An Introduction to Bayesian Multilevel Models Using brms: A Case Study of Gender Effects on Vowel Variability in Standard Indonesian

Preprint · May 2019

DOI: 10.31234/osf.io/guhsa

5 authors, including:

Cédric Batailler
Université Grenoble Alpes

Hélène Loevenbruck
French National Centre for Scientific Research

Paul-Christian Bürkner
Aalto University

An Introduction to Bayesian Multilevel Models Using brms: A Case Study of Gender Effects on Vowel Variability in Standard Indonesian

Ladislas Nalborczyk[1,2], Cédric Batailler[3], Hélène Lœvenbruck[1], Anne Vilain[4,5], & Paul-Christian Bürkner[6]

[1] Univ. Grenoble Alpes, CNRS, LPNC, 38000, Grenoble, France

[2] Department of Experimental Clinical and Health Psychology, Ghent University, Ghent, Belgium

[3] Univ. Grenoble Alpes, LIP/PC2S, 38000, Grenoble, France

[4] Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000, Grenoble, France

[5] Institut Universitaire de France, France

[6] Department of Psychology, University of Münster, Germany

Disclosure: The authors have declared that no competing interests existed at the time of publication. **Funding**: The first author of the manuscript is funded by a fellowship from Univ. Grenoble Alpes.

Author Note

Correspondence concerning this article should be addressed to Ladislas Nalborczyk, Laboratoire de Psychologie et Neurocognition, Univ. Grenoble Alpes, 1251 avenue centrale, 38058 Grenoble Cedex 9, France. E-mail: ladislas.nalborczyk@univ-grenoble-alpes.fr

Abstract

Purpose: Bayesian multilevel models are increasingly used to overcome the limitations of frequentist approaches in the analysis of complex structured data. This paper introduces Bayesian multilevel modelling for the specific analysis of speech data, using the brms package developed in R. **Method**: In this tutorial, we provide a practical introduction to Bayesian multilevel modelling, by reanalysing a phonetic dataset containing formant (F1 and F2) values for five vowels of Standard Indonesian (ISO 639-3:ind), as spoken by eight speakers (four females), with several repetitions of each vowel. **Results**: We first give an introductory overview of the Bayesian framework and multilevel modelling. We then show how Bayesian multilevel models can be fitted using the probabilistic programming language Stan and the R package brms, which provides an intuitive formula syntax. **Conclusions**: Through this tutorial, we demonstrate some of the advantages of the Bayesian framework for statistical modelling and provide a detailed case study, with complete source code for full reproducibility of the analyses (https://osf.io/dpzcb/).

Keywords: Bayesian data analysis, multilevel models, mixed models, brms, Stan

An Introduction to Bayesian Multilevel Models Using brms: A Case Study of Gender Effects on Vowel Variability in Standard Indonesian

wordcount (excluding abstract, references, tables and figures): 8637

1 Introduction

The last decade has witnessed noticeable changes in the way experimental data are analysed in phonetics, psycholinguistics, and speech sciences in general. In particular, there has been a shift from analysis of variance (ANOVA) to *linear mixed models*, also known as *hierarchical models* or *multilevel models* (MLMs), spurred by the spreading use of data-oriented programming languages such as `R` (R Core Team, 2017), and by the enthusiasm of its active and ever growing community. This shift has been further sustained by the current transition in data analysis in social sciences, with researchers evolving from a widely criticised point-hypothesis mechanical testing (e.g., Bakan, 1966; Gigerenzer, Krauss, & Vitouch, 2004; Kline, 2004; Lambdin, 2012; Trafimow et al., 2018) to an approach that emphasises parameter estimation, model comparison, and continuous model expansion (e.g., Cumming, 2012, 2014; Gelman et al., 2013; Gelman & Hill, 2007; Kruschke, 2015; Kruschke & Liddell, 2017a, 2017b; McElreath, 2016).

MLMs offer great flexibility in the sense that they can model statistical phenomena that occur on different levels. This is done by fitting models that include both constant and varying effects (sometimes referred to as *fixed* and *random* effects). Among other advantages, this makes it possible to generalise the results to unobserved levels of the *groups* existing in the data (e.g., stimulus or participant, Janssen, 2012). The multilevel strategy can be especially useful when dealing with repeated measurements (e.g., when measurements are nested into participants) or with unequal sample sizes, and more generally, when handling complex dependency structures in the data. Such complexities are frequently found in the kind of experimental designs used in speech science studies, for which MLMs are therefore particularly well suited.

The standard MLM is usually fitted in a frequentist framework, with the `lme4` package (Bates et al., 2015b) in R (R Core Team, 2017). However, when one tries to include the maximal varying effect structure, this kind of model tends either not to converge, or to give aberrant estimations of the correlation between varying effects (e.g., Bates et al., 2015a)[1]. Yet, fitting the maximal varying effect structure has been explicitly recommended (e.g., Barr, Levy, Scheepers, & Tily, 2013). In contrast, the maximal varying effect structure can generally be fitted in a Bayesian framework (Bates et al., 2015a; Eager & Roy, 2017; Nicenboim & Vasishth, 2016; Sorensen, Hohenstein, & Vasishth, 2016).

Another advantage of Bayesian statistical modelling is that it fits the way researchers intuitively understand statistical results. Widespread misinterpretations of frequentist statistics (like p-values and confidence intervals) are often attributable to the wrong interpretation of these statistics as resulting from a Bayesian analysis (e.g., Dienes, 2011; Gigerenzer et al., 2004; Hoekstra, Morey, Rouder, & Wagenmakers, 2014; Kruschke & Liddell, 2017a; Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2015). However, the intuitive nature of the Bayesian approach might arguably be hidden by the predominance of frequentist teaching in undergraduate statistical courses.

Moreover, the Bayesian approach offers a natural solution to the problem of multiple comparisons, when the situation is adequately modelled in a multilevel framework (Gelman, Hill, & Yajima, 2012; Scott & Berger, 2010), and allows *a priori* knowledge to be incorporated in data analysis via the prior distribution. The latter feature is particularily relevant when dealing with contraint parameters or for the purpose of incorporating expert knowledge.

The aim of the current paper is to introduce Bayesian multilevel models, and to provide an accessible and illustrated hands-on tutorial for analysing typical phonetic data. This paper will be structured in two main parts. First, we will briefly introduce the Bayesian approach to data analysis and the multilevel modelling strategy. Second, we will illustrate how Bayesian MLMs can be implemented in R by using the `brms` package (Bürkner, 2017b) to reanalyse a dataset from McCloy (2014) available in the `phonR` package (McCloy, 2016).

[1] In this context, the *maximal varying effect structure* means that any potential source of systematic influence should be explicitly modelled, by adding appropriate varying effects.

We will fit Bayesian MLMs of increasing complexity, going step by step, providing explanatory figures and making use of the tools available in the `brms` package for model checking and model comparison. We will then compare the results obtained in a Bayesian framework using `brms` with the results obtained using frequentist MLMs fitted with `lme4`. Throughout the paper, we will also provide comments and recommendations about the feasability and the relevance of such analysis for the researcher in speech sciences.

1.1 Bayesian data analysis

The Bayesian approach to data analysis differs from the frequentist one in that each parameter of the model is considered as a random variable (contrary to the frequentist approach which considers parameter values as unknown and fixed quantities), and by the explicit use of probability to model the uncertainty (Gelman et al., 2013). The two approaches also differ in their conception of what *probability* is. In the Bayesian framework, probability refers to the experience of uncertainty, while in the frequentist framework it refers to the limit of a relative frequency (i.e., the relative frequency of an event when the number of trials approaches infinity). A direct consequence of these two differences is that Bayesian data analysis allows researchers to discuss the probability of a parameter (or a vector of parameters) θ, given a set of data y:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

Using this equation (known as Bayes' theorem), a probability distribution $p(\theta|y)$ can be derived (called the *posterior distribution*), that reflects knowledge about the parameter, given the data and the prior information. This distribution is the goal of any Bayesian analysis and contains all the information needed for inference.

The term $p(\theta)$ corresponds to the *prior distribution*, which specifies the prior information about the parameters (i.e., what is known about θ before observing the data) as a probability distribution. The left hand of the numerator $p(y|\theta)$ represents the *likelihood*, also called the *sampling distribution* or *generative model*, and is the function through which the data affect the posterior distribution. The likelihood function indicates how likely the data are to appear, for each possible value of θ.

Finally, $p(y)$ is called the *marginal likelihood*. It is meant to normalise the posterior distribution, that is, to scale it in the "probability world". It gives the "probability of the data", summing over all values of θ and is described by $p(y) = \sum_\theta p(\theta)p(y|\theta)$ for discrete parameters, and by $p(y) = \int p(\theta)p(y|\theta)d\theta$ in the case of continuous parameters.

All this pieced together shows that the result of a Bayesian analysis, namely the posterior distribution $p(\theta|y)$, is given by the product of the information contained in the data (i.e., the likelihood) and the information available before observing the data (i.e., the prior). This constitutes the crucial principle of Bayesian inference, which can be seen as an updating mechanism (as detailed for instance in Kruschke & Liddell, 2017a). To sum up, Bayes' theorem allows a prior state of knowledge to be updated to a posterior state of knowledge, which represents a compromise between the prior knowledge and the empirical evidence.

The process of Bayesian analysis usually involves three steps that begin with setting up a probability model for all the entities at hand, then computing the posterior distribution, and finally evaluating the fit and the relevance of the model (Gelman et al., 2013). In the context of linear regression, for instance, the first step would require to specify a likelihood function for the data and a prior distribution for each parameter of interest (e.g., the intercept or the slope). We will go through these three steps in more details in the application section, but we will first give a brief overview of the multilevel modelling strategy.

1.2 Multilevel modelling

MLMs can be considered as "multilevel" for at least two reasons. First, an MLM can generally be conceived as a regression model in which the parameters are themselves modelled as outcomes of another regression model. The parameters of this second-level regression are known as *hyperparameters*, and are also estimated from the data (Gelman &

Hill, 2007). Second, the multilevel structure can arise from the data itself, for instance when one tries to model the second-language speech intelligibility of a child, who is considered within a particular class, itself considered within a particular school. In such cases, the hierarchical structure of the data itself calls for hierarchical modelling. In both conceptions, the number of levels that can be handled by MLMs is virtually unlimited (McElreath, 2016). When we use the term *multilevel* in the following, we will refer to the structure of the model, rather than to the structure of the data, as non-nested data can also be modelled in a multilevel framework.

As briefly mentioned earlier, MLMs offer several advantages compared to single-level regression models, as they can handle the dependency between units of analysis from the same group (e.g., several observations from the same participant). In other words, they can account for the fact that, for instance, several observations are not independent, as they relate to the same participant. This is achieved by partitioning the total variance into variation due to the groups (level-2) and to the individual (level-1). As a result, such models provide an estimation of the variance component for the second level (i.e., the variability of the participant-specific estimates) or higher levels, which can inform us about the generalisability of the findings (Janssen, 2012; McElreath, 2016).

