Statistical methods for linguistic research: Advanced Tools

Shravan Vasishth

Department of Linguistics
University of Potsdam, Germany

August 9, 2015

Today's goals

In this lecture, my goal is to

1. present two simple examples of how linear mixed models can be fit in JAGS.
2. show you how the same models could be fit in Stan
3. show you how to scale up to more complex models.

Bayesian LMM

Step 1: Set up data

Set up data for JAGS (and Stan). The data must be a list containing vectors.

Bayesian LMM

Step 1: Set up data

```
## data for JAGS and Stan:
headnoun.dat <- list(subj=
                  as.integer(factor(headnoun$subj) ),
                       item=as.integer(factor(headnoun$item)),
                      rrt = headnoun$rrt,
                      x = headnoun$x,
                      I = nrow(headnoun),
                      J =length( unique(headnoun$subj) ),
                      K =length( unique(headnoun$item)))
```

Bayesian LMM

Step 2: Define model

1. We literally write out the model that is assumed to have generated the data:

$$\mathrm{RT}_i = \beta_0 + u_{0j} + w_{0k} + (\beta_1 + u_{1j} + w_{1k})\mathrm{x}_i + \varepsilon_i \tag{1}$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_u\right) \quad \begin{pmatrix} w_{0k} \\ w_{1k} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_w\right) \tag{2}$$

$$\varepsilon_i \sim N(0, \sigma^2) \tag{3}$$

2. We will also need to define priors for the parameters $\beta_0, \beta_1, \Sigma_u, \Sigma_w, \sigma$.

Variance vs Precision

As discussed earlier, in JAGS, instead of variance, we talk about precision, which is the **inverse** of variance. So we can write the variance components as follows.

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Omega_u\right) \quad \begin{pmatrix} w_{0k} \\ w_{1k} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Omega_w\right) \tag{4}$$

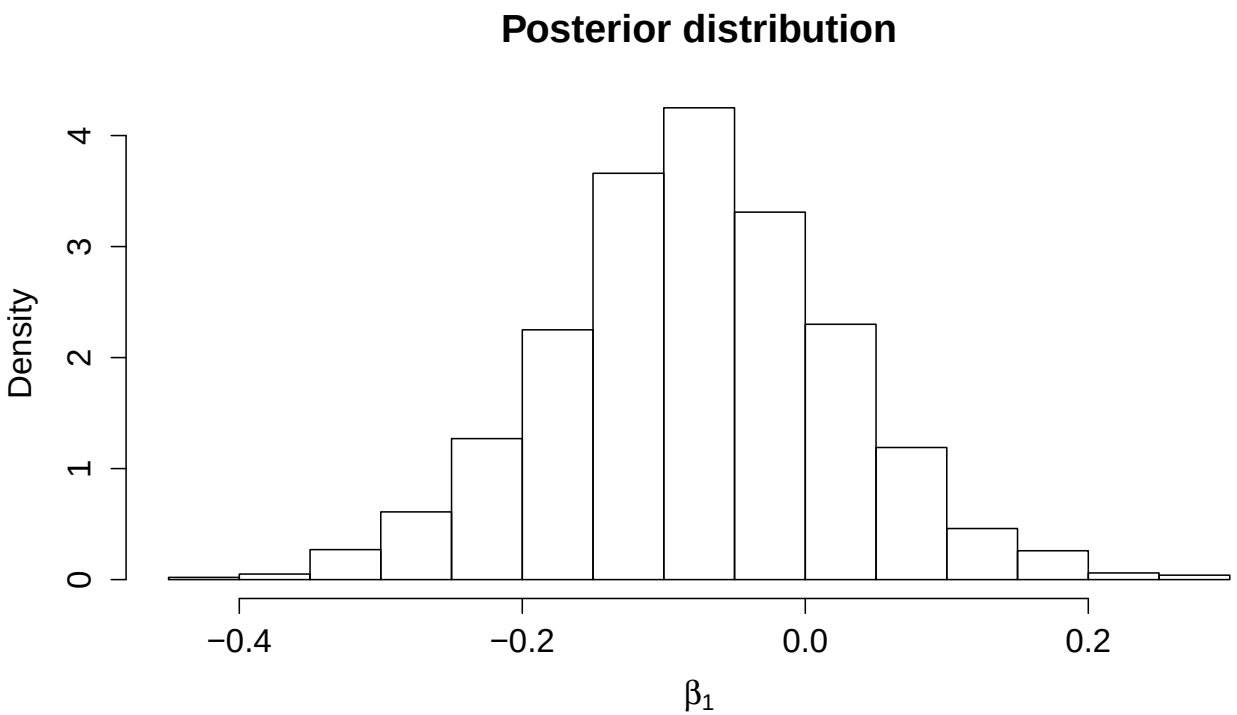
$$\varepsilon_i \sim N(0, \tau^2) \tag{5}$$

Here, $\Omega_u = \Sigma_u^{-1}$, $\Omega_w = \Sigma_w^{-1}$, and $\tau = \frac{1}{\sigma^2}$.
Σ_u^{-1} is the **inverse** of Σ_u, and yields a precision matrix. We will define priors on the precision matrix rather than the variance-covariance matrix.

