MAX PLANCK INSTITUTE FOR PSYCHOLINGUISTICS

A comparison of frequentist and Bayesian models of language variation

The problems of priors and sample size

NATALIA LEVSHINA

LANGUAGE in INTERACTION

NWO

Introduction

- Frequentist statistics aka classical statistics aka sampling theory statistics aka maximum likelihood estimation (especially in the context of regression modelling) aka null hypothesis significance testing…

- “whereas the 20th century was dominated by NHST, the 21st century is becoming Bayesian” (Kruschke 2011: 272)

- Examples of studies:
  - Vasishth et al. (2013) on processing of Chinese relative clauses
  - Scrivner (2015): VO and OV word order patterns in Latin and Old French infinitival complements,
  - Levshina’s (2016) multifactorial analysis of English permissive constructions
  - …

MAX PLANCK

Outline

1. Advantages and main principles of Bayesian inference

2. A case study of help + (to) Infinitive

   - Large sample:
     - A maximum likelihood model
     - A Bayesian model
   - Small sample:
     - A maximum likelihood model
     - A Bayesian model with strong recycled priors

3. Conclusions

MAX PLANCK

Advantages of Bayesian inference

- An opportunity to test the research hypothesis directly, instead of trying to reject the null hypothesis
- It does not rely on p-values and does not encourage binary decisions (Accept – Reject), enriching our knowledge about the impact of the contextual factors.
  - Less p-hacking!
- One can use information from previous research as priors for subsequent models.

MAX PLANCK

What makes Bayesian statistics special?

- Frequentist inference involves null hypothesis significance testing:
  - The null hypothesis is rejected when the probability of observing the test statistic and more extreme values under the null hypothesis is smaller than some pre-determined level (usually 0.05).
  - That is, we are interested in the likelihood of data given the null hypothesis, or $P(Data|H_0)$.
- Bayesian inference allows us to estimate the probability of the research hypothesis given the data, or $P(H_1|Data)$.

How to get P(H|Data)?