Multilevel modelling allows both *fixed* and *random* effects to be incorporated. However, as pointed out by Gelman (2005), we can find at least five different (and sometimes contradictory) ways of defining the meaning of the terms *fixed* and *random* effects. Moreover, Gelman and Hill (2007) remarked that what is usually called a *fixed* effect can generally be conceived as a *random* effect with a null variance. In order to use a consistent vocabulary, we follow the recommendations of Gelman and Hill (2007) and avoid these terms. We instead use the more explicit terms *constant* and *varying* to designate effects that are constant, or that vary by groups[2].

[2] Note that MLMs are sometimes called *mixed models*, as models that comprise both *fixed* and *random* effects.

A question one is frequently faced with in multilevel modelling is to know which parameters should be considered as varying, and which parameters should be considered as constant. A practical answer is provided by McElreath (2016), who states that "any batch of parameters with *exchangeable* index values can be and probably should be pooled". For instance, if we are interested in the categorisation of native versus non-native phonemes and if for each phoneme in each category there are multiple audio stimuli (e.g., multiple repetitions of the same phoneme), and if we do not have any reason to think that, for each phoneme, audio stimuli may differ in intelligibility in any systematic way, then repetitions of the same phoneme should be pooled together. The essential feature of this strategy is that *exchangeability* of the lower units (i.e., the multiple repetitions of the same phoneme) is achieved by conditioning on indicator variables (i.e., the phonemes) that represent groupings in the population (Gelman et al., 2013).

To sum up, multilevel models are useful as soon as there are predictors at different levels of variation (Gelman et al., 2013). One important aspect is that this varying-coefficients approach allows each subgroup to have a different mean outcome level, while still estimating the global mean outcome level. In an MLM, these two estimations inform each other in a way that leads to the phenomenon of *shrinkage*, that will be discussed in more detail below (see section 2.3).

As an illustration, we will build an MLM starting from the ordinary linear regression model, and trying to predict an outcome y_i (e.g., second-language (L2) speech-intelligibility) by a linear combination of an intercept α and a slope β that quantifies the influence of a predictor x_i (e.g., the number of lessons received in this second language):

$$
\begin{aligned}
y_i &\sim \mathrm{Normal}(\mu_i, \sigma_e) \\
\mu_i &= \alpha + \beta x_i
\end{aligned}
$$

This notation is strictly equivalent to the (maybe more usual) following notation:

$$y_i = \alpha + \beta x_i + \epsilon_i$$
$$\epsilon_i \sim \text{Normal}(0, \sigma_e)$$

We prefer to use the first notation as it generalises better to more complex models, as we will see later. In Bayesian terms, these two lines describe the *likelihood* of the model, which is the assumption made about the generative process from which the data is issued. We make the assumption that the outcomes y_i are normally distributed around a mean μ_i with some error σ_e. This is equivalent to saying that the errors are normally distributed around 0, as illustrated by the above equivalence. Then, we can extend this model to the following multilevel model, adding a varying intercept:

$$y_i \sim \text{Normal}(\mu_i, \sigma_e)$$
$$\mu_i = \alpha_{j[i]} + \beta x_i$$
$$\alpha_j \sim \text{Normal}(\alpha, \sigma_\alpha)$$

where we use the notation $\alpha_{j[i]}$ to indicate that each group j (e.g., class) is given a unique intercept, issued from a Gaussian distribution centered on α, the grand intercept[3], meaning that there might be different mean scores for each class. From this notation we can see that in addition to the residual standard deviation σ_e, we are now estimating one more variance component σ_α, which is the standard deviation of the distribution of varying intercepts. We can interpret the variation of the parameter α between groups j by considering the *intra-class correlation* (ICC) $\sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_e^2)$, which goes to 0, if the grouping conveys no information, and to 1, if all observations in a group are identical (Gelman & Hill, 2007, p. 258).

The third line is called a *prior* distribution in the Bayesian framework. This prior

[3] Acknowledging that these individual intercepts can also be seen as adjustments to the grand intercept α, that are specific to group j.

distribution describes the population of intercepts, thus modelling the dependency between these parameters.

Following the same strategy, we can add a varying slope, allowed to vary according to the group j:

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma_e) \\
\mu_i &= \alpha_{j[i]} + \beta_{j[i]} x_i \\
\alpha_j &\sim \text{Normal}(\alpha, \sigma_\alpha) \\
\beta_j &\sim \text{Normal}(\beta, \sigma_\beta)
\end{aligned}
$$

Indicating that the effect of the number of lessons on L2 speech intelligibility is allowed to differ from one class to another (i.e., the effect of the number of lessons might be more beneficial to some classes than others). These varying slopes are assigned a prior distribution centered on the grand slope β, and with standard deviation σ_β.

In this introductory section, we have presented the foundations of Bayesian analysis and multilevel modelling. Bayes' theorem allows prior knowledge about parameters to be updated according to the information conveyed by the data, while MLMs allow complex dependency structures to be modelled. We now move to a detailed case study in order to illustrate these concepts.

Box 1. Where are my random effects ?

In the Bayesian framework, every unknown quantity is considered as a random variable that we can describe using probability distributions. As a consequence, there is no such thing as a "fixed effect" or a "random effects distribution" in a Bayesian framework. However, these semantic quarrels disappear when we write down the model.

Suppose we have a dependent continuous variable y and a dichotomic categorical predictor x (assumed to be contrast-coded). Let y_{ij} denote the score of the i^{th} participant in the j^{th} condition. We can write a "mixed effects" model (as containing both fixed and random effects) as follows:

$$y_{ij} = \alpha + \alpha_i + \beta x_j + e_{ij},\ e_{ij} \sim \text{Normal}(0, \sigma_e^2),\ \alpha_i \sim \text{Normal}(0, \sigma_u^2)$$

Where the terms α and β represent the "fixed effects" and denote the overall mean response and the condition difference in response, respectively. In addition, e_{ij} are random errors assumed to be normally distributed with unknown variance σ_e^2, and α_i's are individual specific random effects normally distributed in the population with unknown variance σ_a^2.

We can rewrite this model to make apparent that the so-called "random effects distribution" can actually be considered a prior distribution (from a Bayesian standpoint), since by definition, distributions on unknown quantities are considered as priors:

$$\begin{aligned} y_{ij} &\sim \text{Normal}(\mu_{ij}, \sigma_e^2) \\ \mu_{ij} &= \alpha_i + \beta x_j \\ \alpha_i &\sim \text{Normal}(\alpha, \sigma_\alpha^2) \end{aligned}$$

where the parameters of this prior are learned from the data. As we have seen, the same mathematical entity can be conceived either as a "random effects distribution" or as a prior distribution, depending on the framework.

1.3 Software programs

Sorensen et al. (2016) provided a detailed and accessible introduction to Bayesian MLMs (BMLMs) applied to linguistics, using the probabilistic language `Stan` (Stan Development Team, 2016). However, discovering BMLMs and the `Stan` language all at once might seem a little overwhelming, as `Stan` can be difficult to learn for users that are not experienced with programming languages. As an alternative, we introduce the `brms` package (Bürkner, 2017b), that implements BMLMs in `R`, using `Stan` under the hood, with an `lme4`-like syntax. Hence, the syntax required by `brms` will not surprise the researcher familiar with `lme4`, as models of the following form:

$$
\begin{aligned}
y_i &\sim \text{Normal}(\mu_i, \sigma_e) \\
\mu_i &= \alpha + \alpha_{subject[i]} + \beta x_i
\end{aligned}
$$

are specified in `brms` (as in `lme4`) with: `y ~ 1 + x + (1|subject)`. In addition to linear regression models, `brms` allows generalised linear and non-linear multilevel models to be fitted, and comes with a great variety of distribution and link functions. For instance, `brms` allows fitting robust linear regression models, or modelling dichotomous and categorical outcomes using logistic and ordinal regression models. The flexibility of `brms` also allows for distributional models (i.e., models that include simultaneous predictions of all response parameters), Gaussian processes or non-linear models to be fitted, among others. More information about the diversity of models that can be fitted with `brms` and their implementation is provided in Bürkner (2017b) and Bürkner (2017a).

2 Application example

To illustrate the use of BMLMs, we reanalysed a dataset from McCloy (2014), available in the `phonR` package (McCloy, 2016). This dataset contains formant (F1 and F2) values for five vowels of Standard Indonesian (ISO 639-3:ind), as spoken by eight speakers (four females), with approximately 45 repetitions of each vowel. The research question we investigated here is the effect of gender on vowel production variability.

2.1 Data pre-processing

Our research question was about the different amount of variability in the respective vowel productions of male and female speakers, due to cognitive or social differences. To answer this question, we first needed to get rid of the differences in vowel production that are due to physiological differences between males and females (e.g., shorter vocal tract length for females). More generally, we needed to eliminate the inter-individual differences due to physiological characteristics in our groups of participants. For that purpose, we first applied the Watt & Fabricius formant normalisation technique (Watt & Fabricius, 2002). The principle of this method is to calculate for each speaker a "centre of gravity" S in the F1/F2 plane, from the formant values of point vowels [i, a , u], and to express the formant values of each observation as ratios of the value of S for that formant.

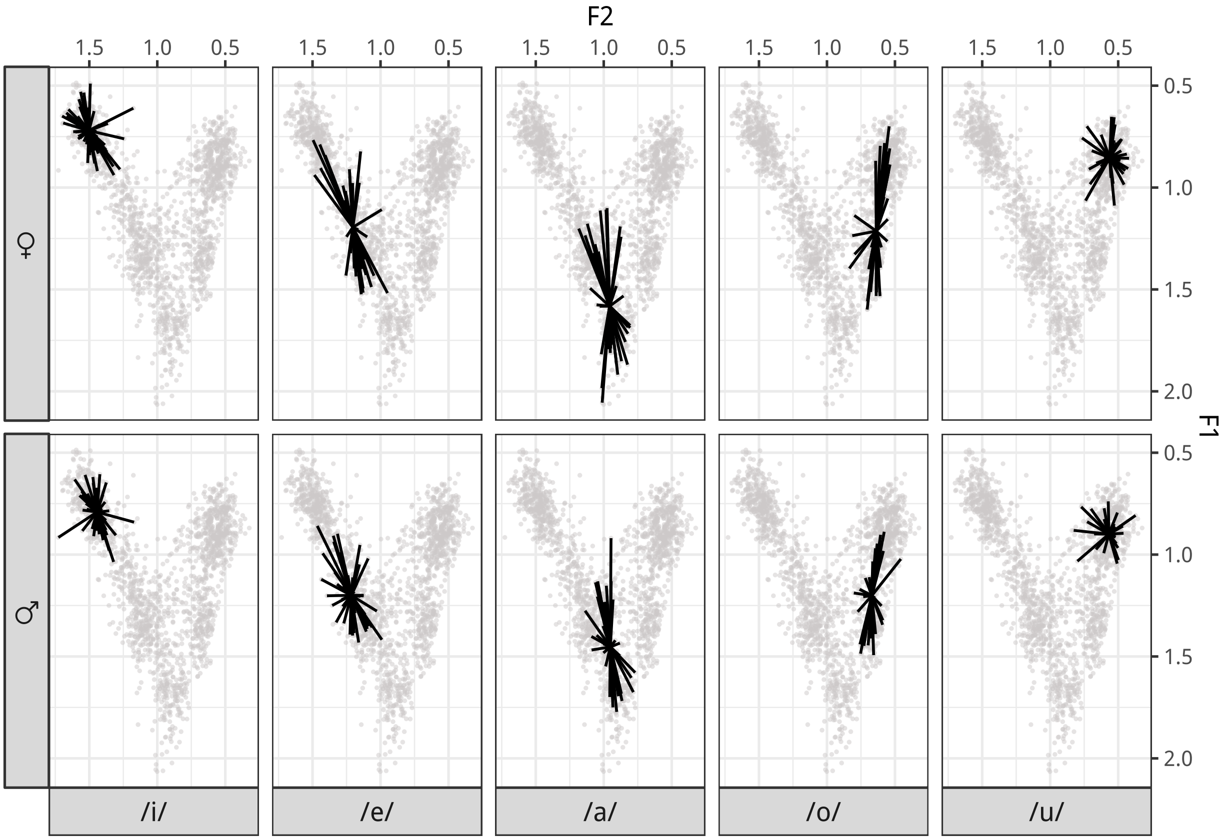

Figure 1. Euclidean distances between each observation and the centres of gravity corresponding to each vowel across all participants, by gender (top row: female, bottom row: male) and by vowel (in column), in the normalised F1-F2 plane. The grey background plots represent the individual data collapsed for all individuals (male and female) and all vowels. Note that, for the sake of clarity, this figure represents a unique center of gravity for each vowel for all participants, whereas in the analysis, one center of gravity was used for each vowel and each participant.