Bayesian LMM

Looking ahead

Our goal will be to determine the **posterior distribution** of β_1, which is the estimate of the effect of relative clause type.
Gibson and Wu expect β_1 to be negative and significantly different from 0. To anticipate the result, using a uniform prior for β_1, what we will get (in the **reciprocal rt** scale) is:



Bayesian LMM

Step 2: Define model

First, write out how the data are assumed to be generated.

$$\mu_i = \beta_0 + u_{0j} + w_{0k} + (\beta_1 + u_{1j} + w_{1k})x_i \tag{6}$$

$$rrt_i \sim N(\mu_i, \sigma_e^2) \tag{7}$$

```
    # Define model for each observational unit
    for( i in 1:N )
{
    mu[i] <- ( beta[1] + u[subj[i],1] + w[item[i],1])
+ ( beta[2] + u[subj[i],2] + w[item[i],2]) * ( x[i] )
    rrt[i] ~ dnorm( mu[i], tau.e )
}
```

Bayesian LMM

Step 2: Define model

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_u \right) \quad \Omega_u = \Sigma_u^{-1} \tag{8}$$

```
data
{    zero[1] <- 0
    zero[2] <- 0
}
    # Intercept and slope for each subj
    for( j in 1:J )
{
    u[j,1:2] ~ dmnorm(zero,Omega.u)
}
```

Bayesian LMM

Step 2: Define model

$$\begin{pmatrix} w_{0k} \\ w_{1k} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_w \right) \quad \Omega_w = \Sigma_w^{-1} \tag{9}$$

```
    # Intercept and slope for each item
    for( k in 1:K )
{
    w[k,1:2] ~ dmnorm(zero,Omega.w)
}
```

Bayesian LMM

Step 2: Define model (priors for fixed effect coefficients)

$$\beta_0 \sim N(\mu = 0, \sigma^2 = 1.0 \times 10^5) \quad \beta_1 \sim N(\mu = 0, \sigma^2 = 1.0 \times 10^5) \tag{10}$$

Recall that in JAGS σ^2 is expressed as precision $\tau = \frac{1}{\sigma^2}$:

```
# Priors:
# Fixed intercept and slope (weakly informative)
beta[1] ~ dnorm(0.0,1.0E-5)
beta[2] ~ dnorm(0.0,1.0E-5)
```

These priors express a belief that the β are likely to be centered around 0 (Note: not reasonable for β_0), but that we are very unsure about this.

Bayesian LMM

Step 2: Define model (priors for variance components)

$$\sigma^2 \sim Uniform(0, 100) \tag{11}$$

```
# Residual variance
tau.e <- sigma.e^(-2)
sigma.e  ~ dunif(0,100)
```

Note: in JAGS, another way to write sigma.e to the power of -2 is

```
pow(sigma.e,-2)
```

Bayesian LMM

Step 2: Define model (priors for variance components)

- Σ_u and Σ_w can be expressed as precision matrices by inverting them: $\Sigma_u^{-1} = \Omega_u$ and $\Sigma_w^{-1} = \Omega_w$.
- We will define a Wishart distribution as a prior for Ω_u and Ω_w.
- The Wishart is the multivariate version of the gamma distribution and is a reasonable prior for precision matrices (see references at the end of these slides for more details).
- The prior will be Wishart(R,2), where R is an initial guess at a variance-covariance matrix, and 2 is the number of dimensions of the matrix:
  $\Omega_u \sim Wishart(R_u, 2)$
  $\Omega_w \sim Wishart(R_w, 2)$

Bayesian LMM

Step 2: Define model (priors for variance components)

The steps for defining the prior for the precision matrix are:

1. State that $\Omega_u \sim Wishart(R_u, 2)$
2. Create R_u by filling in each cell.
3. Define priors for each parameter used to build u R_u:

$$R_u = \begin{bmatrix} \sigma^2_{\text{u0}} & \rho_u \sigma_{u0} \sigma_{u1} \\ \rho_u \sigma_{u0} \sigma_{u1} & \sigma^2_{u1} \end{bmatrix} \tag{12}$$

1. $\sigma_{\text{u0}} \sim Uniform(0, 10)$
2. $\sigma_{\text{u1}} \sim Uniform(0, 10)$
3. $\rho_u \sim Uniform(-1, 1)$.

Bayesian LMM

Step 2: Define model (priors for variance components)

```
## Prior on precision:
Omega.u  ~ dwish( R.u, 2 )
## Fill in R matrix:
R.u[1,1] <- sigma.a^2
R.u[2,2] <- sigma.b^2
R.u[1,2] <- rho.u*sigma.a*sigma.b
R.u[2,1] <- rho.u*sigma.a*sigma.b
## Prior for varying intercepts sd:
sigma.a ~ dunif(0,10)
## prior for varying slopes sd:
sigma.b ~ dunif(0,10)
## prior for correlation:
rho.u ~ dunif(-1,1)
```

Bayesian LMM

Step 2: Define model

See R code accompanying these lectures for full model specification in JAGS.
Also see this tutorial article on Stan (to be discussed later in this course): http://www.ling.uni-potsdam.de/~vasishth/statistics/BayesLMMs.html

Bayesian LMM

Step 3: Fit model

Decide which variables you want to track the posterior distribution of.

```
track.variables<-c("beta","sigma.e",
                   "sigma.a","sigma.b",
                   "sigma.c","sigma.d",
                   "rho.u","rho.w")
```

```
library(rjags)

## Linked to JAGS 3.4.0
## Loaded modules:  basemod,bugs
```

Bayesian LMM

Step 3: Fit model

```
headnoun.mod <- jags.model(
  file="gwmaximal.jag",
  data = headnoun.dat,
  n.chains = 4,
  n.adapt =2000 , quiet=T)
```

Bayesian LMM

Step 4: Generate posterior samples

```
headnoun.res <- coda.samples(headnoun.mod,
                             var = track.variables,
                             n.iter = 10000,
                             thin = 1)
```

Bayesian LMM

Step 4: Generate posterior samples

```
summary(headnoun.res)

##
## Iterations = 2001:12000
## Thinning interval = 1
## Number of chains = 4
## Sample size per chain = 10000
##
## 1. Empirical mean and standard deviation for each variab
##    plus standard error of the mean:
##
##              Mean      SD  Naive SE Time-series SE
## beta[1] -2.66798 0.15240 0.0007620      0.0037267
## beta[2] -0.03842 0.05028 0.0002514      0.0004237
## rho.u   -0.26636 0.50929 0.0025464      0.0120329
```

Bayesian LMM

Step 4: Generate posterior samples

```
## not plotted
#plot(headnoun.res)
```

Frequentist LMM

Comparison with lmer

```
m0<-lmer(rrt~x + (1+x|subj)+(1+x|item),headnoun)
m1<-lmer(rrt~x + (1+x|subj)+(1|item),headnoun)
m2<-lmer(rrt~x + (1|subj)+(1|item),headnoun)
```