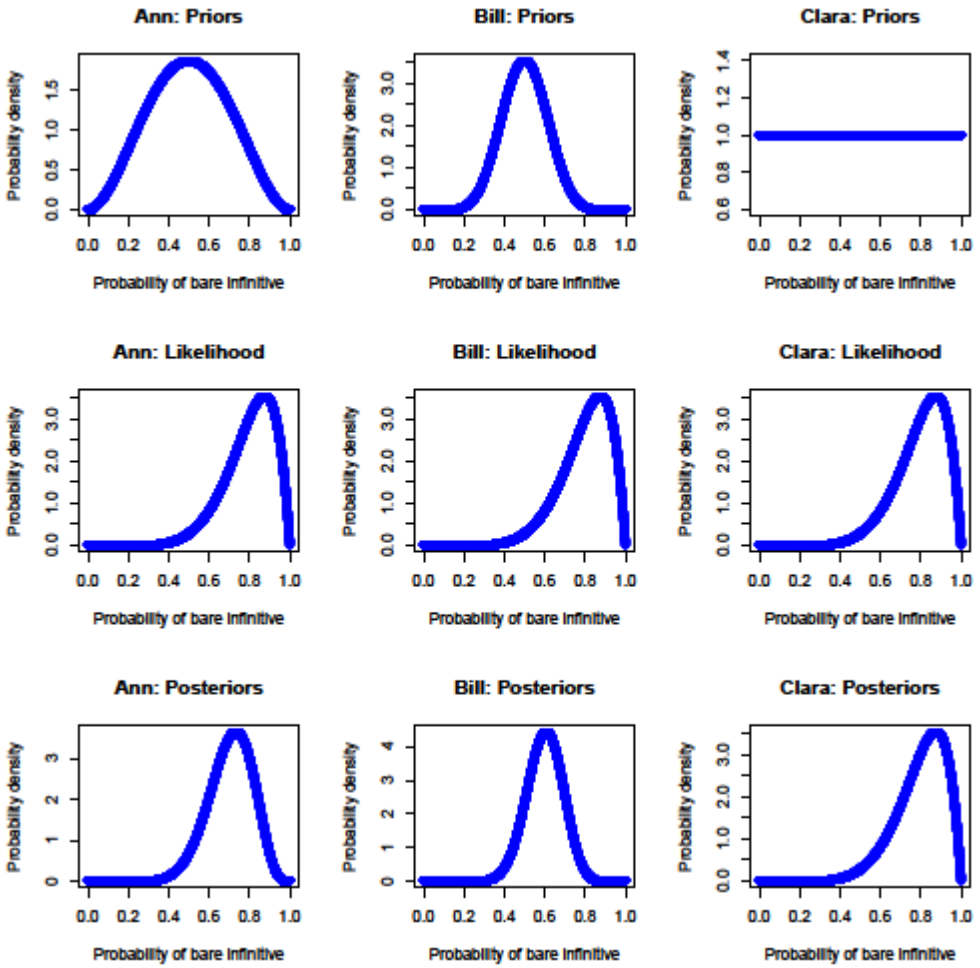
From Bayes rule, it follows that

$$P(H|Data) \propto P(Data|H)\, P(H)$$

or

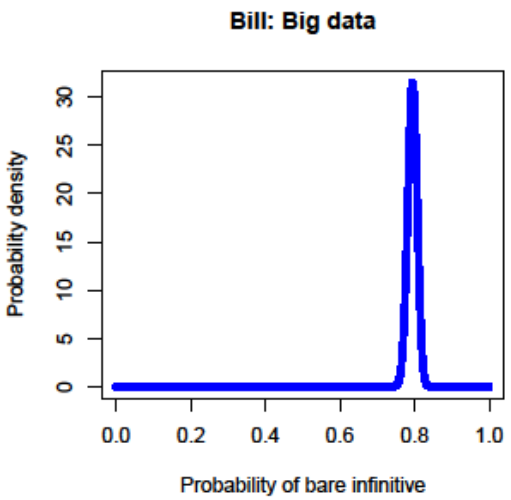
Posterior $\propto$ Prior x Likelihood

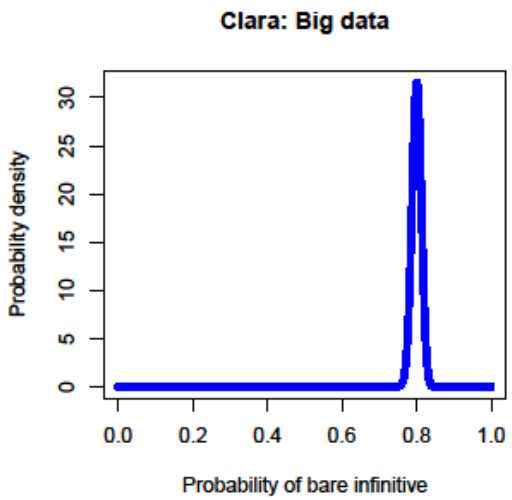
An imaginary example



Big data

- When your dataset is large, the frequentist and Bayesian methods will converge, especially if your priors are weakly or moderately informative.
- It makes sense to perform a prior sensitivity analysis in order to see whether different types of priors have an effect.





Priors help!

- Priors allow the researcher to use smaller samples (good if data are costly, e.g. spoken and multimodal corpora, rare typological phenomena, and experimental linguistics).

- Recycling one's knowledge can help us to overcome the recent crisis of reproducibility (Goodman et al. 2016): the plausibility of the old findings can be estimated in the light of new data when these findings are incorporated into a new model (van de Schoot et al. 2014).

- Priors also help to avoid problems, such as loss of statistical power, overfitting and convergence issues, which often arise when one fits generalized mixed-effect models with complex structure.

MAX PLANCK

Weakly informative priors

- There are practical benefits of using weakly informative priors, similar to Ann's.
- They help to exclude unrealistic values of coefficients in logistic regression, which can be useful in the situations of data sparseness.
- Gelman et al. (2008): it is reasonable to specify that the values of predictor coefficients in a logistic regression model are very likely to be in the interval between -5 and 5, and unlikely to be outside that range.

Markov Chain Monte Carlo

- How to combine the priors and the data and obtain the posteriors? We can approximate them by sampling a large number of representative points from the posterior distribution with the help of a Markov Chain Monte Carlo (MCMC) algorithm.
  - A Monte Carlo simulation is any simulation that draws random values from a distribution.
  - A Markov Chain process is a random walk when the next step does not depend on the steps before the current position.

- Amazingly, an MCMC is supposed to converge to the target distribution regardless of where the initial position was. In Bayesian regression modelling, we usually create several chains, which consist of thousands of iterations.

- Disadvantage: computationally costly.

Summarizing the posterior

Based on 4000 posterior samples

MAX PLANCK

brms

- R package (Bürkner 2017)
- It uses a syntax very similar to the expressions used in the package lme4.
- Based on Stan, a statistical platform in C++ and programming language for Bayesian and other types of inference.