Then, for each vowel and participant, we computed the Euclidean distance between each observation and the centre of gravity of the whole set of observations in the F1-F2 plane for that participant and that vowel. The data obtained by this process are illustrated in Figure 1, and a sample of the final dataset can be found in Table 1.

Table 1

Ten randomly picked rows from the data.

subj	gender	vowel	f1	f2	f1norm	f2norm	distance	repetition
M02	m	/e/	534	1724	1.143	1.113	0.118	7
F09	f	/i/	468	2401	0.943	1.447	0.223	16
F04	f	/a/	885	1413	1.636	0.804	0.223	12
M01	m	/a/	671	1262	1.615	0.823	0.176	25
F04	f	/a/	700	1951	1.294	1.109	0.237	36
F04	f	/e/	614	2100	1.135	1.194	0.070	42
M04	m	/i/	338	2163	0.803	1.432	0.040	16
F04	f	/o/	649	1357	1.200	0.772	0.154	12
M04	m	/a/	524	1573	1.245	1.041	0.146	20
M02	m	/u/	411	762	0.879	0.492	0.134	25

2.2 Constant effect of gender on vowel production variability

We then built a first model with constant effects only and vague priors on α and β, the intercept and the slope. We contrast-coded `gender` (f = -0.5, m = 0.5). Our dependent variable was therefore the distance from each individual vowel centre of gravity, which we will refer to as *formant distance* in the following. The formal model can be expressed as:

$$
\begin{aligned}
\text{distance}_i &\sim \text{Normal}(\mu_i, \sigma_e) \\
\mu_i &= \alpha + \beta \times \text{gender}_i \\
\alpha &\sim \text{Normal}(0, 10) \\
\beta &\sim \text{Normal}(0, 10) \\
\sigma_e &\sim \text{HalfCauchy}(10)
\end{aligned}
$$

where the first two lines of the model describe the likelihood and the linear model[4]. The next three lines define the prior distribution for each parameter of the model, where α and β are given a vague (weakly informative) Gaussian prior centered on 0, and the residual variation is given a Half-Cauchy prior (Gelman, 2006; Polson & Scott, 2012), thus restricting the range of possible values to positive ones. As depicted in Figure 2, the $\text{Normal}(0, 10)$ prior is weakly informative in the sense that it grants a relative high weight to α and β values, between -25 and 25. This corresponds to very large (given the scale of our data) values for, respectively, the mean distance value α, and the mean difference between males and females β. The HalfCauchy(10) prior placed on σ_e also allows very large values of σ_e, as represented in the right panel of Figure 2.

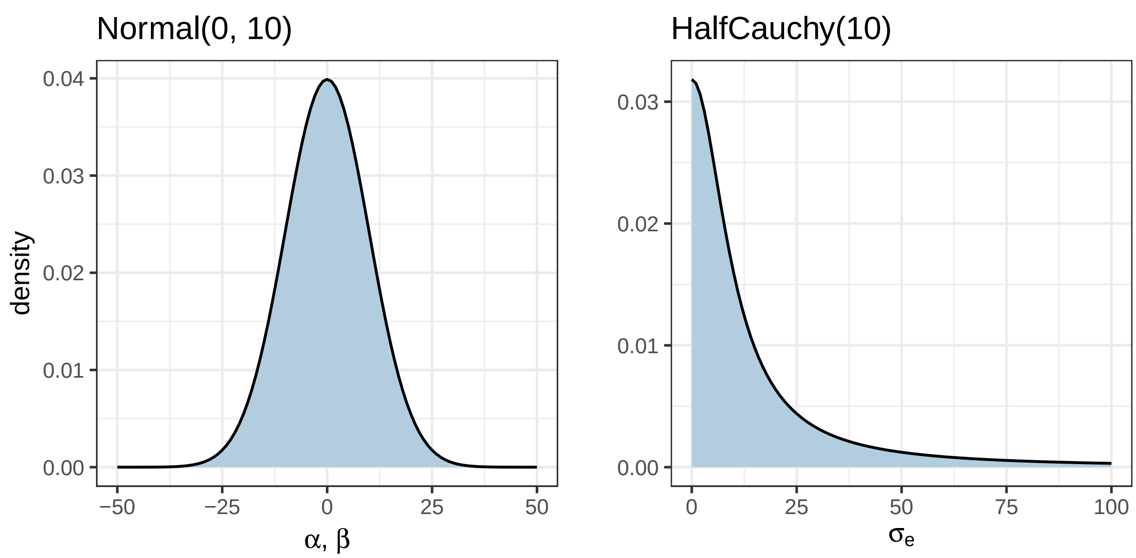

Figure 2. Prior distributions used in the first model, for α and β (left panel) and for the residual variation σ_e (right panel).

These priors can be specified in numerous ways (see `?set_prior` for more details), among which the following:

[4] Note that –for the sake of simplicity– throughout this tutorial we use a Normal likelihood, but other (better) alternatives would include using skew-normal or log-normal models, which are implemented in `brms` with the `skew_normal` and `lognormal` families. We provide examples in the supplementary materials.

```r
prior1 <- c(
    prior(normal(0, 10), class = Intercept),
    prior(normal(0, 10), class = b, coef = gender),
    prior(cauchy(0, 10), class = sigma)
    )
```

where a prior can be defined over a class of parameters (e.g., for all variance components, using the `sd` class) or for a specific one, for instance as above by specifying the coefficient (`coef`) to which the prior corresponds (here the slope of the constant effect of gender).

The model can be fitted with `brms` with the following command:

```r
library(brms)

bmod1 <- brm(
        distance ~ gender,
        data = indo, family = gaussian(),
        prior = prior1,
        warmup = 2000, iter = 5000
    )
```

where `distance` is the distance from the centre of gravity. The `iter` argument serves to specify the total number of iterations of the Markov Chain Monte Carlo (MCMC) algorithm, and the `warmup` argument specifies the number of iterations that are run at the beginning of the process to "calibrate" the MCMC, so that only `iter - warmup` iterations are retained in the end to approximate the shape of the posterior distribution (for more details, see McElreath, 2016).

Figure 3 depicts the estimations of this first model for the intercept α, the slope β, and the residual standard deviation σ_e. The left part of the plot shows histograms of draws taken from the posterior distribution, and from which several summaries can be computed (e.g., mean, mode, quantiles). The right part of Figure 3 shows the behaviour of the two simulations (i.e., the two chains) used to approximate the posterior distribution, where the x-axis represents the number of iterations and the y-axis the value of the parameter. This plot reveals one important aspect of the simulations that should be checked, known as *mixing*. A chain is considered well mixed if it explores many different values for the target parameters and does not stay in the same region of the parameter space. This feature can be evaluated by checking that these plots, usually referred to as *trace plots*, show random scatter around a mean value (they look like a "fat hairy caterpillar").

```
library(tidyverse)

bmod1 %>%
    plot(
        combo = c("hist", "trace"), widths = c(1, 1.5),
        theme = theme_bw(base_size = 10)
        )
```

The estimations obtained for this first model are summarised in Table 2, which includes the mean, the standard error (SE), and the lower and upper bounds of the 95% credible interval (CrI)[5] of the posterior distribution for each parameter. As `gender` was contrast-coded before the analysis (f = -0.5, m = 0.5), the intercept α corresponds to the grand mean of the formant distance over all participants and has its mean around 0.16. The estimate of the slope (β = -0.04) suggests that females are more variable than males in the way they pronounce vowels, while the 95% CrI can be interpreted in a way that there is a

[5] Where a credible interval is the Bayesian analogue of a classical confidence interval, except that probability statements can be made based upon it (e.g., "given the data and our prior assumptions, there is a 0.95 probability that this interval encompasses the population value θ").

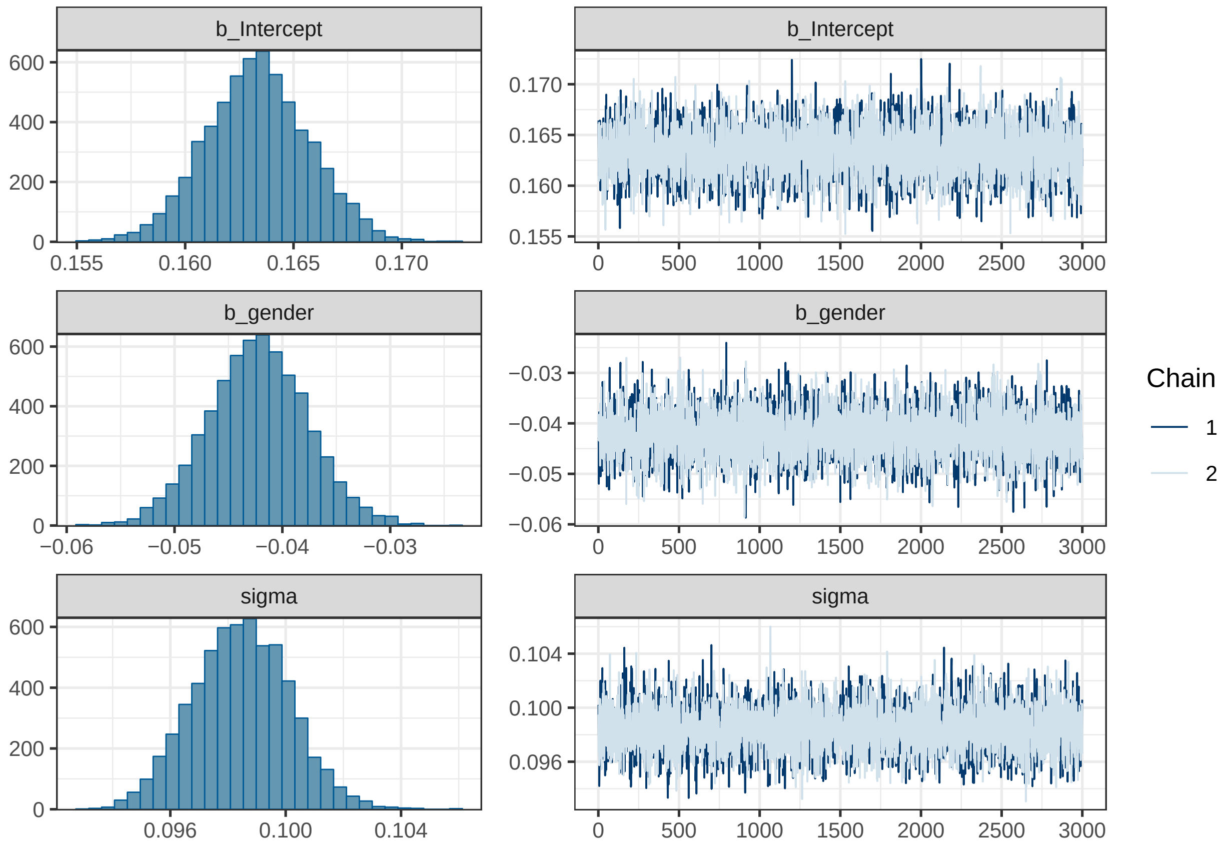

Figure 3. Histograms of posterior samples and trace plots of the intercept, the slope for gender and the standard deviation of the residuals of the constant effects model.