Frequentist LMM

Comparison with lmer

```
anova(m1,m2)

## refitting model(s) with ML (instead of REML)

## Data: headnoun
## Models:
## m2: rrt ~ x + (1 | subj) + (1 | item)
## m1: rrt ~ x + (1 + x | subj) + (1 | item)
##    Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## m2  5 1603.5 1625.0 -796.76   1593.5
## m1  7 1605.6 1635.7 -795.78   1591.6 1.9509      2      0.377
```

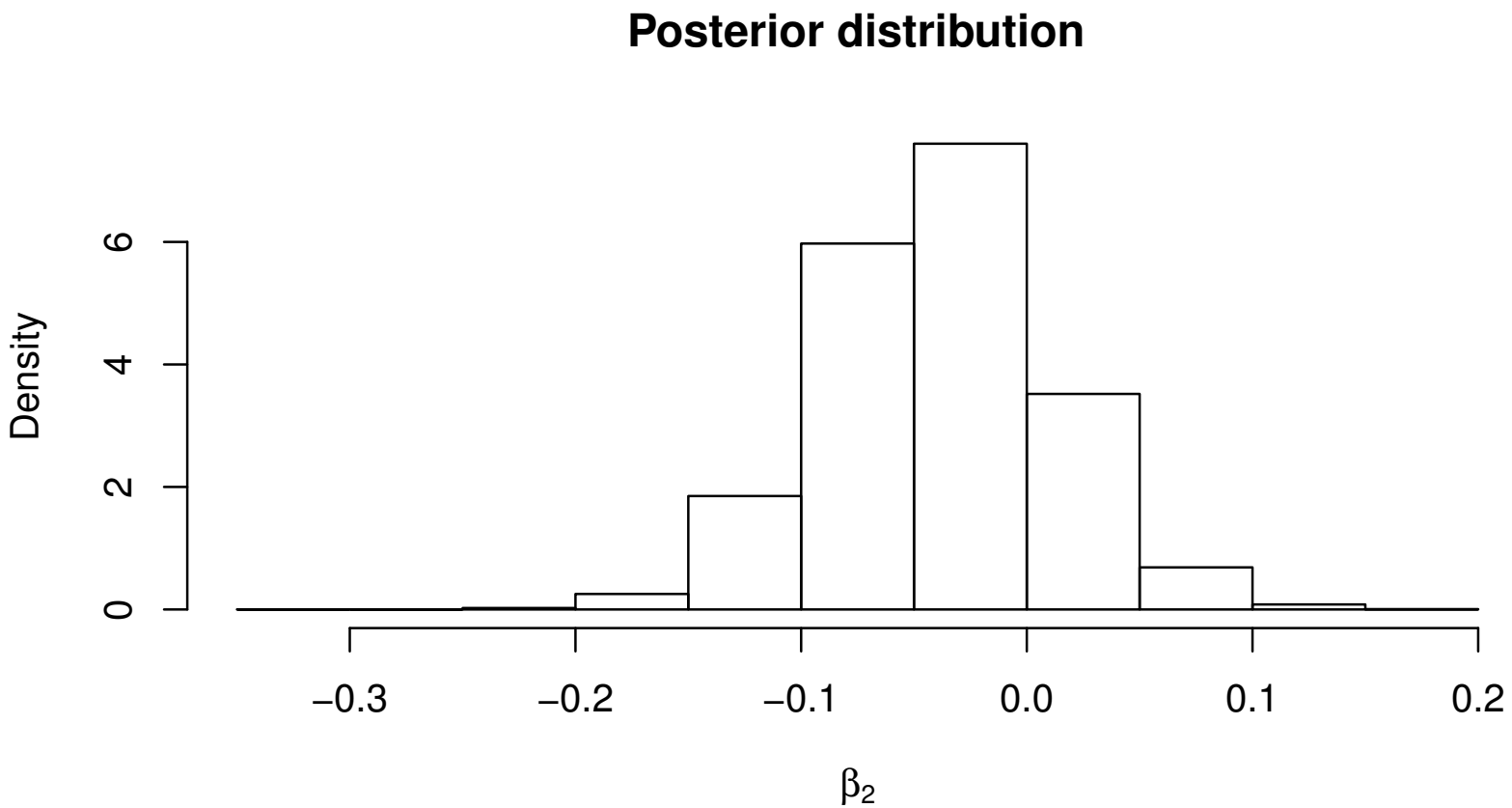
Frequentist LMM

Comparison with lmer

```
summary(m0)$coef[2,1] - 2 * summary(m0)$coef[2,2]

## [1] -0.1316779

summary(m0)$coef[2,1] + 2 * summary(m0)$coef[2,2]

## [1] 0.0540964
```

