Generalizing the linear mixed model to factorial designs

Tanner Sorensen

University of Potsdam, Potsdam, Germany

Shravan Vasishth

University of Potsdam, Potsdam, Germany, and

School of Mathematics and Statistics, University of Sheffield, Sheffield, UK

July 22, 2015

The Gibson and Wu (2013) data-set has a two-condition design. This section presents a varying intercepts, varying slopes model for a 2×2 factorial design. Because of the more general matrix formulation we use here, the Stan code can be deployed with minimal changes for much more complex designs, including correlational studies.

Our example is the 2×2 repeated measures factorial design of Husain, Vasishth, and Srinivasan (2014, Experiment 1), also a self-paced reading study on relative clauses. The dependent variable was the reading time `rt` of the relative clause verb. The factors were relative clause type, which we code with the predictor `so` (`so` $= +1$ for object relatives and `so` $= -1$ for subject relatives) and distance between the head noun and the relative clause verb, which we code with the predictor `dist` (`dist` $= +1$ for far and `dist` $= -1$ for near). Their interaction is the product of the `dist` and `so` contrast vectors, and labeled as the predictor `int`. The 60 subjects were speakers of Hindi, an Indo-Aryan language spoken primarily in India. The 24 items were presented in a standard, fully balanced Latin square design. This resulted in a total of 1440 data points ($60 \times 24 = 1440$). The first few lines from the data frame are shown below.

The theoretical interest is in determining whether relative clause type and distance

row	subj	item	so	dist	rt
1	1	14	s	n	1561
2	1	16	o	n	959
3	1	15	o	f	582
4	1	18	s	n	294
5	1	4	o	n	438
6	1	17	s	f	286
⋮	⋮	⋮	⋮	⋮	⋮
1440	9	13	s	f	516

Table 1
The first six rows, and the last row, of the data-set of Husain et al. (2014, Experiment 1), as they appear in the data frame.

influence reading time, and whether there is an interaction between these two factors. We use Stan to determine the posterior probability distribution of the fixed effect β_1 for relative clause type, the fixed effect β_2 for distance, and their interaction β_3.

We fit a varying intercepts, varying slopes model to this data-set. The grand mean β_0 of log `rt` is adjusted by subject and by item through the varying intercepts u_0 and w_0, which are unique values for each subject and item respectively. Likewise, the three fixed effects β_1, β_2, and β_3 which are associated with the predictors `so`, `dist`, and `int`, respectively, are adjusted by the by-subject varying slopes u_1, u_2, and u_3 and by-item varying slopes w_1, w_2, and w_3.

It is more convenient to represent this model in matrix form. We build up the model specification by first noting that, for each subject, the by-subject varying intercept u_0 and slopes u_1, u_2, and u_3 have a multivariate normal prior distribution with mean zero and covariance matrix Σ_u. Similarly, for each item, the by-item varying intercept w_0 and slopes w_1, w_2, and w_3 have a multivariate normal prior distribution with mean zero and covariance matrix Σ_w. We can write this as follows:

$$\begin{pmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma_u \right) \quad \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma_w \right) \tag{1}$$

```
rDat<-read.table("HusainEtAlexpt1data.txt",header=TRUE)
rDat$subj <- with(rDat,factor(subj))
rDat$item <- with(rDat,factor(item))

X <- unname(model.matrix(~1+so+dist+int, rDat))

stanDat <- within(list(),
{
  N<-nrow(X)
  P <- n_u <- n_w <- ncol(X)
  X <- X
  Z_u <- X
  Z_w <- X
  J <- length(levels(rDat$subj))
  K <- length(levels(rDat$item))
  rt <- rDat$rt
  subj <- as.integer(rDat$subj)
  item <- as.integer(rDat$item)
}
)
factorialFit <- stan(file="factorialModel.stan",
                     data=stanDat,
                     iter=2000, chains=4)
```

Listing 1: Preparation of data for analyzing the Husain et al. data-set, and running the model.

The error ε is assumed to have a normal distribution with mean zero and standard deviation σ_e.