0.95 probability that the value of the intercept lies in the [-0.05, -0.03] interval.

Table 2

Posterior mean, standard error, 95% credible interval and $\hat{R}$ statistic for each parameter of the constant effect model bmod1.

parameter	mean	SE	lower bound	upper bound	Rhat
α	0.163	0.002	0.159	0.168	1.000
β	-0.042	0.005	-0.051	-0.033	1.000
σ_e	0.098	0.002	0.095	0.102	1.000

The `Rhat` value corresponds to the *potential scale reduction factor* $\hat{R}$ (Gelman & Rubin, 1992), that provides information about the convergence of the algorithm. This index can be conceived as equivalent to the F-ratio in ANOVA. It compares the between-chains variability (i.e., the extent to which different chains differ one from each other) to the within-chain variability (i.e., how widely a chain explores the parameter space), and, as such, gives an index of the convergence of the chains. An overly large between-chains variance (as compared to the within-chain variability) would be a sign that chain-specific characteristics, like the starting value of the algorithm, have a strong influence on the final result. Ideally, the value of `Rhat` should be close to 1, and should not exceed 1.1. Otherwise, one might consider running more iterations or defining stronger priors (Bürkner, 2017b; Gelman et al., 2013).

2.3 Varying intercept model

The first model can be improved by taking into account the dependency between vowel formant measures for each participant. This is handled in MLMs by specifying unique intercepts $\alpha_{subject[i]}$ and by assigning them a common prior distribution. This strategy corresponds to the following by-subject varying-intercept model, `bmod2`:

$$
\begin{aligned}
\text{distance}_i &\sim \text{Normal}(\mu_i, \sigma_e) \\
\mu_i &= \alpha + \alpha_{subject[i]} + \beta \times \text{gender}_i \\
\alpha_{subject} &\sim \text{Normal}(0, \sigma_{subject}) \\
\alpha &\sim \text{Normal}(0, 10) \\
\beta &\sim \text{Normal}(0, 10) \\
\sigma_{subject} &\sim \text{HalfCauchy}(10) \\
\sigma_e &\sim \text{HalfCauchy}(10)
\end{aligned}
$$

This model can be fitted with `brms` with the following command (where we specify the HalfCauchy prior on $\sigma_{subject}$ by applying it on parameters of class `sd`):

```r
prior2 <- c(
    prior(normal(0, 10), class = Intercept),
    prior(normal(0, 10), class = b, coef = gender),
    prior(cauchy(0, 10), class = sd),
    prior(cauchy(0, 10), class = sigma)
    )

bmod2 <- brm(
    distance ~ gender + (1|subj),
    data = indo, family = gaussian(),
    prior = prior2,
    warmup = 2000, iter = 10000
    )
```

As described in the first part of the present paper, we now have two sources of variation in the model: the standard deviation of the residuals σ_e and the standard deviation of the by-subject varying intercepts $\sigma_{subject}$. The latter represents the standard deviation of the population of varying intercepts, and is also learned from the data. It means that the estimation of each unique intercept will inform the estimation of the population of intercepts, which, in return, will inform the estimation of the other intercepts. We call this sharing of information between groups the *partial pooling* strategy, in comparison with the *no pooling* strategy, where each intercept is estimated independently, and with the *complete pooling* strategy, in which all intercepts are given the same value (Gelman et al., 2013; Gelman & Hill, 2007; McElreath, 2016). This is one of the most essential features of MLMs, and what leads to better estimations than single-level regression models for repeated measurements or unbalanced sample sizes. This pooling of information is made apparent through the phenomenon of *shrinkage*, which is illustrated in Figure 4, and later on, in Figure 6.

Figure 4 shows the posterior distribution as estimated by this second model for each participant, in relation to the raw mean of its category (i.e., females or males), represented by the vertical dashed lines. We can see for instance that participants `M02` and `F09` have smaller average distance than the means of their groups, while participants `M03` and `F08` have larger ones. The arrows represent the amount of *shrinkage*, that is, the deviation between the mean in the raw data (represented by a cross underneath each density) and the estimated mean of the posterior distribution (represented by the peak of the arrow). As shown in Figure 4, this *shrinkage* is always directed toward the mean of the considered group (i.e., females or males) and the amount of *shrinkage* is determined by the deviation of the individual mean from its group mean. This mechanism acts like a safeguard against overfitting, preventing the model from overly trusting each individual datum.

The marginal posterior distribution of each parameter obtained with `bmod2` is summarised in Table 3, where the `Rhat` values close to 1 suggest that the model has converged. We see that the estimates of α and β are similar to the estimates of the first model, except that the SE is now slightly larger. This result might seem surprising at first sight, as we expected to improve the first model by adding a by-subject varying intercept. In fact, it reveals an underestimation of the SE when using the first model. Indeed, the first model assumes independence of observations, which is violated in our case. This highlights the general need for careful consideration of the model's assumptions when interpreting its estimations. The first model seemingly gives highly certain estimates, but these estimations are only valid in the "independence of observations" world (see also the distinction between the *small world* and the *large world* in McElreath, 2016). Moreover, estimating an intercept by subject (as in the second model) increases the precision of estimation, but it also makes the average estimation less certain, thus resulting in a larger SE.

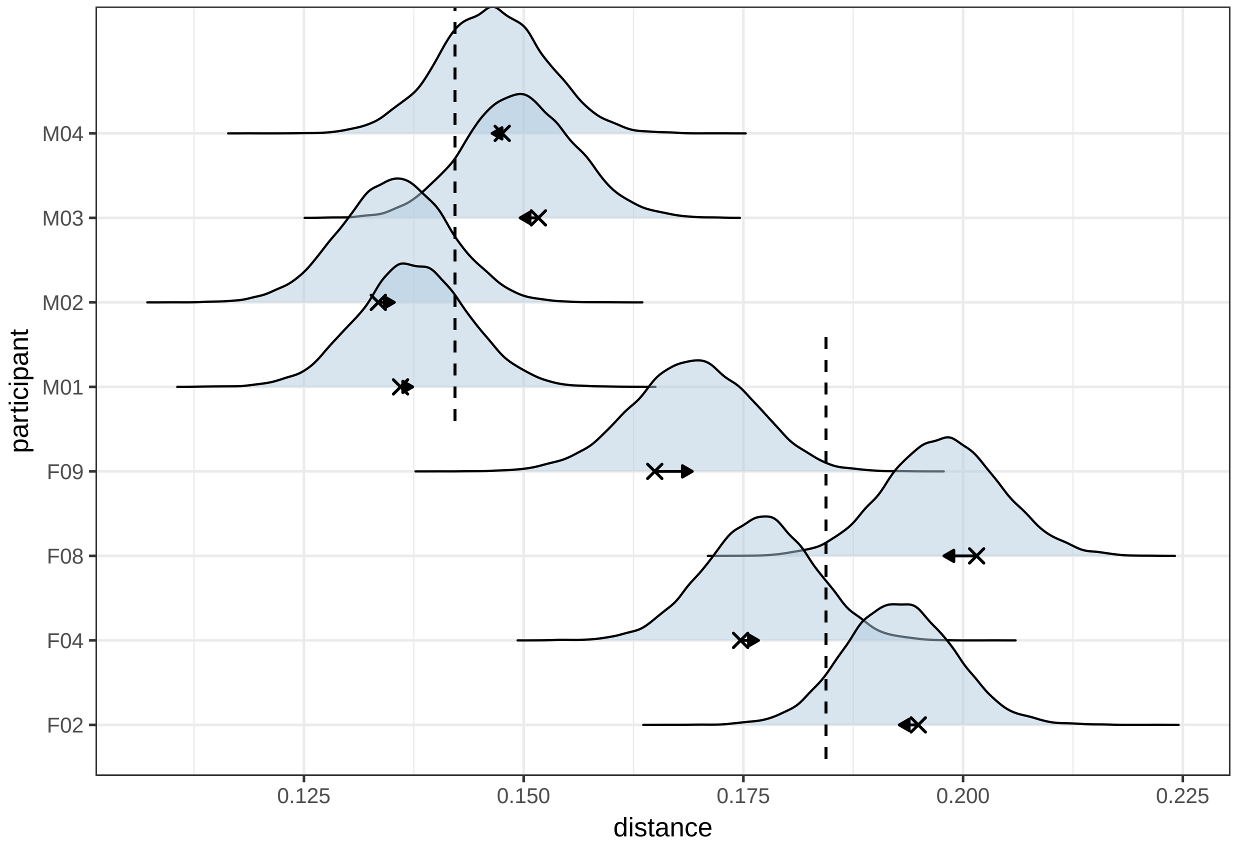

Figure 4. Posterior distributions by subject, as estimated by the `bmod2` model. The vertical dashed lines represent the means of the formant distances for the female and male groups. Crosses represent the mean of the raw data, for each participant. Arrows represent the amount of shrinkage, between the raw mean and the estimation of the model (the mean of the posterior distribution).

