifically reasonable—for example, if you know of some way in which (say) two of the studies with \( x = 3 \) differ from each other that’s relevant to the outcome of interest, then you should include this in the model as a study-level covariate along with \( x \).
An MLEB Drawback

Bryk and Raudenbush used MLEB methods, based on the EM algorithm, to fit this model.

As in Section 5.2, this estimation method combines the two levels of model (9) to construct a single likelihood for the $y_i$, and then maximizes this likelihood as usual in the ML approach.

They obtained $(\hat{\alpha}, \hat{\beta}) = (.407 \pm .087, -.157 \pm .036)$ and $\hat{\sigma}_\theta^2 = 0$, naively indicating that all of the study-level variability has been “explained” by the covariate $x$.

However, from a Bayesian point of view, this model is missing a third layer:

$$(\alpha, \beta, \sigma_\theta^2) \sim p(\alpha, \beta, \sigma_\theta^2)$$

$$(\theta_i | \alpha, \beta, \sigma_\theta^2) \overset{\text{IID}}{\sim} N(\alpha + \beta(x_i - \bar{x}), \sigma_\theta^2)$$

$$(y_i | \theta_i) \overset{\text{indep}}{\sim} N(\theta_i, V_i).$$

(it will help convergence of the sampling-based MCMC methods to make $\alpha$ and $\beta$ uncorrelated by centering the $x_i$ at 0 rather than at $\bar{x}$).

As will subsequently become clear, the trouble with MLEB is that in Bayesian language it assumes in effect that the posterior for $\sigma_\theta^2$ is point-mass on the MLE. This is bad (e.g., Morris, 1983) for two reasons:

• If the posterior for $\sigma_\theta^2$ is highly skewed, the mode will be a poor summary; and

• Whatever point-summary you use, pretending the posterior SD for $\sigma^2$ is zero fails to propagate uncertainty about $\sigma_\theta^2$ through to uncertainty about $\alpha, \beta$, and the $\theta_i$.

The best way to carry out a fully Bayesian analysis of model (10) is with MCMC methods.
WinBUGS Implementation

For $p(\alpha, \beta, \sigma^2_\theta)$ in model (10) I've chosen the usual WinBUGS **diffuse prior** $p(\alpha)p(\beta)p(\sigma^2_\theta)$: since $\alpha$ and $\beta$ live on the whole real line I've taken marginal Gaussian priors for them with mean 0 and precision $10^{-6}$, and since $\tau_\theta = \frac{1}{\sigma^2}$ is positive I use a $\Gamma(0.001, 0.001)$ prior for it.

Model (10) treats the variances $V_i$ of the $y_i$ as **known** (and equal to the squares of column 4 in Table 5.4); I've converted these into precisions in the data file (e.g., $\tau_1 = \frac{1}{0.125^2} = 64.0$).
WinBUGS Implementation

A burn-in of 1,000 (certainly longer than necessary) from default initial values \((\alpha, \beta, \tau_\theta) = (0.0, 0.0, 1.0)\) and a monitoring run of 10,000 yield the following preliminary MCMC results.

Because this is a random-effects model we don’t expect anything like IID mixing: the output for \(\alpha\) behaves like an AR\(_1\) time series with \(\hat{\rho}_1 = 0.86\).

The posterior mean for \(\alpha, 0.135\) (with an MCSE of 0.002), shows that \(\alpha\) in model (10) and \(\alpha\) in model (9) are not comparable because of the recentering of the predictor \(x\) in model (10): the MLE of \(\alpha\) in (9) was \(0.41 \pm 0.09\).
WinBUGS Implementation

But $\beta$ means the same thing in both models (9) and (10): its posterior mean in (10) is $-0.161 \pm 0.002$, which is not far from the MLE $-0.157$.

Note, however, that the posterior SD for $\beta$, 0.0396, is 10% larger than the standard error of the maximum likelihood estimate of $\beta$ (0.036).