We proceed to implement the model in Stan. First we read in the data-set (see Listing 1). Instead of passing the predictors `so`, `dist`, and their interaction `int` to `stan` as vectors, as we did with `so` earlier, we make `so`, `dist`, and `int` into a design matrix `X` using the function `model.matrix` available in R.[1] The first column of the design matrix `X` consists of all ones. The second column is the predictor `so` which codes the factor for relative clause type. The third column the predictor `dist` which codes the factor for distance. The fourth column is the predictor `int` which codes the interaction between relative clause type and distance. The model matrix thus consists of a fully factorial 2×2 design, with blocks of this design repeated for each subject. For the full data-set, we could write it very compactly in matrix form as follows:

[1] Here, we would like to acknowledge the contribution of Douglas Bates in specifying the model in this general matrix form.

```
data {
  int<lower=0> N;               //no trials
  int<lower=1> P;               //no fixefs
  int<lower=0> J;               //no subjects
  int<lower=1> n_u;             //no subj ranefs
  int<lower=0> K;               //no items
  int<lower=1> n_w;             //no item ranefs
  int<lower=1,upper=J> subj[N]; //subject indicator
  int<lower=1,upper=K> item[N]; //item indicator
  row_vector[P] X[N];           //fixef design matrix
  row_vector[n_u] Z_u[N];       //subj ranef design matrix
  row_vector[n_w] Z_w[N];       //item ranef design matrix
  vector[N] rt;                 //reading time
}
parameters {
  vector[P] beta;               //fixef coefs
  cholesky_factor_corr[n_u] L_u;  //cholesky factor of subj ranef corr matrix
  cholesky_factor_corr[n_w] L_w;  //cholesky factor of item ranef corr matrix
  vector<lower=0>[n_u] sigma_u; //subj ranef std
  vector<lower=0>[n_w] sigma_w; //item ranef std
  real<lower=0> sigma_e;        //residual std
  vector[n_u] z_u[J];           //subj ranef
  vector[n_w] z_w[K];           //item ranef
}
transformed parameters {
  vector[n_u] u[J];             //subj ranefs
  vector[n_w] w[K];             //item ranefs
  {
    matrix[n_u,n_u] Sigma_u;    //subj ranef cov matrix
    matrix[n_w,n_w] Sigma_w;    //item ranef cov matrix
    Sigma_u <- diag_pre_multiply(sigma_u,L_u);
    Sigma_w <- diag_pre_multiply(sigma_w,L_w);
    for(j in 1:J)
      u[j] <- Sigma_u * z_u[j];
    for(k in 1:K)
      w[k] <- Sigma_w * z_w[k];
  }
}
model {
  //priors
  L_u ~ lkj_corr_cholesky(2.0);
  L_w ~ lkj_corr_cholesky(2.0);
  for (j in 1:J)
    z_u[j] ~ normal(0,1);
  for (k in 1:K)
    z_w[k] ~ normal(0,1);
  //likelihood
  for (i in 1:N)
    rt[i] ~ lognormal(X[i] * beta +
                      Z_u[i] * u[subj[i]] +
                      Z_w[i] * w[item[i]],
                      sigma_e);
}
```

Listing 2: Stan code for Husain et al data.

$$\log(\mathbf{rt}) = \mathbf{X}\beta + \mathbf{Z}_u\mathbf{u} + \mathbf{Z}_w\mathbf{w} + \varepsilon \tag{2}$$

Here, $\mathbf{X}$ is the $N \times P$ model matrix (with $N = 1440$, since we have 1440 data points; and $P = 4$ since we have the intercept plus three other fixed effects), β is a $P \times 1$ vector of fixed effects parameters, $\mathbf{Z}_u$ and $\mathbf{Z}_w$ are the subject and item model matrices ($N \times P$), and u and w are the by-subject and by-item adjustments to the fixed effects estimates; these are identical to the design matrix $\mathbf{X}$ in the model with varying intercepts and varying slopes included. For more examples of similar model specifications in Stan, see the R package `RePsychLing` on github (https://github.com/dmbates/RePsychLing).

Having defined the model, we proceed to assemble the list `stanDat` of data, relying on the above matrix formulation; please refer to Listing 1. The number `N` of observations, the number `J` of subjects and `K` of items, the reading times `rt`, and the subject and item indicator variables `subj` and `item` are familiar from the previous models presented. The integer `P` is the number of fixed effects (four including the intercept). Model 2 includes a varying intercept u_0 and varying slopes u_1, u_2, u_3 for each subject, and so the number `n_u` of by-subject random effects equals `P`. Likewise, Model 2 includes a varying intercept w_0 and varying slopes w_1, w_2, w_3 for each item, and so the number `n_w` of by-item random effects also equals `P`. The data block contains the corresponding variables. We declare the fixed effects design matrix `X` as an array of `N` row vectors whose components are the predictors associated with the `N` reading times. Likewise for the subject and item random effects design matrices `Z_u` and `Z_w`, which correspond to $\mathbf{Z}_u$ and $\mathbf{Z}_w$ respectively in Model 2. The vector `beta` contains the fixed effects β_0, β_1, β_2, and β_3. The matrices `L_u`, `L_w` and the arrays `z_u`, `z_w` of vectors (not to be confused with the design matrices `Z_u` and `Z_w`) will generate the varying intercepts and slopes $u_0, \ldots, u_3$ and $w_0, \ldots, w_3$. The vector `sigma_u` contains the standard deviations of the by-subject varying intercepts and slopes $u_0, \ldots, u_3$, and the vector `sigma_w` contains the standard deviations of the by-item varying intercepts and slopes $w_0, \ldots, w_3$. The variable `sigma_e` is the standard deviation σ_e of the error ε. The transformed parameters block generates the by-subject intercepts and slopes $u_0, \ldots, u_3$ and the by-item intercepts and slopes $w_0, \ldots, w_3$.