Table 3

Posterior mean, standard error, 95% credible interval and $\hat{R}$ statistic for each parameter of model bmod2 with a varying intercept by subject.

parameter	mean	SE	lower bound	upper bound	Rhat
α	0.163	0.006	0.150	0.176	1.001
β	-0.042	0.013	-0.068	-0.017	1.001
$\sigma_{subject}$	0.016	0.008	0.006	0.035	1.000
σ_e	0.098	0.002	0.095	0.101	1.000

This model (`bmod2`), however, is still not adequate to describe the data, as the dependency between repetitions of each vowel is not taken into account. In `bmod3`, we added a by-vowel varying intercept, thus also allowing each vowel to have a different general level of variability.

$$
\begin{aligned}
\text{distance}_i &\sim \text{Normal}(\mu_i, \sigma_e) \\
\mu_i &= \alpha + \alpha_{subject[i]} + \alpha_{vowel[i]} + \beta \times \text{gender}_i \\
\alpha_{subj} &\sim \text{Normal}(0, \sigma_{subject}) \\
\alpha_{vowel} &\sim \text{Normal}(0, \sigma_{vowel}) \\
\alpha &\sim \text{Normal}(0, 10) \\
\beta &\sim \text{Normal}(0, 10) \\
\sigma_e &\sim \text{HalfCauchy}(10) \\
\sigma_{subject} &\sim \text{HalfCauchy}(10) \\
\sigma_{vowel} &\sim \text{HalfCauchy}(10)
\end{aligned}
$$

This model can be fitted with `brms` with the following command:

```r
prior3 <- c(
    prior(normal(0, 10), class = Intercept),
    prior(normal(0, 10), class = b, coef = gender),
    prior(cauchy(0, 10), class = sd),
    prior(cauchy(0, 10), class = sigma)
    )

bmod3 <- brm(
    distance ~ gender + (1|subj) + (1|vowel),
    data = indo, family = gaussian(),
```

```
    prior = prior3,
    warmup = 2000, iter = 10000
    )
```

where the same Half-Cauchy is specified for the two varying intercepts, by applying it directly to the `sd` class.

Table 4

Posterior mean, standard error, 95% credible interval and $\hat{R}$ statistic for each parameter of model bmod3 with a varying intercept by subject and by vowel.

parameter	mean	SE	lower bound	upper bound	Rhat
α	0.164	0.040	0.086	0.244	1.000
β	-0.042	0.013	-0.069	-0.014	1.000
$\sigma_{subject}$	0.017	0.008	0.007	0.036	1.000
σ_{vowel}	0.075	0.048	0.031	0.196	1.000
σ_e	0.088	0.002	0.085	0.091	1.000

The marginal posterior distribution of each parameter is summarised in Table 4. We can compute the intra-class correlation (ICC, see section 1.2) to estimate the relative variability associated with each varying effect: $ICC_{subject}$ is equal to 0.03 and ICC_{vowel} is equal to 0.42. The rather high ICC for vowels suggests that observations are highly correlated within each vowel, thus stressing the relevance of allocating a unique intercept by vowel[6].

[6] But please note that we do not mean to suggest that the varying intercept for subjects should be removed because its ICC is low.

2.4 Including a correlation between varying intercept and varying slope

One can legitimately question the assumption that the differences between male and female productions are identical for each vowel. To explore this issue, we thus added a varying slope for the effect of gender, allowing it to vary by vowel. Moreover, we can exploit the correlation between the baseline level of variability by vowel, and the amplitude of the difference between males and females in pronouncing them. For instance, we can observe that the pronunciation of /a/ is more variable in general. We might want to know whether females tend to pronounce vowels that are situated at a specific location in the F1-F2 plane with less variability than males. In other words, we might be interested in knowing whether the effect of `gender` is correlated with the baseline level of variability. This is equivalent to investigating the *dependency*, or the correlation between the varying intercepts and the varying slopes. We thus estimated this correlation by modelling α_{vowel} and β_{vowel} as issued from the same multivariate normal distribution (a multivariate normal distribution is a generalisation of the usual normal distribution to more than one dimension), centered on 0 and with some covariance matrix $\mathbf{S}$, as specified on the third line of the following model:

$$
\begin{aligned}
\text{distance}_i &\sim \text{Normal}(\mu_i, \sigma_e)\\
\mu_i &= \alpha + \alpha_{subject[i]} + \alpha_{vowel[i]} + (\beta + \beta_{vowel[i]}) \times \text{gender}_i\\
\begin{bmatrix} \alpha_{\text{vowel}} \\ \beta_{\text{vowel}} \end{bmatrix} &\sim \text{MVNormal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{S} \right)\\
\mathbf{S} &= \begin{pmatrix} \sigma^2_{\alpha_{vowel}} & \sigma_{\alpha_{vowel}} \sigma_{\beta vowel} \rho \\ \sigma_{\alpha_{vowel}} \sigma_{\beta vowel} \rho & \sigma^2_{\beta_{vowel}} \end{pmatrix}\\
\alpha_{subject} &\sim \text{Normal}(0, \sigma_{subject})\\
\alpha &\sim \text{Normal}(0, 10)\\
\beta &\sim \text{Normal}(0, 10)\\
\sigma_e &\sim \text{HalfCauchy}(10)\\
\sigma_{\alpha_{vowel}} &\sim \text{HalfCauchy}(10)\\
\sigma_{\beta_{vowel}} &\sim \text{HalfCauchy}(10)\\
\sigma_{subject} &\sim \text{HalfCauchy}(10)\\
\mathbf{R} &\sim \text{LKJcorr}(2)
\end{aligned}
$$

where $\mathbf{R}$ is the correlation matrix $\mathbf{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ and ρ is the correlation between intercepts and slopes, used in the computation of $\mathbf{S}$. This matrix is given the LKJ-Correlation prior (Lewandowski, Kurowicka, & Joe, 2009) with a parameter ζ (zeta) that controls the strength of the correlation[7]. When $\zeta = 1$, the prior distribution on the correlation is uniform between -1 and 1. When $\zeta > 1$, the prior distribution is peaked around a zero correlation, while lower values of ζ ($0 < \zeta < 1$) allocate more weight to extreme values (i.e., close to -1 and 1) of ρ (see Figure 5).

[7] The LKJ prior is the default prior for correlation matrices in `brms`.

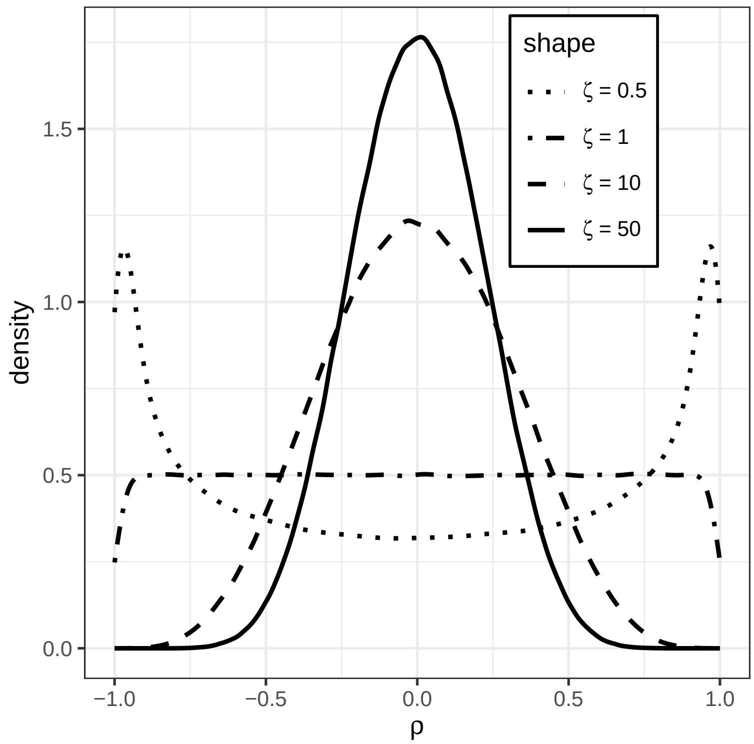

Figure 5. Visualisation of the LKJ prior for different values of the shape parameter ζ.

```
prior4 <- c(
    prior(normal(0, 10), class = Intercept),
    prior(normal(0, 10), class = b, coef = gender),
    prior(cauchy(0, 10), class = sd),
    prior(cauchy(0, 10), class = sigma),
    prior(lkj(2), class = cor)
    )

bmod4 <- brm(
    distance ~ gender + (1|subj) + (1 + gender|vowel),
    data = indo, family = gaussian(),
    prior = prior4,
    warmup = 2000, iter = 10000
```

```
)
```

Estimates of this model are summarised in Table 5. This summary reveals a negative correlation between the intercepts and slopes for vowels, meaning that vowels with a large "baseline level of variability" (i.e., with a large average `distance` value) tend to be pronounced with more variability by females than by males. However, we notice that this model's estimation of β is even more uncertain than that of the previous models, as shown by the associated standard error and the width of the credible interval.

Table 5

Posterior mean, standard error, 95% credible interval and $\hat{R}$ statistic for each parameter of model bmod4 with a varying intercept and varying slope by vowel.

parameter	mean	SE	lower bound	upper bound	Rhat
α	0.164	0.036	0.096	0.237	1.001
β	-0.042	0.030	-0.099	0.016	1.000
$\sigma_{subject}$	0.016	0.008	0.007	0.036	1.000
$\sigma_{\alpha_{vowel}}$	0.067	0.043	0.029	0.171	1.000
$\sigma_{\beta_{vowel}}$	0.052	0.031	0.022	0.132	1.000
ρ	-0.497	0.356	-0.951	0.371	1.001
σ_e	0.086	0.001	0.084	0.089	1.000

Figure 6 illustrates the negative correlation between the by-vowel intercepts and the by-vowel slopes, meaning that vowels that tend to have higher "baseline variability" (i.e., /e/, /o/, /a/), tend to show a stronger effect of `gender`. This figure also illustrates the amount of shrinkage, here in the parameter space. We can see that the *partial pooling* estimate is shrunk somewhere between the *no pooling* estimate and the *complete pooling* estimate (i.e., the grand mean). This illustrates again the mechanism by which MLMs balance the risk of overfitting and underfitting (McElreath, 2016).