Bayesian LMM

Step 5: Inference



Bayesian LMM

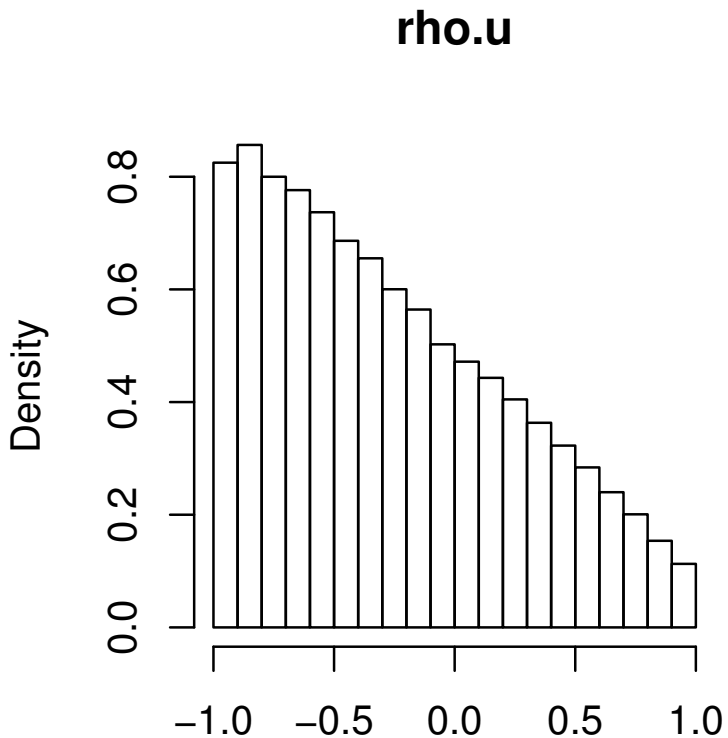
Step 5: Inference

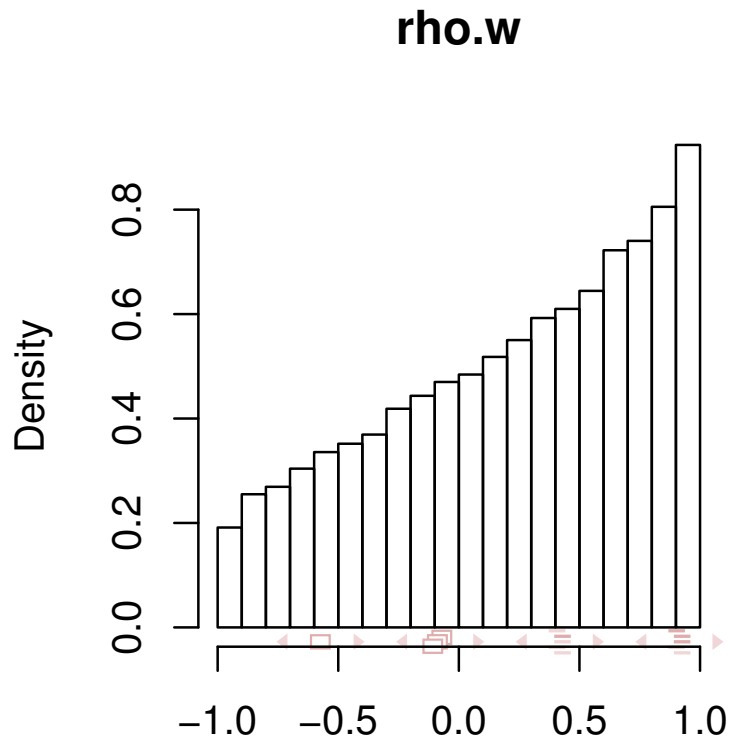
```
mean(mcmcChain[,2]<0)

## [1] 0.78535
```

Bayesian LMM

Step 5: Inference

```
op<-par(mfrow=c(1,2),pty="s")
hist(mcmcChain[,3],freq=F,main="rho.u")
hist(mcmcChain[,4],freq=F,main="rho.w")
```





Bayesian LMM

Step 5: Inference

```
## posterior probability of beta_1 < 0
## given data:
(meanbeta1<-mean(mcmcChain[,2]<0))

## [1] 0.78535
```

The conclusion here is that Gibson and Wu's claim seems to be weakly supported by the data: there is a 0.79 probability of β_1 being less than 0.
Lack of power (here, small sample size) and replicability are still the key issues.

Bayesian LMM

Checking convergence:

- The Gelman-Rubin (or Brooks-Gelman-Rubin) diagnostic involves sampling from multiple chains and then comparing between and within group variability. It's analogous to the F-score in anova.
- Within variance is represented by the mean width of the 95% posterior Credible Intervals (CrI) of all chains, from the final T iterations.
- Between variance is represented by the width of the 95% CrI using all chains pooled together (for the T iterations). If the ratio $\hat{R} = B/W$ is approximately 1, we have convergence.

Bayesian LMM

Checking convergence:

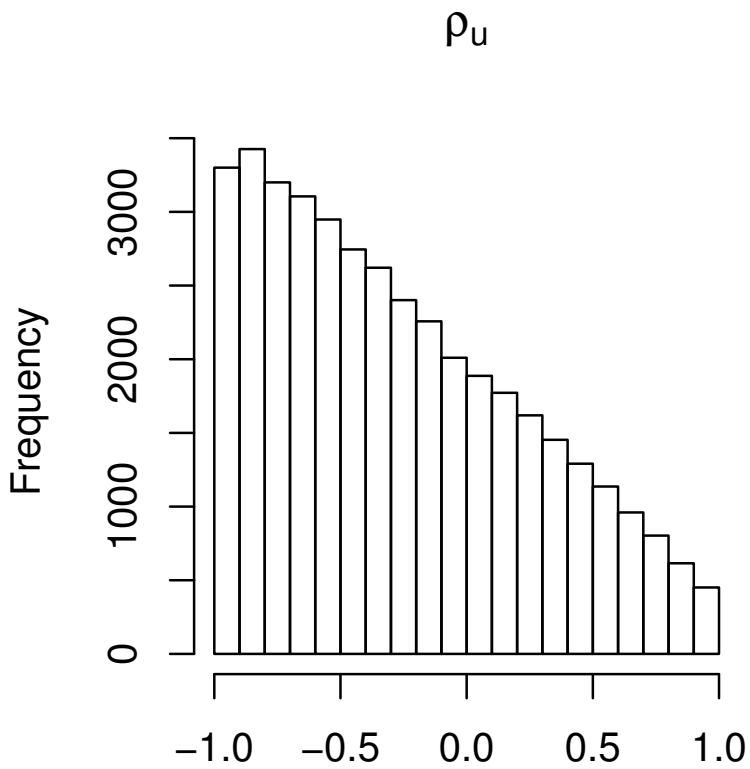
```
gelman.diag(headnoun.res)

## Potential scale reduction factors:
##
##          Point est. Upper C.I.
## beta[1]        1.00       1.00
## beta[2]        1.00       1.00
## rho.u          1.01       1.02
## rho.w          1.00       1.01
## sigma.a        1.00       1.00
## sigma.b        1.02       1.07
## sigma.c        1.00       1.00
## sigma.d        1.01       1.03
## sigma.e        1.00       1.00
##
```