Outline

1. Advantages and main principles of Bayesian inference

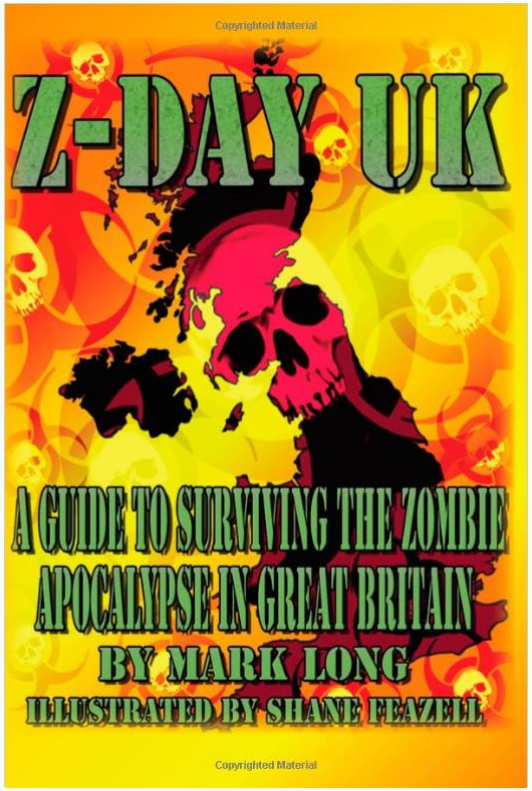
2. A case study of help + (to) Infinitive

   - Large sample:
     - A maximum likelihood model
     - A Bayesian model
   - Small sample:
     - A maximum likelihood model
     - A Bayesian model with strong recycled priors

3. Conclusions

MAX PLANCK

Z-DAY UK



A GUIDE TO SURVIVING THE ZOMBIE APOCALYPSE IN GREAT BRITAIN

BY MARK LONG

ILLUSTRATED BY SHANE FEAZELL

IF THIS BOOK DOES NOT HELP YOU TO SURVIVE THE ZOMBIE APOCALYPSE, A FULL REFUND WILL BE GIVEN IN CASH OR SHOTGUN SHELLS....

This book is a practical guide to dealing with the worst case survival scenario: Z-day, the zombie apocalypse. Unlike other books, this addresses the specific problems of surviving the apocalypse in the United Kingdom.

It includes:

- How to survive if you have made no preparations prior to Z-day
- How to prepare
- Getting around including how to fly a light aircraft
- Working with other survivors
- How to get food and water
- How to prevent infection and treat wounds
- Useful equipment including weapons, improvised, found or made
- How to build a long term community including information on farming, wells, latrines, blacksmithing and other essential skills

While no one book can contain all that you need, this book gives you a practical core of knowledge on how to avoid becoming one of the walking dead and how to live in a very changed Britain.

ISBN 9781490389875

*Here is how items in your emergency kit can **help** you **survive** the zombie apocalypse.* https://www.redcross.ca/blog/2013/2/preparedness-101-zombie-apocalypse

Preparedness 101: Zombie Apocalypse

Posted February 01, 2013 by Jamie-Leigh Cuthbertson - Photo aficionado

With the release of a zombie rom-com movie this weekend and AMC's The Walking Dead returning to our television screens shortly, we thought it might be a great time to share our zombie preparedness tips. Fuelled by a blog about zombies and emergency preparedness from the US Centers for Disease Control and Prevention, North America has become almost obsessed with imagining how they would fare if the walking dead started roaming their neighbourhood.

Photo Credit: http://www.cdc.gov/phpr/zombies

Did you know that your emergency kit contains tools and supplies that can also be used in the event of an outbreak of th zombie virus? Here is how items in your emergency kit can help you survive the zombie apocalypse:

Previous research

- Descriptive studies (e.g. Rohdenburg 1996; Biber et al. 1999; Mair 2002; McEnery & Xiao 2005; Rohdenburg 2009).
- Multivariate models (Lohmann 2011; Levshina 2018)
- Multiple contextual factors + regional and stylistic variation!
- The case of “soft constraints” (Bresnan et al. 2001) studied by “probabilistic linguistics” (Bod, Hay and Jannedy 2003) and “probabilistic grammar” (Grafmiller et al. 2018)
- Cf. complex statistical models of language users’ behaviour in Gries 2003; Bresnan et al. 2007; Szmrecsanyi et al. 2016, etc.)

Form of *help*

- The base form *help* occurs the most frequently with the bare infinitive,
- The form *helping* shows the highest proportion of *to*-infinitives.

a. *It is his job to see through the contracts that will **help rebuild** Iraq.*
b. *Many are partners with South African companies on projects that are **helping to rebuild** the country's infrastructure.*

Distance between *help* and Infinitive

- The more words between help and the infinitive, the higher the chances of the *to*-infinitive:

  HUD will provide $70 million for 1,300 rental vouchers to ***help*** *people in public housing projects in Los Angeles, Chicago, Boston, Baltimore, and New York* ***to move*** *into surrounding middle-class and affluent suburbs.*

- This tendency can be explained by the principle of minimization of cognitive complexity:
  - "[i]n the case of more or less explicit grammatical options the more explicit one(s) will tend to be favoured in cognitively more complex environments" (Rohdenburg 1996: 151).