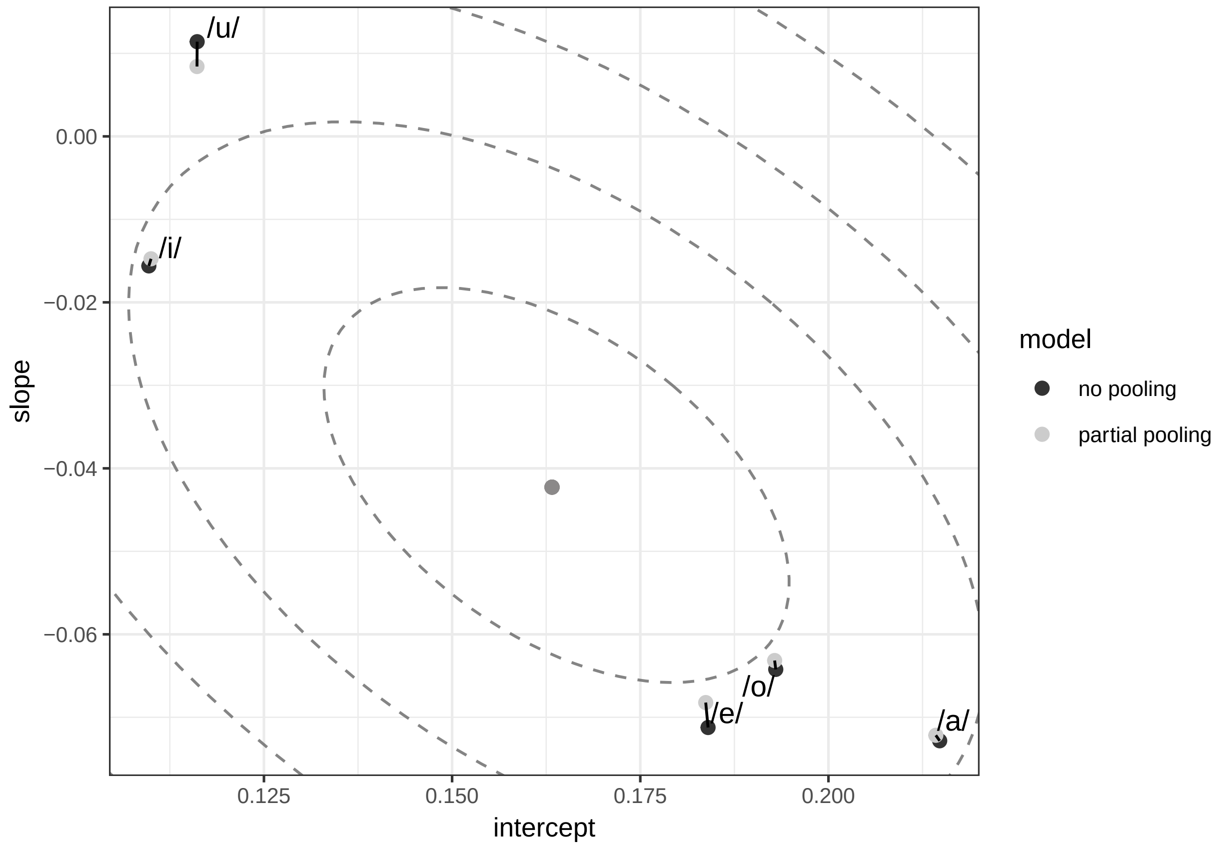

Figure 6. Shrinkage of estimates in the parameter space, due to the pooling of information between clusters (based on the `bmod4` model). The ellipses represent the contours of the bivariate distribution, at different degrees of confidence 0.1, 0.3, 0.5 and 0.7.

2.5 Varying intercept and varying slope model, interaction between subject and vowel

So far, we modelled varying effects of subjects and vowels. In this study, these varying factors were crossed, meaning that every subject had to pronounce every vowel. Let us now imagine a situation in which Subject 4 systematically mispronounced the /i/ vowel. This would be a source of systematic variation over replicates which is not considered in the model (`bmod4`), because this model can only adjust parameters for either vowel or participant, but not for a specific vowel for a specific participant.

In building the next model, we added a varying intercept for the interaction between subject and vowel, that is, we created an index variable that allocates a unique value at each crossing of the two variables (e.g., Subject1-vowel/a/, Subject1-vowel/i/, etc.), resulting in 8 $\times$ 5 = 40 intercepts to be estimated (for a review of multilevel modeling in various experimental designs, see Judd, Westfall, & Kenny, 2017). This varying intercept for the interaction between subject and vowel represents the systematic variation associated with a specific subject pronouncing a specific vowel. This model can be written as follows, for any observation i:

$$
\begin{aligned}
\text{distance}_i &\sim \text{Normal}(\mu_i, \sigma_e) \\
\mu_i &= \alpha + \alpha_{subject[i]} + \alpha_{vowel[i]} + \alpha_{subject:vowel[i]} + (\beta + \beta_{vowel[i]}) \times \text{gender}_i \\
\begin{bmatrix} \alpha_{\text{vowel}} \\ \beta_{\text{vowel}} \end{bmatrix} &\sim \text{MVNormal}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{S} \right) \\
\mathbf{S} &= \begin{pmatrix} \sigma^2_{\alpha_{vowel}} & \sigma_{\alpha_{vowel}} \sigma_{\beta vowel} \rho \\ \sigma_{\alpha_{vowel}} \sigma_{\beta vowel} \rho & \sigma^2_{\beta_{vowel}} \end{pmatrix} \\
\alpha_{subject} &\sim \text{Normal}(0, \sigma_{subject}) \\
\alpha_{subject:vowel} &\sim \text{Normal}(0, \sigma_{subject:vowel}) \\
\alpha &\sim \text{Normal}(0, 10) \\
\beta &\sim \text{Normal}(0, 10) \\
\sigma_e &\sim \text{HalfCauchy}(10) \\
\sigma_{subject} &\sim \text{HalfCauchy}(10) \\
\sigma_{subject:vowel} &\sim \text{HalfCauchy}(10) \\
\sigma_{\alpha_{vowel}} &\sim \text{HalfCauchy}(10) \\
\sigma_{\beta_{vowel}} &\sim \text{HalfCauchy}(10) \\
\mathbf{R} &\sim \text{LKJcorr}(2)
\end{aligned}
$$

This model can be fitted with the following command:

```
prior5 <- c(
    prior(normal(0, 10), class = Intercept),
    prior(normal(0, 10), class = b, coef = gender),
    prior(cauchy(0, 10), class = sd),
    prior(cauchy(0, 10), class = sigma),
    prior(lkj(2), class = cor)
    )

bmod5 <- brm(
    distance ~ gender + (1|subj) + (1 + gender|vowel) + (1|subj:vowel),
    data = indo, family = gaussian(),
    prior = prior5,
    warmup = 2000, iter = 10000
    )
```

Estimates of this model are summarised in Table 6. From this table, we first notice that the more varying effects we add, the more the model is uncertain about the estimation of α and β, which can be explained in the same way as in section 2.2. Second, we see the opposite pattern for σ_e, the residuals standard deviation, which has decreased by a considerable amount compared to the first model, indicating a better fit.

Table 6

Posterior mean, standard error, 95% credible interval and $\hat{R}$ statistic for each parameter of model bmod5 with a varying intercept and a varying slope by vowel and a varying intercept for the interaction between subject and vowel.

parameter	mean	SE	lower bound	upper bound	Rhat
α	0.163	0.038	0.087	0.236	1.000
β	-0.042	0.030	-0.100	0.018	1.000
$\sigma_{subject}$	0.012	0.009	0.001	0.033	1.000
$\sigma_{subject:vowel}$	0.024	0.005	0.016	0.034	1.000
$\sigma_{\alpha_{vowel}}$	0.070	0.046	0.029	0.183	1.000
$\sigma_{\beta_{vowel}}$	0.050	0.038	0.013	0.144	1.000
ρ	-0.433	0.380	-0.946	0.454	1.000
σ_e	0.085	0.001	0.082	0.088	1.000

3 Model comparison

Once we have built a set of models, we need to know which model is the more accurate and should be used to draw conclusions. It might be a little tricky to select the model that has the better absolute fit on the actual data (using for instance R^2), as this model will not necessarily perform as well on new data. Instead, we might want to choose the model that has the best predictive abilities, that is, the model that performs the best when it comes to predicting data that have not yet been observed. We call this ability the out-of-sample predictive performance of the model (McElreath, 2016). When additional data is not available, cross-validation techniques can be used to obtain an approximation of the model's predictive abilities, among which the Bayesian leave-one-out-cross-validation (LOO-CV, Vehtari, Gelman, & Gabry, 2017). Another useful tool, and asymptotically equivalent to the

LOO-CV, is the Watanabe Akaike Information Criterion (WAIC, Watanabe, 2010), which can be conceived as a generalisation of the Akaike Information Criterion (AIC, Akaike, 1974)[8].

Both WAIC and LOO-CV indexes are easily computed in `brms` with the `WAIC` and the `LOO` functions, where n models can be compared with the following call: `LOO(model1, model2, ..., modeln)`. These functions also provide an estimate of the uncertainty associated with these indexes (in the form of a SE), as well as a difference score ΔLOOIC, which is computed by taking the difference between each pair of information criteria. The `WAIC` and the `LOO` functions also provide a SE for these delta values (ΔSE). A comparison of the five models we fitted can be found in Table 7.

[8] More details on model comparison using cross-validation techniques can be found in Nicenboim and Vasishth (2016). See also Gelman, Hwang, and Vehtari (2014) for a complete comparison of information criteria.

Table 7

Model comparison with LOOIC.

Model	LOOIC	SE	ΔLOOIC	ΔSE	right side of the formula
bmod5	-3600.29	68.26	0.00	0.00	gender + (1 | subj) + (1 + gender | vowel) + (1 | subj:vowel)
bmod4	-3544.66	66.92	55.63	14.94	gender + (1 | subj) + (1 + gender | vowel)
bmod3	-3484.21	67.15	116.08	20.22	gender + (1 | subj) + (1 | vowel)
bmod2	-3119.41	65.32	480.88	39.50	gender + (1 | subj)
bmod1	-3103.43	66.72	496.86	40.52	gender

We see from Table 7 that `bmod5` (i.e., the last model) is performing much better than the other models, as it has the lower LOOIC. We then based our conclusions (see last section) on the estimations of this model. We also notice that each addition to the initial model brought improvement in terms of predictive accuracy, as the set of models is ordered from the first to the last model. This should not be taken as a general rule though, as successive additions made to an original model could also lead to *overfitting*, corresponding to a situation in which the model is over-specified in regards to the data, which makes the model good to explain the data at hand, but very bad to predict non-observed data. In such cases, information criteria and indexes that rely exclusively on goodness-of-fit (such as R^2) would point to different conclusions.

4 Comparison of brms and lme4 estimations

Figure 7 illustrates the comparison of `brms` (Bayesian approach) and `lme4` (frequentist approach) estimates for the last model (`bmod5`), fitted in `lme4` with the following command.

```
lmer_model <- lmer(
    distance ~ gender + (1|subj) + (1 + gender|vowel) + (1|subj:vowel),
    REML = FALSE, data = indo
    )
```

Densities represent the posterior distribution as estimated by `brms` along with 95% credible intervals, while the crosses underneath represent the *maximum likelihood estimate* (MLE) from `lme4` along with 95% confidence intervals, obtained with parametric bootstrapping.