This is a reflection of the underpropagation of uncertainty about $\sigma_\theta$ in maximum likelihood mentioned on page 15.
In these preliminary results $\sigma_\theta$ has posterior mean $0.064 \pm 0.002$ and SD $0.036$, providing **clear evidence** that the MLE $\hat{\sigma}_\theta = 0$ is a **poor summary**.

Note, however, that the likelihood for $\sigma_\theta$ may be **appreciable in the vicinity of 0** in this case, meaning that some **sensitivity analysis** with diffuse priors other than $\Gamma(0.001, 0.001)$—such as $U(0, c)$ for $c$ around 0.5—would be in order.
WinBUGS Implementation

When you specify node theta in the Sample Monitor Tool and then look at the results, you see that WinBUGS presents **parallel findings with a single click** for all elements of the vector \( \theta \).

Some of the \( \theta_i \) are evidently mixing better than others.
WinBUGS Implementation

The marginal density traces of the $\theta_i$ look rather like $t$ distributions with fairly low degrees of freedom (fairly heavy tails).
WinBUGS Implementation

Many of the $\theta_i$ have **posterior probability concentrated near 0**, but not all; $\theta_4, \theta_5, \theta_9, \theta_{11}$, and $\theta_{12}$ are particularly large (looking back on page 12, what's **special** about the corresponding studies?).
WinBUGS Implementation

Some of the $\theta_i$ are not far from white noise; others are mixing quite slowly.
It’s also useful to monitor the $\mu_i = \alpha + \beta(x_i - \bar{x})$, because they represent an important part of the shrinkage story with model (10).
5.3.1 Shrinkage Estimation

In a manner parallel to the situation with the simpler model (1), the posterior means of the underlying study effects $\theta_i$ should be at least approximately related to the raw effect sizes $y_i$ and the $\mu_i$ via the shrinkage equation

$$E(\theta_i|y) = (1 - \hat{B}_i) y_i + \hat{B}_i E(\mu_i|y);$$  \hspace{1cm} (11)

here $\hat{B}_i = \frac{V_i}{V_i + \hat{\sigma}_\theta^2}$ and $\hat{\sigma}_\theta^2$ is the posterior mean of $\sigma^2_\theta$.

This is easy to check in R:

```r
> mu <- c( 0.09231, -0.06898, -0.06898, 0.4149, 0.4149, -0.06898, -0.06898, 
          -0.06898, 0.4149, 0.2536, 0.4149, 0.4149, 0.2536, 0.09231, -0.06898, 
          -0.06898, 0.2536, 0.09231, -0.06898 )

> y <- c( 0.03, 0.12, -0.14, 1.18, 0.26, -0.06, -0.02, -0.32, 0.27, 0.80, 
          0.54, 0.18, -0.02, 0.23, -0.18, -0.06, 0.30, 0.07, -0.07 )

> theta <- c( 0.08144, -0.03455, -0.07456, 0.4377, 0.4076, -0.0628, 
             -0.05262, -0.08468, 0.3934, 0.289, 0.4196, 0.3938, 0.2393, 0.1014, 
             -0.08049, -0.06335, 0.2608, 0.08756, -0.06477 )

> V <- 1 / tau

> B.hat <- V / ( V + 0.064^2 )

> theta.approx <- ( 1 - B.hat ) * y + B.hat * mu
```
The Shrinkage Story (continued)

> cbind( y, theta, mu, sigma.2, B.hat, theta.approx )