MAX PLANCK

Horror aequi

- A universal tendency to avoid repetition of identical elements.
- If the verb *help* is preceded by *to*, the second infinitive is usually without *to*:

> *With yoga, find a teacher who will make adjustments <u>to</u> **help** **prevent** injuries.*

Helpee

- The chances of the marked infinitive are higher if the Helpee is implicit:

a. […] *physical therapy has* ***helped*** *<u>me</u>* ***improve*** *my posture*…

b. His encouragement and guidance ***helped*** *[Ø]* ***to improve*** *my health and self-confidence.*

Data extraction

- Magazines section from the Corpus of Contemporary American English (COCA)
- All sentences with the forms *help*, *helps*, *helped* and *helping* in upper or lower case in the text-only version of the corpus.
- These instances were parsed syntactically with the Universal Dependencies parser using the R package *udpipe* (Strakov & Straková 2017).
- All sentences in which there was a lemma *help* with the part of speech VERB, and this lemma had a dependency *xcomp (*infinitival complements).
- A Python script to extract all subjects of *help* (the Helper), the objects of *help* (the Helpee) and other relevant contextual variables.

MAX PLANCK

Two samples

- A large sample with approximately 2300 occurrences. After cleaning the data manually and excluding some spurious hits, there were 2050 examples left.

- A small sample, which resulted in 400 occurrences after the manual cleaning (for additional manual annotation)

Variables

- Response variable (labelled as *Response* in the R code and output): the bare or *to*-infinitive.
- Year (*Year_new*): the year when the text was published, from 1990 to 2012.
    - In order to have an interpretable intercept value corresponding to 0, I subtracted 1990 from every number. As a result, I had numbers from 0 (1990) to 22 (2012).
    - The research hypothesis: the proportion of the *to*-infinitive slightly decreases with time.

Variables (continued)

- *Horror*: whether there is *to* immediately before *help* or not.
    - The presence of *to* in front of *help* is expected to decrease the chances of *to* before the next infinitive, especially if the distance between *help* and the infinitive (see below) is small.
- Log-transformed distance (*Distance_log*): the number of words between *help* and the infinitive, disregarding *to*.
    - According to the principle of cognitive complexity, I expected the chances of the *to*-infinitive increase with distance.

Variables (continued)

- Morphological form of *help* (*MorphForm*): *help*, *helps*, *helping* and *helped*.
    - I expected the highest chances of *to* after the form *helping;* the base form *help* is expected to have the highest chances of being used with the bare infinitive.

- *Helpee*: whether the Helpee is expressed explicitly in the sentence as a pronoun or nominal phrase or not.
    - One could expect the *to*-infinitive to be more likely when the Helpee is absent and less likely when the Helpee is present.

Variables (continued)

- The Helper's semantic class (*Helper*): animate (humans, organizations, animals), inanimate or missing.
  - *She even went through a rehabilitation program in an effort **to help** her **walk** again.*
  - I expected higher likelihood of the bare infinitive when the Helper was animate in comparison with inanimate Helpers.

- The individual verbs (*Verb*), which fill in the infinitival slot of the construction (random intercepts).

Stress patterns (only small sample)

- Schlüter (2003): the principle of rhythmic alternation, a prosodic tendency to avoid sequences of stressed syllables (so-called stress clash), as well as sequences of unstressed syllables (stress lapse).
- Wasow et al. (2015) on the use and omission of *to* in construction *All we want to do is (to) celebrate*.
- Written texts can have this effect as a spillover from spoken discourse.
- The variable is called *Stress* and has the values ‘Clash’, ‘Lapse’ and ‘Good’ (neither stress, nor clash).
- Manually annotated (a very time-consuming task!)

Outline

1. Advantages and main principles of Bayesian inference

2. A case study of help + (to) Infinitive

   - Large sample:
     - A maximum likelihood model
     - A Bayesian model
   - Small sample:
     - A maximum likelihood model
     - A Bayesian model with strong recycled priors