We can see that the estimations of `brms` and `lme4` are for the most part very similar. The differences we observe for $\sigma_{\alpha_{vowel}}$ and $\sigma_{\beta_{vowel}}$ might be explained by the skewness of the posterior distribution. Indeed, in these cases (i.e., when the distribution is not symmetric), the mode of the distribution would better coincide with the `lme4` estimate. This figure also

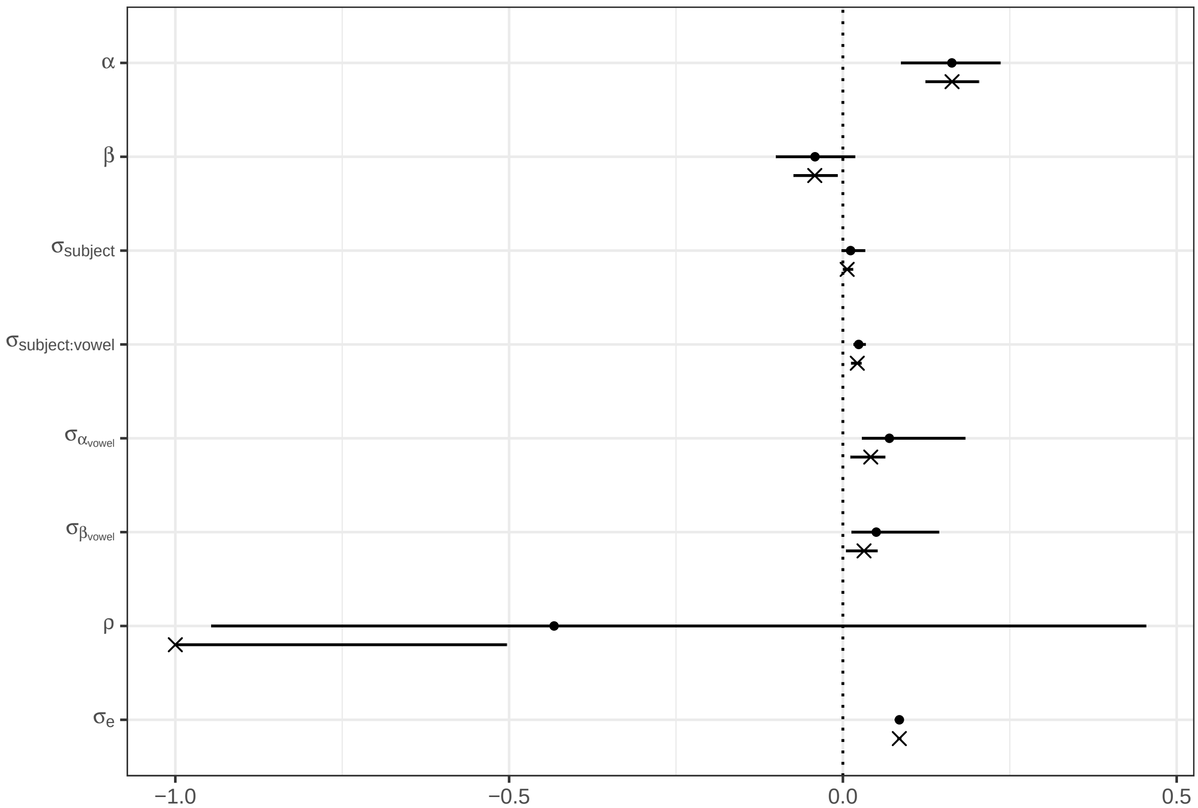

Figure 7. Comparison of estimations from `brms` and `lme4`. Dots represent means of posterior distribution along with 95% CrIs, as estimated by the `bmod5` model. Crosses represent estimations of `lme4` along with bootstrapped 95% CIs.

illustrates a limitation of frequentist MLMs that we discussed in the first part of the current paper. If we look closely at the estimates of `lme4`, we can notice that the MLE for the correlation ρ is at its boundary, as $\rho = -1$. This might be interpreted in (at least) two ways. The first interpretation is what Eager and Roy (2017) call the *parsimonious convergence hypothesis* (PCH) and consists in saying that this aberrant estimation is caused by the over-specification of the random structure (e.g., Bates et al., 2015a). In other words, this would correspond to a model that contains too many varying effects to be “supported” by a certain dataset (but this does not mean that with more data, this model would not be a correct model). However, the PCH has been questioned by Eager and Roy (2017), who have shown that under conditions of unbalanced datasets, non-linear models fitted with `lme4` provided more prediction errors than Bayesian models fitted with `Stan`. The second interpretation considers failures of convergence as a problem of frequentist MLMs *per se*, which is resolved in the Bayesian framework by using weakly informative priors (i.e., the LKJ prior) for the correlation between varying effects (e.g., Eager & Roy, 2017; Nicenboim & Vasishth, 2016), and by using the full posterior for inference.

One feature of the Bayesian MLM in this kind of situation is to provide an estimate of the correlation that incorporates the uncertainty caused by the weak amount of data (i.e., by widening the posterior distribution). Thus, the `brms` estimate of the correlation coefficient has its posterior mean at $\rho = -0.433$, but this estimate comes with a huge uncertainty, as expressed by the width of the credible interval (95% CrI = $[-0.946, 0.454]$).

5 Inference and conclusions

Regarding our initial question, which was to know whether there is a gender effect on vowel production variability in standard Indonesian, we can base our conclusions on several parameters and indices. However, the discrepancies between the different models we fitted deserve some discussion first. As already pointed out previously, if we had based our conclusions on the results of the first model (i.e., the model with constant effects only), we would have confidently concluded on a positive effect of gender. However, when we included the appropriate error terms in the model to account for repeated measurements by subject and by vowel, as well as for the by-vowel specific effect of gender, the large variability of this effect among vowels lead the model to adjust its estimation of β, resulting in more uncertainty about it. The last model then estimated a value of β = -0.04 with quite a large uncertainty (95% CrI = [-0.10, 0.02]), and considering 0 as well as some positive values as credible. This result alone makes it difficult to reach any definitive conclusion concerning the presence or absence of a gender effect on the variability of vowels pronunciation in Indonesian, and should be considered (at best) as suggestive.

Nevertheless, it is useful to recall that in the Bayesian framework, the results of our analysis is a (posterior) probability distribution which can be, as such, summarised in multiple ways. This distribution is plotted in Figure 8, which also shows the mean and the 95% CrI, as well as the proportion of the distribution below and above a particular value[9]. This figure reveals that 94.1% of the distribution is below 0, which can be interpreted as suggesting that there is a 0.94 probability that males have a lower mean formant distance than females (recall that female was coded as -0.5 and male as 0.5), given the data at hand, and the model.

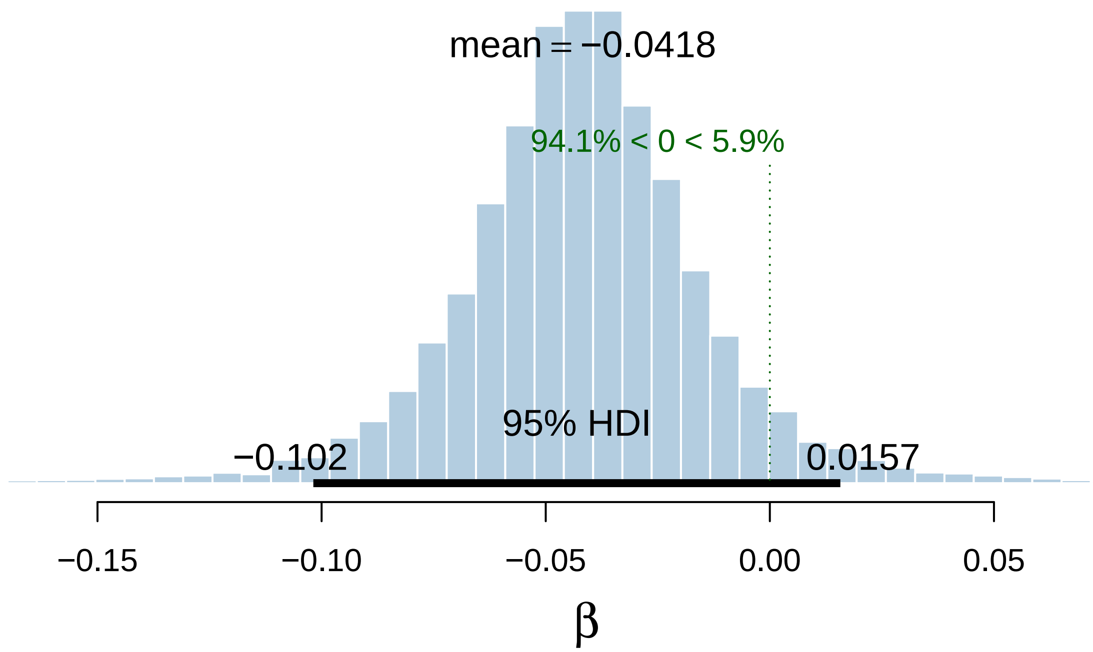

Figure 8. Histogram of posterior samples of the slope for `gender`, as estimated by the last model.

This quantity can be easily computed from the posterior samples:

```
post <- posterior_samples(bmod5) # extracting posterior samples
mean(post$b_gender < 0) # computing p(beta<0)
```

```
## [1] 0.940625
```

[9] We compare the distribution with 0 here, but it should be noted that this comparison could be made with whatever value.

Of course, this estimate can (and should) be refined using more data from several experiments, with more speakers. In this line, it should be pointed out that `brms` can easily be used to extend the multilevel strategy to meta-analyses (e.g., Bürkner et al., 2017; Williams & Bürkner, 2017). Its flexibility makes it possible to fit multilevel hierarchical Bayesian models at two, three, or more levels, enabling researchers to model the heterogeneity between studies as well as dependencies between experiments of the same study, or between studies carried out by the same research team. Such a modelling strategy is usually equivalent to the ordinary frequentist random-effect meta-analysis models, while offering all the benefits inherent to the Bayesian approach.

Another useful source of information comes from the examination of effects sizes. One of the most used criteria is Cohen's d standardized effect size, that expresses the difference between two groups in terms of their pooled standard deviation:

$$\text{Cohen's d} = \frac{\mu_1 - \mu_2}{\sigma_{pooled}} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{2}}}$$

However, as the total variance is partitioned into multiple sources of variation in MLMs, there is no unique way of computing a standardised effect size. While several approaches have been suggested (e.g., dividing the mean difference by the standard deviation of the residuals), the more consensual one involves taking into account all of the variance sources of the model (Hedges, 2007). One such index is called the δ_t (where the t stands for "total"), and is given by the estimated difference between group means, divided by the square root of the sum of all variance components:

$$\delta_t = \frac{\beta}{\sqrt{\sigma^2_{subject} + \sigma^2_{subject:vowel} + \sigma^2_{\alpha_{vowel}} + \sigma^2_{\beta_{vowel}} + \sigma^2}}$$

As this effect size is dependent on the parameters estimated by the model, one can derive a probability distribution for this index as well. This is easily done in `R`, computing it from the posterior samples:

```
delta_t <-
    # extracting posterior samples from bmod5
    posterior_samples(bmod5, pars = c("^b_", "sd_", "sigma") ) %>%
    # taking the square of each variance component
    mutate_at(.vars = 3:7, .funs = funs(.^2) ) %>%
    # dividing the slope estimate by the square root of the sum of
    # all variance components
    mutate(delta = b_gender / sqrt(rowSums(.[3:7]) ) )
```

This distribution is plotted in Figure 9, and reveals the large uncertainty associated with the estimation of δ_t.