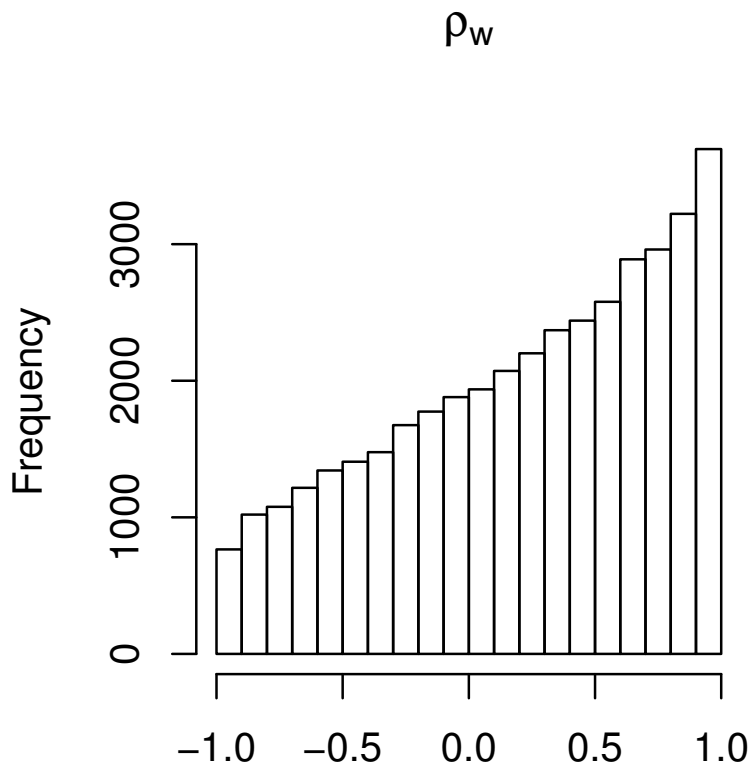
Comparison of lmer and JAGS fit

Parameter estimate	lmer	JAGS
$\hat{\beta}_0$	-2.67 (0.14)	-2.68 (0.13)
$\hat{\beta}_1$	-0.08 (0.10)	-0.08 (0.10)
$\hat{\sigma}_{subj,int}$	0.61	0.78
$\hat{\sigma}_{subj,sl}$	0.23	0.20
$\hat{\rho}_{subj}$	-0.51	-0.09 (0.55)
$\hat{\sigma}_{item,int}$	0.33	0.39
$\hat{\sigma}_{item,sl}$	0.10	0.19
$\hat{\rho}_{item}$	**1.00***	-0.11 (0.58)

* degenerate var-cov matrix, one reason why you should not fit a maximal model here with lmer.

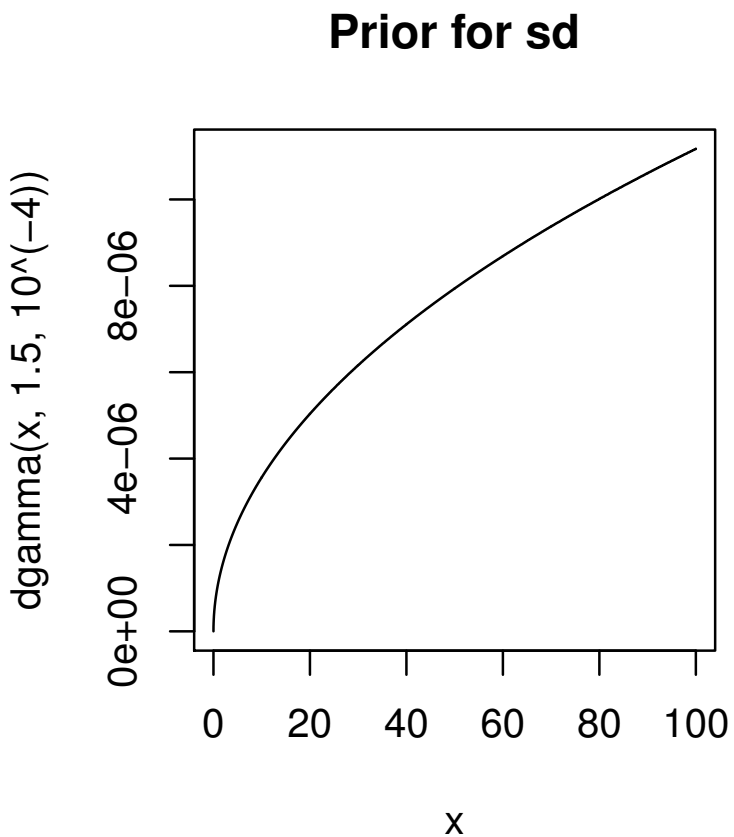
The posterior distributions of the correlations

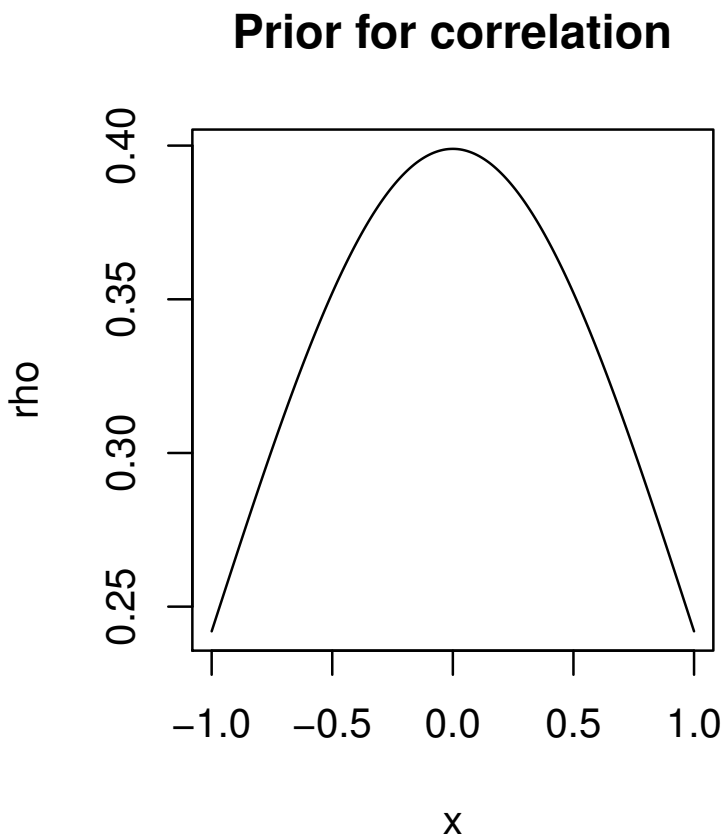




What to do about ±1 correlation estimates?