We place lkj priors on the random effects correlation matrices through the `lkj_corr_cholesky(2.0)` priors on their Cholesky factors `L_u` and `L_w`. We implicitly place uniform priors on the fixed effects $\beta_0, \ldots, \beta_3$, the random effects standard deviations $\sigma_{u0}, \ldots, \sigma_{u3}$, and $\sigma_{w0}, \ldots, \sigma_{w3}$ and the error standard deviation σ_e by omitting any prior specifications for them in the model block. We specify the likelihood with the probability statement that `rt[i]` is distributed log-normally with mean `X[i] * beta + Z_u[i] * u[subj[i]] + Z_w[i] * w[item[i]]` and standard deviation `sigma_e`. The next step towards model-fitting is to pass the list `stanDat` to `stan`, which compiles a C++ program to sample from the posterior distribution of the model parameters.

Figure 1 plots histograms of the marginal posterior distribution of the fixed effects. The HPD interval of the fixed effect $\hat{\beta}_1$ for relative clause type is entirely below zero. This is evidence that object relatives are read faster than subject relatives. The HPD interval of the fixed effect $\hat{\beta}_2$ for distance is also entirely below zero. This is evidence of a slowdown when the verb (where reading time was measured) is closer to the head noun of the relative clause. The HPD of the interaction $\hat{\beta}_3$ between relative clause type and distance is greater than zero, which is evidence for a greater slowdown on subject relatives when the distance between the verb and head noun is short.

A major advantage of the above matrix formulation is that we do not need to write a new Stan model for a future repeated measures factorial design. All we have to do now is define the design matrix X appropriately, and include it (along with appropriately defined Z_u and Z_w for the subjects and items random effects) as part of the data specification that is passed to Stan.

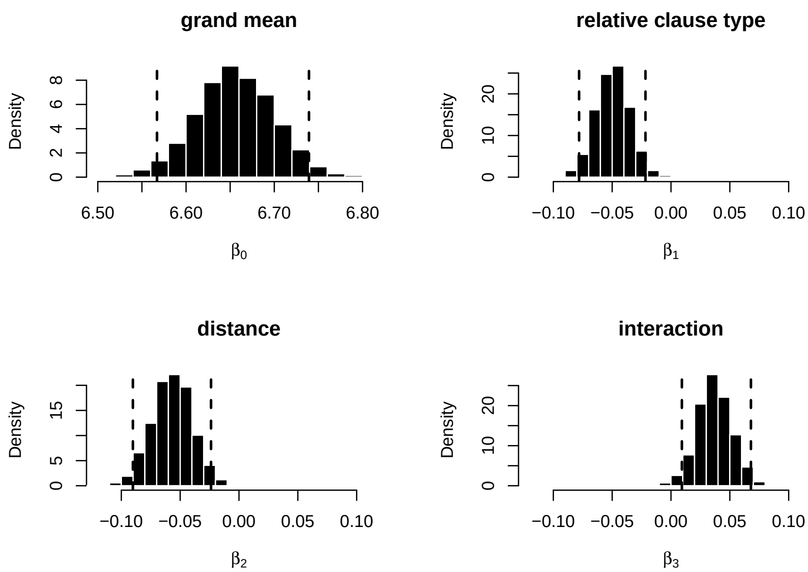

Figure 1. Marginal posterior distribution and HPD intervals of the fixed effects grand mean β_0, slope β_1 for relative clause type, slope β_2 for distance, and interaction β_3. All fixed effects are on the log-scale.

References

Gibson, E., & Wu, H.-H. I. (2013). Processing chinese relative clauses in context. *Language and Cognitive Processes*, *28*(1-2), 125–155.

Husain, S., Vasishth, S., & Srinivasan, N. (2014). Strong expectations cancel locality effects: Evidence from Hindi. *PLoS ONE*, *9*(7), 1–14.