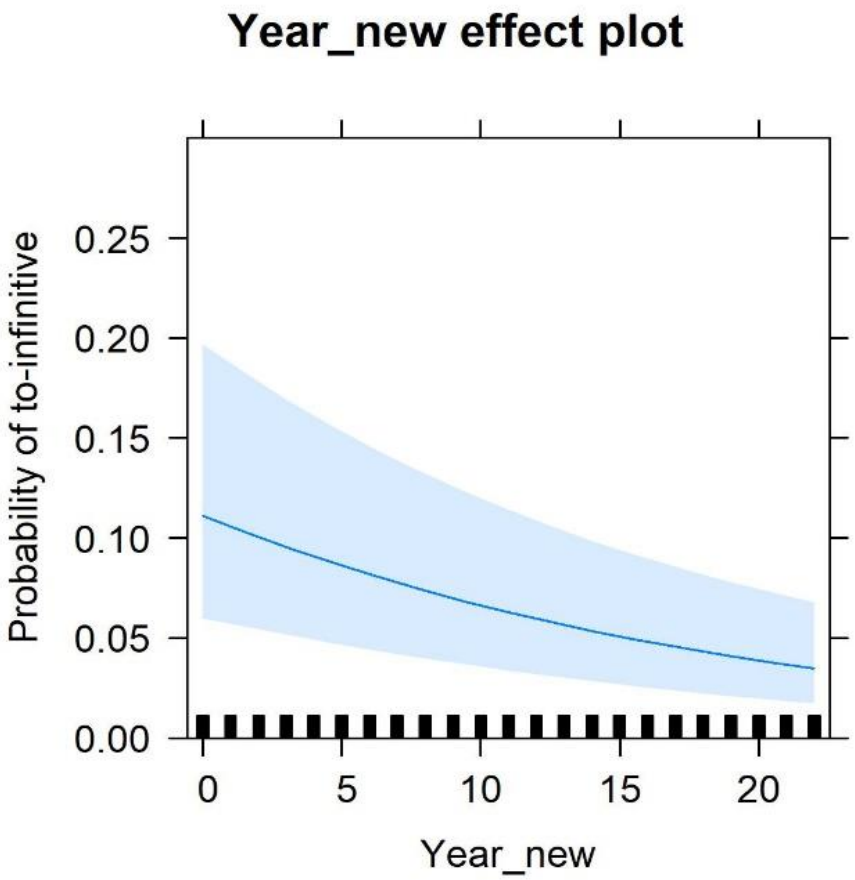
3. Conclusions

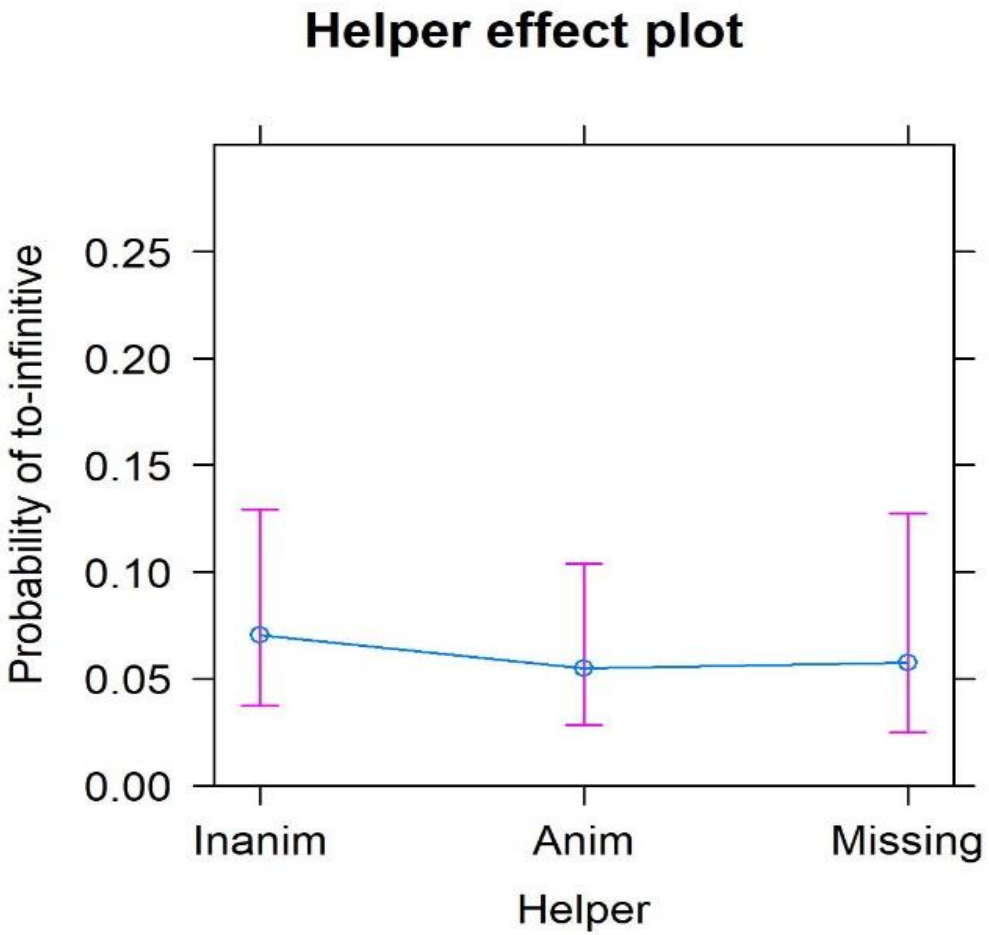
MAX PLANCK

Mixed-effects GLMM

```
Fixed effects:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -1.03992    0.17451  -5.959 2.54e-09 ***
Year_new            -0.05659    0.01048  -5.401 6.63e-08 ***
HorrorYes           -6.33335    1.58414  -3.998 6.39e-05 ***
Distance_log         0.83140    0.41989   1.980  0.04770 *
MorphFormhelped      0.58174    0.19855   2.930  0.00339 **
MorphFormhelping     3.26033    0.29139  11.189  < 2e-16 ***
MorphFormhelps       1.23052    0.24966   4.929 8.27e-07 ***
HelpeeYes           -0.94924    0.39858  -2.382  0.01724 *
HelperAnim          -0.26949    0.16123  -1.672  0.09462 .
HelperMissing       -0.21942    0.36459  -0.602  0.54729
```

