Journal of Phonetics 71 (2018) 147–161

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Special Issue: Emerging Data Analysis in Phonetic Sciences, eds. Roettger, Winter & Baayen

Bayesian data analysis in the phonetic sciences: A tutorial introduction

Shravan Vasishth [a,*], Bruno Nicenboim [a], Mary E. Beckman [b,*], Fangfang Li [c], Eun Jong Kong [d]

[a] Department of Linguistics, University of Potsdam, Germany
[b] Department of Linguistics, Ohio State University, United States
[c] Department of Psychology, University of Lethbridge, Canada
[d] Department of English, Korea Aerospace University, Republic of Korea

ARTICLE INFO

Article history:
Received 4 October 2017
Received in revised form 6 July 2018
Accepted 25 July 2018

Keywords:
Bayesian data analysis
Linear mixed models
Voice onset time
Gender effects
Vowel duration

ABSTRACT

This tutorial analyzes voice onset time (VOT) data from Dongbei (Northeastern) Mandarin Chinese and North American English to demonstrate how Bayesian linear mixed models can be fit using the programming language Stan via the R package `brms`. Through this case study, we demonstrate some of the advantages of the Bayesian framework: researchers can (i) flexibly define the underlying process that they believe to have generated the data; (ii) obtain direct information regarding the uncertainty about the parameter that relates the data to the theoretical question being studied; and (iii) incorporate prior knowledge into the analysis. Getting started with Bayesian modeling can be challenging, especially when one is trying to model one's own (often unique) data. It is difficult to see how one can apply general principles described in textbooks to one's own specific research problem. We address this barrier to using Bayesian methods by providing three detailed examples, with source code to allow easy reproducibility. The examples presented are intended to give the reader a flavor of the process of model-fitting; suggestions for further study are also provided. All data and code are available from: https://osf.io/g4zpv.

1. Introduction

In phonetics and other related areas of the language sciences, the vast majority of studies are designed to elicit several data points from each participant for each level of the linguistic variable of interest. This design poses difficulties for classic ANOVA models, which can accommodate only one random effect at a time, so that either the sets of data-points for each participant or the sets of data-points for each item must be replaced with the mean values (Clark, 1973). Over the last two decades, phoneticians have addressed these difficulties by turning to other methods, and linear mixed models—sometimes referred to as multilevel or hierarchical linear models—have become a standard tool, perhaps *the* standard tool for analyzing repeated measures data. The `lme4` package (Baayen, Davidson, & Bates, 2008; Bates, Maechler, Bolker, & Walker, 2015b; Pinheiro & Bates, 2000) in R has greatly simplified model specification and data analysis for repeated measures designs. Even more recently, a Bayesian alternative to frequentist linear mixed models has become available, largely due to the emergence of a new programming language, Stan (version 1.17.3) (Stan Development Team, 2017b). In this article, we provide a tutorial introduction to fitting Bayesian linear mixed models. In order to make it easy for the newcomer to Bayesian data analysis to fit models, we use the popular and powerful R package `brms`, version 2.1.9 (Bürkner, 2016), which uses `lme4` syntax that researchers in linguistics and psychology are familiar with.

Fitting Bayesian models takes more time and effort than their frequentist analogues. Why bother to learn this relatively complex approach? We feel that there are several important advantages to fitting Bayesian models. Perhaps the most important one is that it gives us a degree of flexibility in defining models that is difficult to match with frequentist tools (Lee, 2011; Nicenboim & Vasishth, 2016). We discuss an example below. A second advantage of Bayesian modeling is that we can focus our attention on quantifying our uncertainty about the magnitude of an effect. Instead of drawing a conclusion like "gender affects voice onset time", using the Bayesian framework we can identify a credible interval of plausible values representing the effect. In other words, we can present a probability distribution of plausible values, instead of focusing on whether a particular confidence interval does or does not contain the value 0. Such quantitative summaries of an effect tell us much more

* Corresponding authors.
E-mail addresses: vasishth@uni-potsdam.de (S. Vasishth), beckman.2@osu.edu (M.E. Beckman).

https://doi.org/10.1016/j.wocn.2018.07.008
0095-4470/© 2018 Elsevier Ltd. All rights reserved.

about the research question than binary statements like “effect present” or “effect absent.” A third advantage of Bayesian data analysis is that we can incorporate prior knowledge or beliefs in the model in an explicit way with the use of so-called informative priors. Such a use of priors is not widespread, but could be a powerful tool for building on what we already know about a research question. Finally, frequentist tools like `lme4` can run into convergence problems when an attempt is made to fit a “maximal” random-effects structure (Barr, Levy, Scheepers, & Tily, 2013).[1] Bayesian linear mixed models will always converge once so-called regularizing priors are used; we explain this point below. In this tutorial, we will provide an informal introduction to Bayesian data analysis, and then present three examples involving retrospective measurements of productions in a large cross-linguistic phonetic corpus. These examples are intended to provide a practical first entry to Bayesian data analysis. We do not aim to cover all aspects of Bayesian modeling here, but suggestions for further reading are provided at the end. In our examples, we will focus on (generalized) linear mixed models (Pinheiro & Bates, 2000; Baayen et al., 2008; Bates et al., 2015b), because they are a standard tool today in experimental research in linguistics and the psychological sciences. We assume in this paper that the reader knows how to fit linear mixed models using the R package `lme4` (Bates et al., 2015b). Accessible introductions to linear mixed models are in Gelman and Hill (2007) and McElreath (2016).