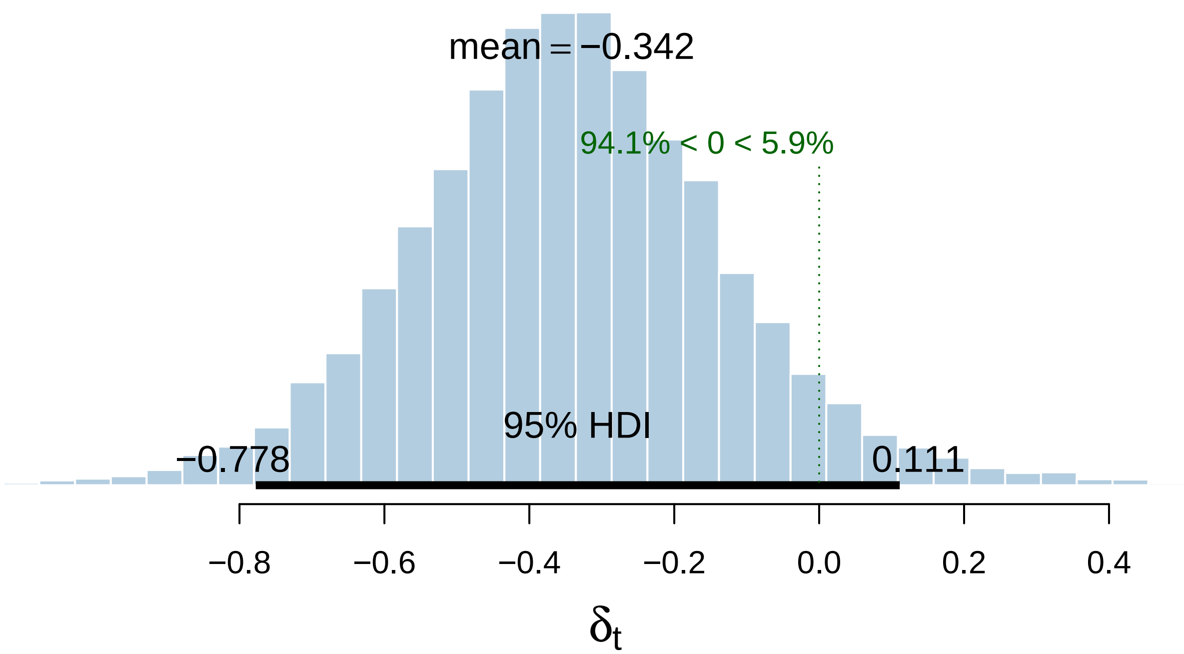

Figure 9. Posterior distribution of δ_t.

In the same fashion, undirected effect sizes (e.g., R^2) can be computed directly from the posterior samples, or included in the model specification as a parameter of the model, in a way that at each iteration of the MCMC, a value of the effect size is sampled, resulting in an estimation of its full posterior distribution (see for instance Gelman & Pardoe, 2006 for measures of explained variance in MLMs and @Marsman2017 for calculations in ANOVA designs). A Bayesian version of the R^2 is also available in `brms` using the `bayes_R2` method, for which the calculations are based on Gelman, Goodrich, Gabry, and Ali (2017).

```
bayes_R2(bmod5)
```

```
##    Estimate  Est.Error   2.5%ile  97.5%ile
## R2 0.295614 0.01589917 0.2635006 0.3262617
```

In brief, we found a weak effect of gender on vowel production variability in Indonesian (β = -0.04, 95% CrI = [-0.10, 0.02], δ_t = -0.34, 95% CrI = $[-0.78, 0.11]$), this effect being associated with a large uncertainty (as expressed by the width of the credible interval). This result seems to show that females tend to pronounce vowels with more variability than males, while the variation observed across vowels (as suggested by $\sigma_{\beta_{vowel}}$) suggests that there might exist substantial inter-vowel variability, that should be subsequently properly studied. A follow-up analysis specifically designed to test the effect of gender on each vowel should help better describe inter-vowel variability (we give an example of such an analysis in the supplementary materials).

To sum up, we hope that this introductive tutorial has helped the reader to understand the foundational ideas of Bayesian MLMs, and to appreciate how straightforward the interpretation of the results is. Moreover, we hope to have demonstrated that although Bayesian data analysis may still sometimes (wrongfully) sound difficult to grasp and to use, the development of recent tools like `brms` helps to build and fit Bayesian MLMs in an intuitive way. We believe that this shift in practice will allow more reliable statistical inferences to be drawn from empirical research.

6 Supplementary materials

Supplementary materials, reproducible code and figures are available at: osf.io/dpzcb. A lot of useful packages have been used for the writing of this paper, among which the `papaja` and `knitr` packages for writing and formatting (Aust & Barth, 2017; Xie, 2015), the `ggplot2`, `viridis`, `ellipse`, `BEST`, and `ggridges` packages for plotting (Garnier, 2017; Kruschke & Meredith, 2017; Murdoch & Chow, 2013; Wickham, 2009; Wilke, 2017), as well as the `tidyverse` and `broom` packages for code writing and formatting (Robinson, 2017; Wickham, 2017).

Acknowledgements

We thank Brice Beffara for helpful comments on a previous version of this manuscript, as well as Shravan Vasishth and one anonymous reviewer for insightful suggestions during the review process.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/tac.1974.1100705

Aust, F., & Barth, M. (2017). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. https://doi.org/10.1037/h0020412

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015a). *Parsimonious mixed models*. Retrieved from https://arxiv.org/pdf/1506.04967.pdf

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015b). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bürkner, P.-C. (2017a). *Advanced bayesian multilevel modeling with the R package brms.* Retrieved from https://arxiv.org/pdf/1705.11123

Bürkner, P.-C. (2017b). brms: An R package for bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P.-C., Williams, D. R., Simmons, T. C., & Woolley, J. D. (2017). Intranasal oxytocin may improve high-level social cognition in schizophrenia, but not social cognition or neurocognition in general: A multilevel Bayesian meta-analysis. *Schizophrenia Bulletin*, *43*(6), 1291–1303. https://doi.org/10.1093/schbul/sbx053

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* New York: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. https://doi.org/10.1177/1745691611406920

Eager, C., & Roy, J. (2017). *Mixed effects models are sometimes terrible.* Retrieved from https://arxiv.org/pdf/1701.04858.pdf

Garnier, S. (2017). *viridis: Default color maps from 'matplotlib'.* Retrieved from https://CRAN.R-project.org/package=viridis

Gelman, A. (2005). Analysis of variance — why it is more important than ever. *The Annals of Statistics*, *33*(1), 1–53. https://doi.org/10.1214/009053604000001048

Gelman, A. (2006). Prior distributions for variance parameter in hierarchical models.

Bayesian Analysis, *1*(3), 515–534. https://doi.org/10.1214/06-ba117a

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis, third edition*. CRC Press.

Gelman, A., Goodrich, B., Gabry, J., & Ali, I. (2017). *R-squared for Bayesian regression models*. Retrieved from https://github.com/jgabry/bayes_R2/blob/master/bayes_R2.pdf

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, New York.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*, 189–211. https://doi.org/10.1080/19345747.2011.618213

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Gelman, A., & Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, *48*(2), 241–251. https://doi.org/10.1198/004017005000000517

Gelman, A., & Rubin, D. B. (1992). Inference from Iterative Simuation Using Multiple Sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. *The Sage Handbook of Methodology for the Social Sciences*, 391–408. https://doi.org/10.4135/9781412986311.n21

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and*

Behavioral Statistics, *32*(4), 341–370. https://doi.org/10.3102/1076998606298043

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

Janssen, D. P. (2012). Twice random, once mixed: Applying mixed models to simultaneously analyze random effects of language and participants. *Behavior Research Methods*, *44*(1), 232–247. https://doi.org/10.3758/s13428-011-0145-1

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*, 601–625. https://doi.org/10.1146/annurev-psych-122414-033702

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research.* (pp. 61–91). Washington, DC: APA Book. https://doi.org/10.1037/10693-003

Kruschke, J. K. (2015). *Doing Bayesian data analysis, Second Edition: A tutorial with R, JAGS, and Stan.* Burlington, MA: Academic Press / Elsevier.

Kruschke, J. K., & Liddell, T. M. (2017a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, 1–23. https://doi.org/10.3758/s13423-017-1272-1

Kruschke, J. K., & Liddell, T. M. (2017b). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 1–29. https://doi.org/10.3758/s13423-016-1221-4

Kruschke, J. K., & Meredith, M. (2017). *BEST: Bayesian estimation supersedes the t-test.* Retrieved from https://CRAN.R-project.org/package=BEST

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical–significance tests are not. *Theory & Psychology*, *22*(1), 67–90. https://doi.org/10.1177/0959354311429854

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Marsman, M., Waldorp, L., Dablander, F., & Wagenmakers, E.-J. (2017). *Bayesian estimation of explained variance in ANOVA designs*. Retrieved from http://maartenmarsman.com/wp-content/uploads/2017/04/MarsmanEtAl_R2.pdf

McCloy, D. R. (2014). Phonetic effects of morphological structure in Indonesian vowel reduction. In *Proceedings of meetings on acoustics* (Vol. 12, pp. 1–14). https://doi.org/10.1121/1.4870068

McCloy, D. R. (2016). *phonR: Tools for phoneticians and phonologists*. Retrieved from https://cran.r-project.org/web/packages/phonR/

McElreath, R. (2016). *Statistical Rethinking*. Chapman; Hall/CRC.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2015). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*, 103–123. https://doi.org/10.3758/s13423-015-0947-8

Murdoch, D., & Chow, E. D. (2013). *ellipse: Functions for drawing ellipses and ellipse-like confidence regions*. Retrieved from https://CRAN.R-project.org/package=ellipse

Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational Ideas – Part II. *Language and Linguistics Compass*, *10*(11), 591–613. https://doi.org/10.1111/lnc3.12207

Polson, N. G., & Scott, J. G. (2012). On the half-cauchy prior for a global scale parameter. *Bayesian Analysis*, *7*(4), 887–902. https://doi.org/10.1214/12-BA730

R Core Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Robinson, D. (2017). *broom: Convert statistical analysis objects into tidy data frames.* Retrieved from https://CRAN.R-project.org/package=broom

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, *38*(5), 2587–2619. https://doi.org/10.1214/10-AOS792

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Methods for Psychology*, *12*(3), 175–200. https://doi.org/10.20982/tqmp.12.3.p175

Stan Development Team. (2016). Stan modeling language users guide and reference manual. Retrieved from http://mc-stan.org

Trafimow, D., Amrhein, V., Areshenkoff, C. N., Barrera-Causil, C. J., Beh, E. J., Bilgiç, Y. K., … Marmolejo-Ramos, F. (2018). Manipulating the alpha level cannot cure significance testing. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00699

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*(5), 1413–1432. https://doi.org/10.1007/s11222-016-9696-4

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*, 3571–3594.

Watt, D., & Fabricius, A. (2002). Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1~F2 plane. *Leeds Working Papers in Linguistics and Phonetics*, *9*(9), 159–173.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H. (2017). *tidyverse: Easily install and load 'tidyverse' packages*. Retrieved from https://CRAN.R-project.org/package=tidyverse

Wilke, C. O. (2017). *ggridges: Ridgeline plots in 'ggplot2'*. Retrieved from https://CRAN.R-project.org/package=ggridges

Williams, D. R., & Bürkner, P.-C. (2017). Psychoneuroendocrinology Effects of intranasal oxytocin on symptoms of schizophrenia: A multivariate Bayesian meta-analysis. *Psychoneuroendocrinology*, *75*, 141–151. https://doi.org/10.1016/j.psyneuen.2016.10.013

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.name/knitr/