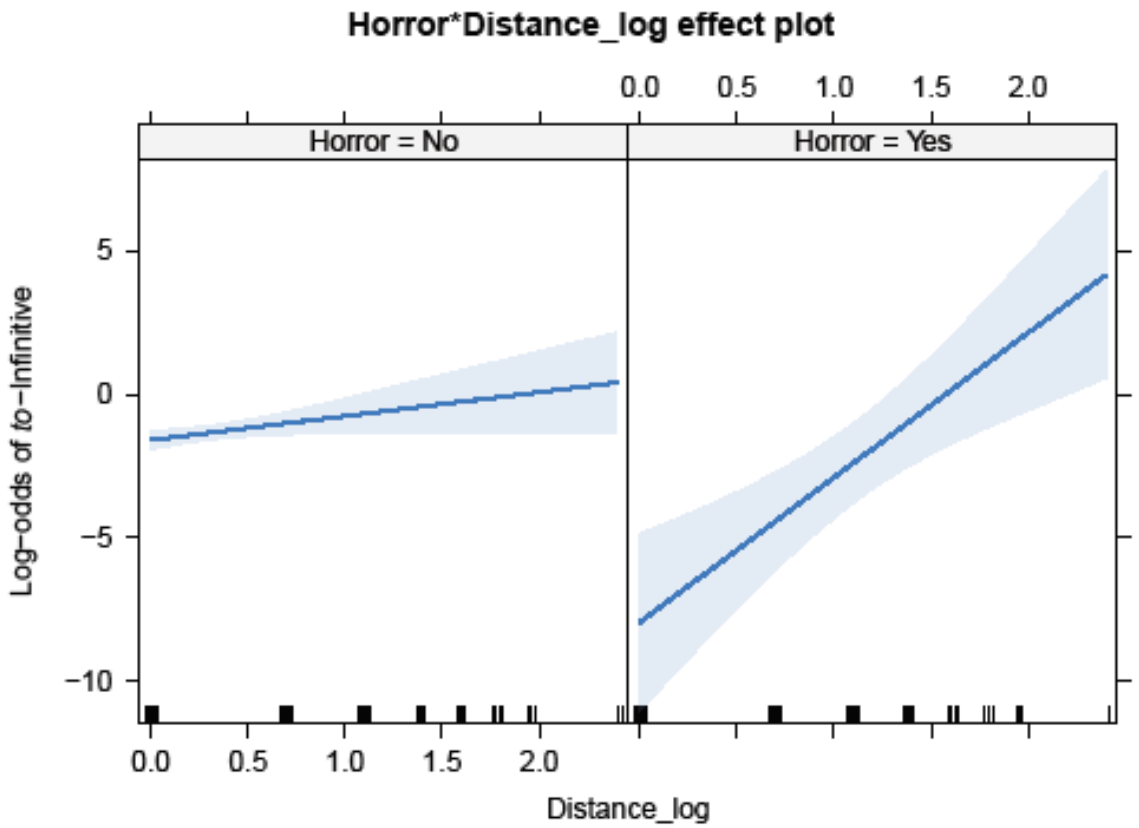
…

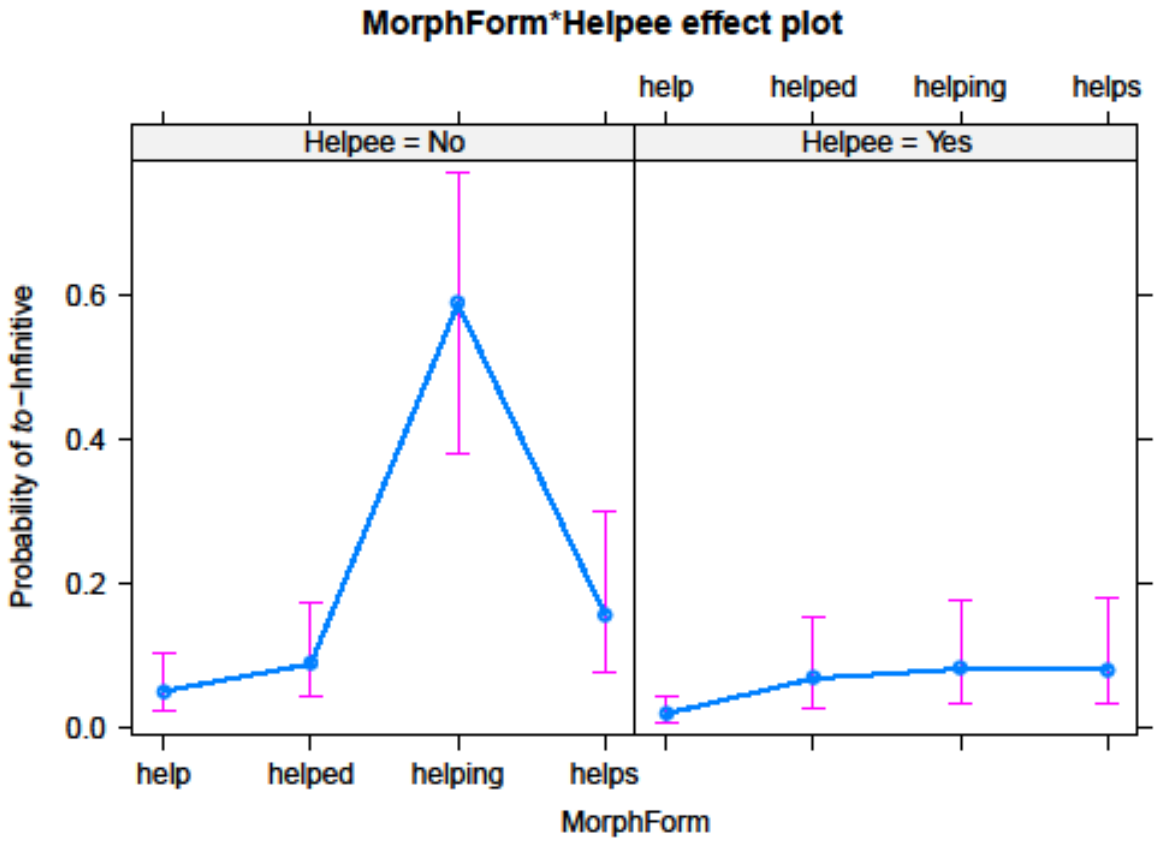

Year_new effect plot
Probability of to-infinitive
0.25
0.20
0.15
0.10
0.05
0.00
0
5
10
15
20
Year_new


Helper effect plot
Probability of to-infinitive
0.25
0.20
0.15
0.10
0.05
0.00
Inanim
Anim
Missing
Helper

Interactions

	Estimate	Std. Error	z value	Pr(>|z|)	
…					
HorrorYes:Distance_log	4.19892	1.33709	3.140	0.00169	**
MorphFormhelped:HelpeeYes	0.67234	0.37968	1.771	0.07660	.
MorphFormhelping:HelpeeYes	-1.81554	0.40431	-4.491	7.11e-06	***
MorphFormhelps:HelpeeYes	0.20592	0.42235	0.488	0.62586	


Horror*Distance_log effect plot
Horror = No
Horror = Yes
Log−odds of to−Infinitive
Distance_log
0.0
0.5
1.0
1.5
2.0
5
0
−5
−10


MorphForm*Helpee effect plot