<table>
<thead>
<tr>
<th>y</th>
<th>theta</th>
<th>mu</th>
<th>V</th>
<th>B.hat</th>
<th>theta.approx</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>0.03</td>
<td>0.08144</td>
<td>0.09231</td>
<td>0.015625</td>
<td>0.7923026</td>
</tr>
<tr>
<td>[2,]</td>
<td>0.12</td>
<td>-0.03455</td>
<td>-0.06898</td>
<td>0.021609</td>
<td>0.8406536</td>
</tr>
<tr>
<td>[3,]</td>
<td>-0.14</td>
<td>-0.07456</td>
<td>-0.06898</td>
<td>0.027889</td>
<td>0.8719400</td>
</tr>
<tr>
<td>[4,]</td>
<td>1.18</td>
<td>0.43770</td>
<td>0.41490</td>
<td>0.139129</td>
<td>0.9714016</td>
</tr>
<tr>
<td>[5,]</td>
<td>0.26</td>
<td>0.40760</td>
<td>0.41490</td>
<td>0.136161</td>
<td>0.9707965</td>
</tr>
<tr>
<td>[6,]</td>
<td>-0.06</td>
<td>-0.06280</td>
<td>-0.06898</td>
<td>0.010609</td>
<td>0.7214553</td>
</tr>
<tr>
<td>[7,]</td>
<td>-0.02</td>
<td>-0.05262</td>
<td>-0.06898</td>
<td>0.010609</td>
<td>0.7214553</td>
</tr>
<tr>
<td>[8,]</td>
<td>-0.32</td>
<td>-0.08468</td>
<td>-0.06898</td>
<td>0.048400</td>
<td>0.9219750</td>
</tr>
<tr>
<td>[9,]</td>
<td>0.27</td>
<td>0.39340</td>
<td>0.41490</td>
<td>0.026896</td>
<td>0.8678369</td>
</tr>
<tr>
<td>[10,]</td>
<td>0.80</td>
<td>0.28900</td>
<td>0.25360</td>
<td>0.063001</td>
<td>0.9389541</td>
</tr>
<tr>
<td>[11,]</td>
<td>0.54</td>
<td>0.41960</td>
<td>0.41490</td>
<td>0.091204</td>
<td>0.9570199</td>
</tr>
<tr>
<td>[12,]</td>
<td>0.18</td>
<td>0.39380</td>
<td>0.41490</td>
<td>0.049729</td>
<td>0.9239015</td>
</tr>
<tr>
<td>[13,]</td>
<td>-0.02</td>
<td>0.23930</td>
<td>0.25360</td>
<td>0.083521</td>
<td>0.9532511</td>
</tr>
<tr>
<td>[14,]</td>
<td>0.23</td>
<td>0.10140</td>
<td>0.09231</td>
<td>0.084100</td>
<td>0.9535580</td>
</tr>
<tr>
<td>[15,]</td>
<td>-0.18</td>
<td>-0.08049</td>
<td>-0.06898</td>
<td>0.025281</td>
<td>0.8605712</td>
</tr>
<tr>
<td>[16,]</td>
<td>-0.06</td>
<td>-0.06335</td>
<td>-0.06898</td>
<td>0.027889</td>
<td>0.8719400</td>
</tr>
<tr>
<td>[17,]</td>
<td>0.30</td>
<td>0.26080</td>
<td>0.25360</td>
<td>0.019321</td>
<td>0.8250843</td>
</tr>
<tr>
<td>[18,]</td>
<td>0.07</td>
<td>0.08756</td>
<td>0.09231</td>
<td>0.008836</td>
<td>0.6832663</td>
</tr>
<tr>
<td>[19,]</td>
<td>-0.07</td>
<td>-0.06477</td>
<td>-0.06898</td>
<td>0.030276</td>
<td>0.8808332</td>
</tr>
</tbody>
</table>

You can see that equation (11) is indeed a good approximation to what’s going on: the posterior means of the $\theta_i$ (column 3 of this table, counting the leftmost column of study indices) all fall between the $y_i$ (column 2) and the posterior means of the $\mu_i$ (column 4), with the closeness to $y_i$ or $E(\mu_i|y)$ expressed through the shrinkage factor $\hat{B}_i$.

Since $\hat{\sigma}_\theta^2$ is small (i.e., most—but not quite all—of the between-study variation has been explained by the covariate $x$), the raw $y_i$ values are shrunken almost all of the way toward the regression line $\alpha + \beta(x_i - \bar{x})$. 

25
5.4 References