Suggestion from Chung et al (unpublished MS)





Bayesian LMM

Model with regularization for correlation

The only innovation now is to have more informative priors for correlations.
We write a new model (see gwmaximal2.jag in accompanying R code).

Bayesian LMM

Run model

```
headnoun.mod2 <- jags.model(
  file="gwmaximal2.jag",
  data = headnoun.dat,
  n.chains = 4,
  n.adapt =2000 , quiet=T)
```

Bayesian LMM

Generate posterior samples

```
headnoun.res2 <- coda.samples(headnoun.mod2,
                  var = track.variables,
                  n.iter = 10000,
                  thin = 20)
```

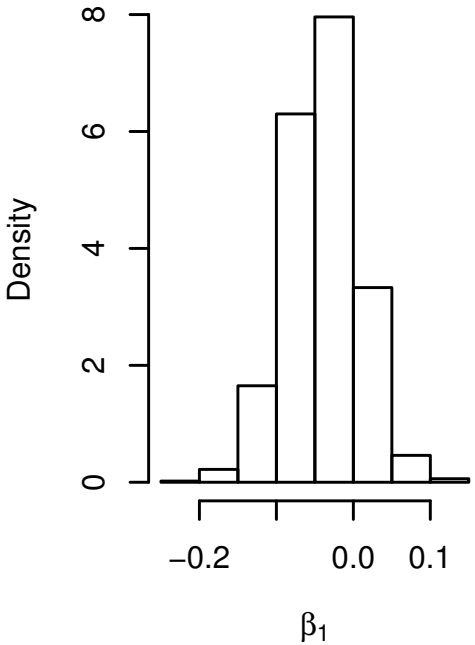
Bayesian LMM

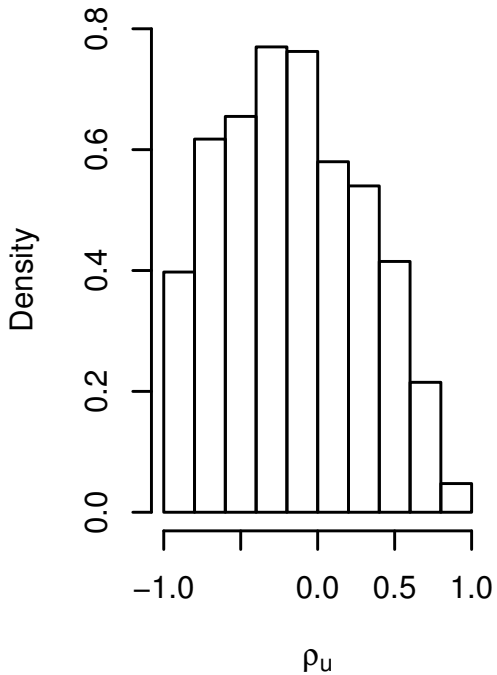
Plot posterior distributions

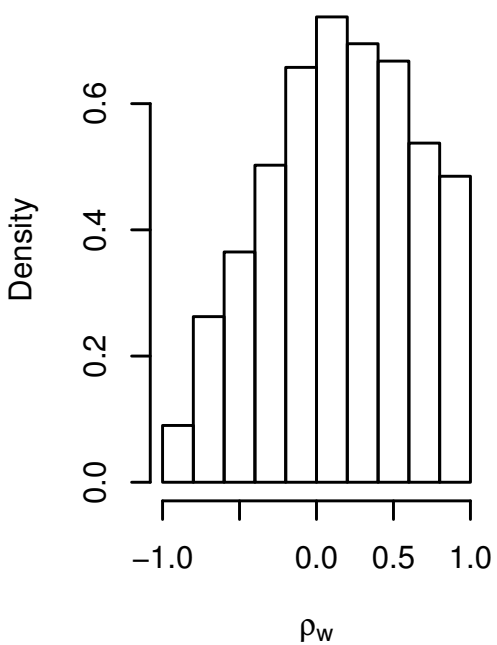
You would need to have a lot of data to shift the posterior for the ρ away from 0, but you could do that, in principle.