help
helped
helping
helps
Helpee = No
Helpee = Yes
Probability of to−Infinitive
0.0
0.2
0.4
0.6
help
helped
helping
helps
MorphForm

Outline

1. Advantages and main principles of Bayesian inference

2. A case study of help + (to) Infinitive

   - Large sample:
     - A maximum likelihood model
     - A Bayesian model
   - Small sample:
     - A maximum likelihood model
     - A Bayesian model with strong recycled priors

3. Conclusions

MAX PLANCK

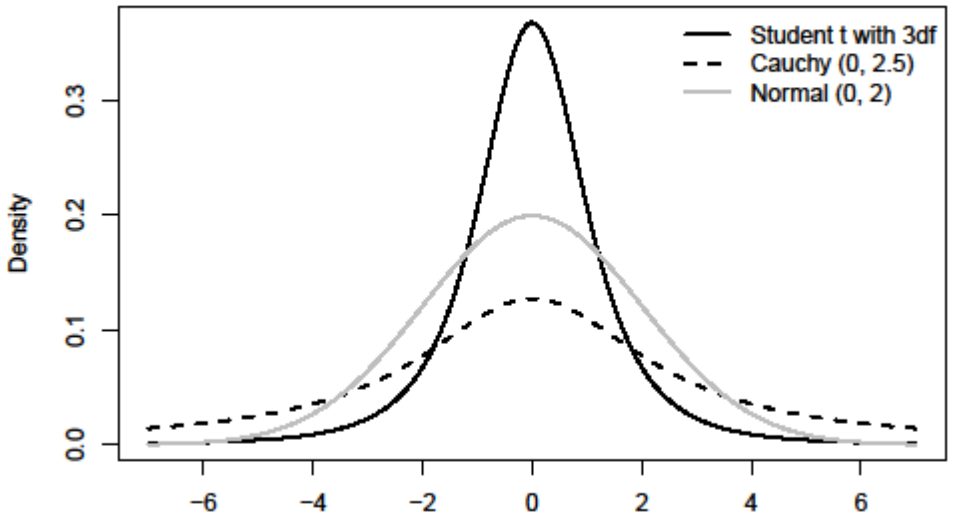
Bayesian model in brms

- For residual variances and covariances, the default non-informative priors should be used. This is done because the residuals pick up omitted variables, which almost by definition are unknown (van de Schoot et al. 2014).