All data and code are available from https://osf.io/g4zpv. The additional code examples provided there cover some further issues not discussed in this paper.

2. An informal introduction to Bayesian data analysis

Consider a simple case where we carry out an experiment in which we measure voice onset time in milliseconds in recordings of word-initial stops such as Mandarin /t^h/ and /k^h/ produced by male and female participants. Participants in each gender category are asked to produce multiple stop-initial words, resulting in repeated measurements of VOT from each participant. The first few lines and last few lines of an example data-frame is shown in Listing 1.

Listing 1: Example data-set from English.

```
subject    item          gender    VOT
F01        kh^.l&r       0.5       105
F01        kh^.tIxN      0.5       120
F01        khA9          0.5       104
F01        khE.tS&p      0.5       127
F01        khek          0.5       141
F01        khev          0.5       106
...
M20        thu.n&        −0.5      101
M20        thub          −0.5      66
M20        thuT          −0.5      67
M20        twhI.stIxd    −0.5      69
M20        twhi.z&rz     −0.5      93
M20        twhIn         −0.5      85
```

For $i = 1, \ldots, I$ participants and $j = 1, \ldots, J$ items, we often want to fit a so-called varying intercepts and varying slopes linear mixed model of the type specified in (1) – the equation for a frequentist linear mixed model for the effect of gender on VOT. A notational convention we use here: a varying intercept always has index 0, and a varying slope has index 1 (or higher, in the case the case of multiple regression). Thus, a varying intercept for item j is written $w_{0,j}$ and a varying slope is written $w_{1,j}$ (or $w_{2,j}$, for a second predictor, and so on). Fixed intercepts and slopes also have the same numerical subscript convention of 0 for intercepts, and 1 (or a higher index) for the slope (with increasing numbers in the case of multiple predictors).

Using these notational conventions, a frequentist linear mixed model for the effect of gender on VOT could be specified as follows:

$$VOT_{ij} = \beta_0 + u_{0,i} + w_{0,j} + (\beta_1 + w_{1,j}) \times \text{gender}_{ij} + \epsilon_{ij} \tag{1}$$

Assuming that the categorical variable is sum-coded (e.g., $+0.5$ for female, -0.5 for male), the intercept β_0 represents the grand mean, and the slope β_1 the difference in means between the two levels of gender. These are the so-called fixed effects. The terms $u_{0,i}$ and $w_{0,j}$ are, respectively, the by-participant and by-item adjustments to the intercept coefficient β_0, and $w_{1,j}$ is the by-item adjustment to the slope term for gender, β_1. The varying intercepts for subjects, $u_{0,i}$, are assumed to be distributed as $Normal(0, \sigma_{u0})$; similarly, the varying intercepts for items $w_{0,j}$ have the distribution $Normal(0, \sigma_{w0})$, and the varying slopes for item by gender, $w_{1,j}$ have the distribution $Normal(0, \sigma_{w1})$. The residual error, ϵ, is assumed to have the distribution $Normal(0, \sigma_e)$. Finally, the varying intercepts and slopes for item, $w_{0,j}, w_{1,j}$ are assumed to have correlation ρ_w. In `lme4` syntax, the above model corresponds to the following (`datE_stops` refers to the data frame):

```
lmer(VOT ~ 1 + gender + (1 | subject) + (1 + gender |
item), dat = datE_stops)
```

Because `lme4` assumes an intercept term, the `1 +` can be omitted, as in:

```
lmer(VOT ~ gender + (1 | subject) + (gender | item),
dat = datE_stops)
```

The above model requires the estimation of the parameters listed in 2. (Note that in Bayesian linear mixed models, u_{0i}, w_{0j}, w_{1j} are also parameters; but these are not of primary interest in studies such as this example which address questions only about group effects rather than about patterns of differences across individuals or across items.)

$$\beta_0, \beta_1, \sigma_{u0}, \sigma_{w0}, \sigma_{w1}, \rho_w, \sigma_e \tag{2}$$

Again, the intercept β_0 represents the grand mean VOT. Note that it does not make sense to fit varying slopes for gender by participants in this model because gender is a between-participants factor (i.e., we can’t investigate the effect of gender on the participants). Gender is, however, a within-items factor, so varying slopes for gender can be fit by items (i.e., we *can* investigate the effect of gender on the items).

In the frequentist framework, we would just need to run the `lmer` function as shown above. However, in the Bayesian linear mixed model, some more work is needed before we can

[1] For issues relating to the fitting of “maximal models” see the discussions in Bates, Kliegl, Vasishth, and Baayen (2015a), Baayen, Vasishth, Kliegl, and Bates (2017) and Matuschek, Kliegl, Vasishth, Baayen, and Bates (2017).