Bayesian LMM

Plot posterior distributions

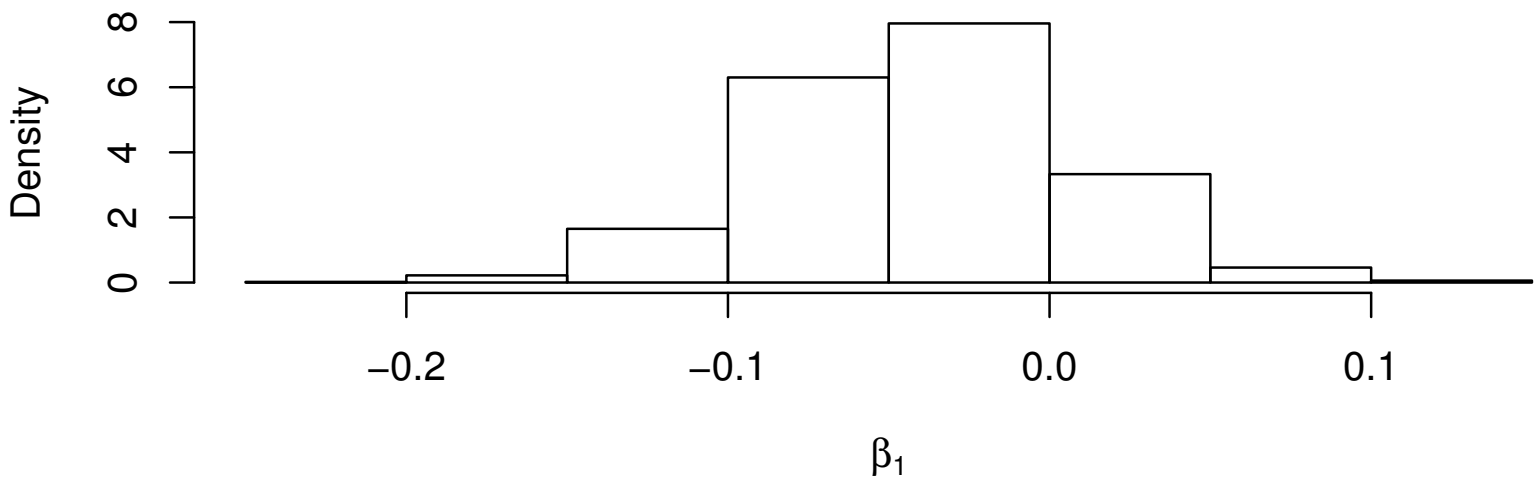






Bayesian LMM

Probability of $\beta_1 < 0$



Bayesian LMM

Probability of $\beta_1 < 0$

```
mean(MCMCchain[,2]<0)

## [1] 0.8075
```

Thus, *given this data-set*, there is some reason to believe that β_1 is less than 0, as predicted by Gibson and Wu.

Comparison of lmer and JAGS fit (model 2)

Parameter estimate	lmer	JAGS
$\hat{\beta}_0$	-2.67 (0.14)	-2.69 (0.11)
$\hat{\beta}_1$	-0.08 (0.10)	-0.09 (0.10)
$\hat{\sigma}_{subj,int}$	0.61	0.02
$\hat{\sigma}_{subj,sl}$	0.23	0.01
$\hat{\rho}_{subj}$	-0.51	-0.02 (0.47)
$\hat{\sigma}_{item,int}$	0.33	0.01
$\hat{\sigma}_{item,sl}$	0.10	0.01
$\hat{\rho}_{item}$	**1.00***	-0.07 (0.49)

* degenerate var-cov matrix

Why ignore prior knowledge?

Suppose (just hypothetically) that you have good reason to believe (based on theory or data) that
$\beta_1 \sim N(\mu = 0.10, \sigma^2 = 0.10^2)$
We can take this prior knowledge into account in the model by simply making this our prior for β_1.
Here, the probability that $\beta_1 < 0$ given the data is only:

```
mean(MCMCchain[,2]<0)

## [1] 0.6375
```

Of course, with enough data, you could in principle shift the posterior distribution in either direction, i.e., change your belief in the face of enough evidence!

Meta-analyses

The controversy about Chinese relative clauses

source	coef.	SE	n	method
Gibson Wu 2012	-123.20	46.84	36	SPR
Vasishth et al 2013 expt 3	-109.40	54.80	40	SPR
Lin et al 2011 expt 1	-100.00	30.00	48	SPR
Lin et al 2011 expt 2	-30.00	32.05	40	SPR
Qiao et al 2012 expt 2	-28.00	23.80	24	LMaze
Qiao et al 2012 expt 1	-16.00	44.26	32	GMaze
Wu et al 2011	50.00	40.00	48	SPR
Hsiao and Gibson 2003	50.00	25.00	35	SPR
Wu et al 2009	50.00	23.00	40	SPR
Jaeger et al 2013 expt 1	55.62	65.14	49	SPR
Chen et al 2008	75.00	35.50	39	SPR
Jaeger et al 2013 expt 2	81.92	36.25	49	ET
Vasishth et al 2013 expt 2	82.60	41.20	61	SPR
Vasishth et al 2013 expt 1	148.50	50.90	60	SPR

Synthesizing the evidence

A Bayesian meta-analysis

1. Let Y_i be the effect size in the i-th study, where i ranges from 1 to k (here, k=14). The unit is milliseconds; a positive sign means a subject relative advantage and a negative sign an object relative advantage.
2. Let d be the underlying effect size, to be estimated by the model.
3. Let v_i^2 be the estimated within-study variance.
4. Then, our model is:

$$Y_i \sim N(\delta_i, v_i^2) \quad i = 1, \ldots, k \tag{13}$$

where

$$\delta_i \sim N(d, \tau^2) \quad i = 1, \ldots, k \tag{14}$$

The variance parameter τ^2 represents between study variance. The prior for $\sqrt{\tau}$ could be a uniform distribution, or in inverse gamma.

Synthesizing the evidence

A Bayesian meta-analysis

Plausible values of the subject/object relative clause advantage can be assumed to range between -300 and 300 ms. But we will assume three different levels of uncertainty: The 95% credible intervals are

1. $(-1.96 \times 100, 1.96 \times 100) = (-196, 196)$;
2. $(-1.96 \times 200, 1.96 \times 200) = (-392, 392)$; and
3. $(-1.96 \times 300, 1.96 \times 300) = (-588, 588)$.

We therefore try three priors for d: $N(0, \sigma^2)$, with $\sigma = 100, 200, 300$. These priors correspond to an agnostic starting point with increasing levels of uncertainty about the range of plausible values for the relative clause processing difference.

Synthesizing the evidence

Analysis with all the data

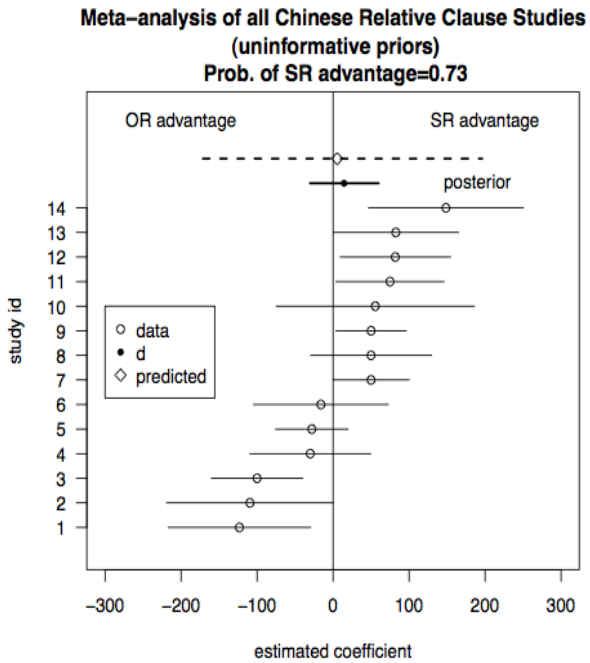


Figure 2: Meta-analysis with all data, $d \sim N(0, 1/300^2)$ and $\tau \sim Ga(0.001, 0.001)$.

Synthesizing the evidence

Analysis using existing data as prior

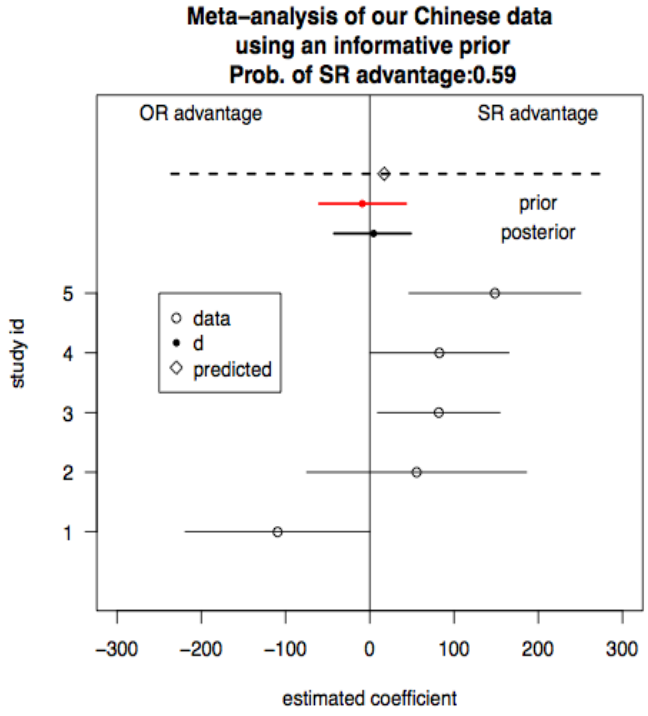


Figure 3: Evaluation of my data using posterior of previous data as my prior.

Synthesizing the evidence

Concluding remarks

Given existing evidence, even believers in the object-relative advantage for Chinese would have to be skeptical about their belief: Prob(Object Relative Advantage | data)=0.41 to 0.27, depending on what prior we have.
Two key advantages of Bayesian LMMs in this example are that

1. We can assign a probability to our belief given the data. **Quantifying uncertainty is the central goal, not a binary reject-accept decision**.
2. We can use prior knowledge in our analyses.

Fitting LMMs of greater complexity, using Stan

I will discuss the following papers if there is time:

1. Sorensen, Hohenstein, Vasishth, Bayesian Linear Mixed Models using Stan: A tutorial for psychologists, linguists, and cognitive scientists
   http://www.ling.uni-potsdam.de/~vasishth/statistics/BayesLMMs.html
2. Bates, Kliegl, Vasishth, Baayen, Parsimonious Mixed Models. ArXiv preprint: http://arxiv.org/abs/1506.04967

Our articles using Stan or JAGS

1. Dario Paape and Shravan Vasishth. Local coherence and preemptive digging-in effects in German. Language and Speech, 2015. Accepted pending minor revisions.
2. Samar Husain, Shravan Vasishth, and Narayanan Srinivasan. Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. Journal of Eye Movement Research, 8(2):1-12, 2015.
3. Stefan L. Frank, Thijs Trompenaars, and Shravan Vasishth. Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? submitted, 2015.
4. Samar Husain, Shravan Vasishth, and Narayanan Srinivasan. Strong Expectations Cancel Locality Effects: Evidence from Hindi. PLoS ONE, 9(7):1-14, 2014.
5. Philip Hofmeister and Shravan Vasishth. Distinctiveness and encoding effects in online sentence comprehension. page n/a, 2014. accepted in Frontiers Special Issue, http://journal.frontiersin.org/ResearchTopic/1545
6. Shravan Vasishth, Zhong Chen, Qiang Li, and Gueilan Guo. Processing Chinese Relative Clauses: Evidence for the Subject-Relative Advantage. PLoS ONE, 8(10):1-14, 10 2013.

Recommended reading

1. Lynch SM (2007) Introduction to applied Bayesian statistics and estimation for social scientists. Springer.
2. Lunn et al. (2012) The BUGS book: A practical introduction to Bayesian analysis. CRC Press.
3. Gelman A, & Hill J (2007) Data analysis using regression and multilevel/hierarchical models. Cambridge, UK: Cambridge University Press.
4. Lee, M.D., & Wagenmakers, E.-J. (2013). Bayesian Cognitive Modeling: A Practical Course. Cambridge University Press. http://faculty.sites.uci.edu/mdlee/bgm/
5. Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). Bayesian data analysis (Vol. 2). London: Chapman & Hall/CRC.

You can also get a lot of help from the JAGS and Stan mailing lists.

In closing

1. Don't be seduced by the illusion that computing p-values $\neq$ doing science.
2. The first goal is to build a reasonable model for the data at hand. Inference is the next step.
3. Aim for high power and replicability above all else.

In closing

1. Don't look for or blindly follow guidelines (except this one!), instead aim for developing understanding of statistical methods.
2. It's not about Bayes vs Frequentist methods; both are useful depending on context. When you have a lot of data, Frequentist methods can be adequate. When you have sparse data, Bayesian methods are very powerful.
3. If you want flexibility in model specification, the Bayesian approach is the way to go.
4. You can of course use Bayesian methods without exception, even for standard models. This is what I do.

In closing

"Do the best experiments you can, and always tell the truth. That's all."

Sydney Brenner