- For the standard deviation of the random intercepts, weakly informative priors are used, and the number is fixed to be non-negative because variance cannot be negative by definition.

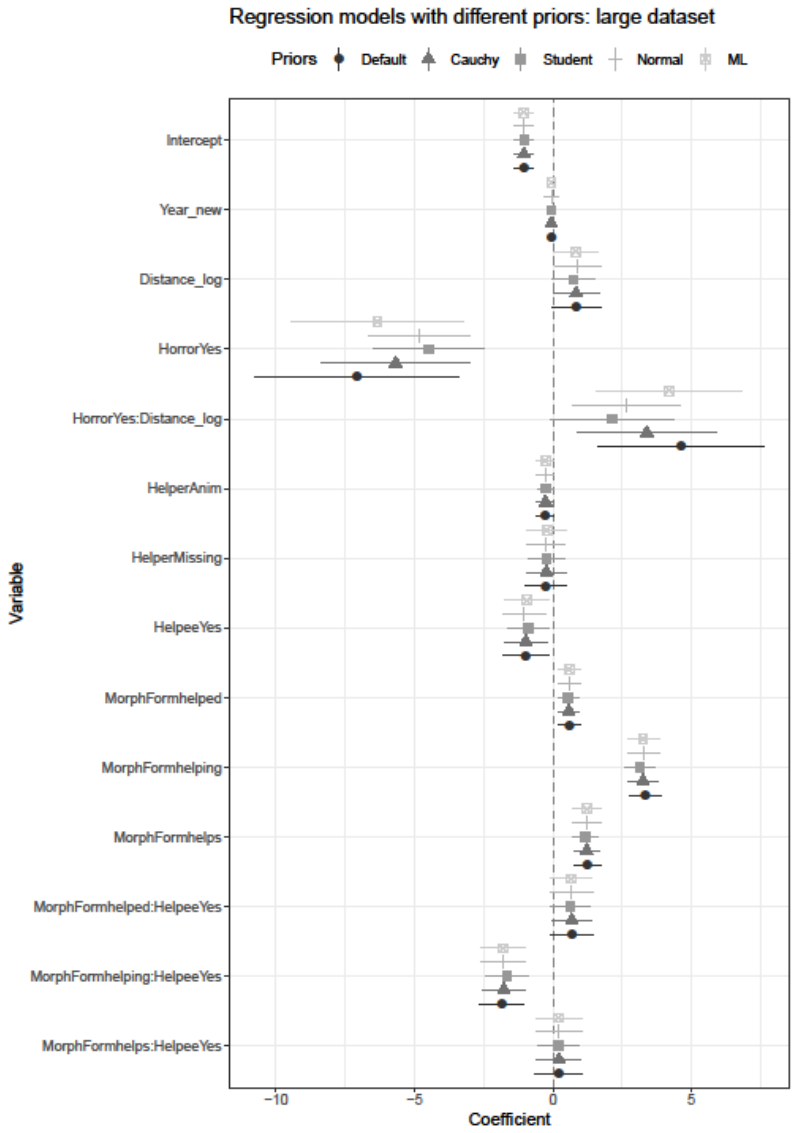
- As for the fixed effects, different options are available…

Different priors



MAX PLANCK

Sensitivity analysis + ML model



Why bother?

Regression parameter	**Probability of *b* > 0**
Intercept	0%
Year_new	0%
Horror = Yes (when Distance_log = 0)	0%
Distance_log (when Horror = No)	98.3%
MorphForm = helped (when Helpee = No)	99.9%
MorphForm = helping (when Helpee = No)	100%
MorphForm = helps (when Helpee = No)	100%
Helpee = Yes (when MorphForm = help)	0.7%
Helper = Animate (vs. Inanimate)	4.8%
Helper = Missing (vs. Inanimate)	26.7%
Horror= Yes : Distance_log	99.8%
MorphForm = helped:Helpee = Yes	96.8%
MorphForm = helping:Helpee = Yes	0%
MorphForm = helps:Helpee = Yes	69.7%

The probabilities of the positive effects of the contextual factors on the chances of the to-infinitive, based on the model with Cauchy priors.

MAX PLANCK

Interim summary

1. The Bayesian regression with default flat priors is the most similar to the ML model.

2. Informative priors help to shrink the most extreme coefficients.

3. The chances of animate Helpers having a positive effect are only 4.8%. This means that their chances of having a negative effect are 95.2%, which is quite substantial. Recall that we would have to throw away this variable based on its p-value.

Outline

1. Advantages and main principles of Bayesian inference

2. A case study of help + (to) Infinitive

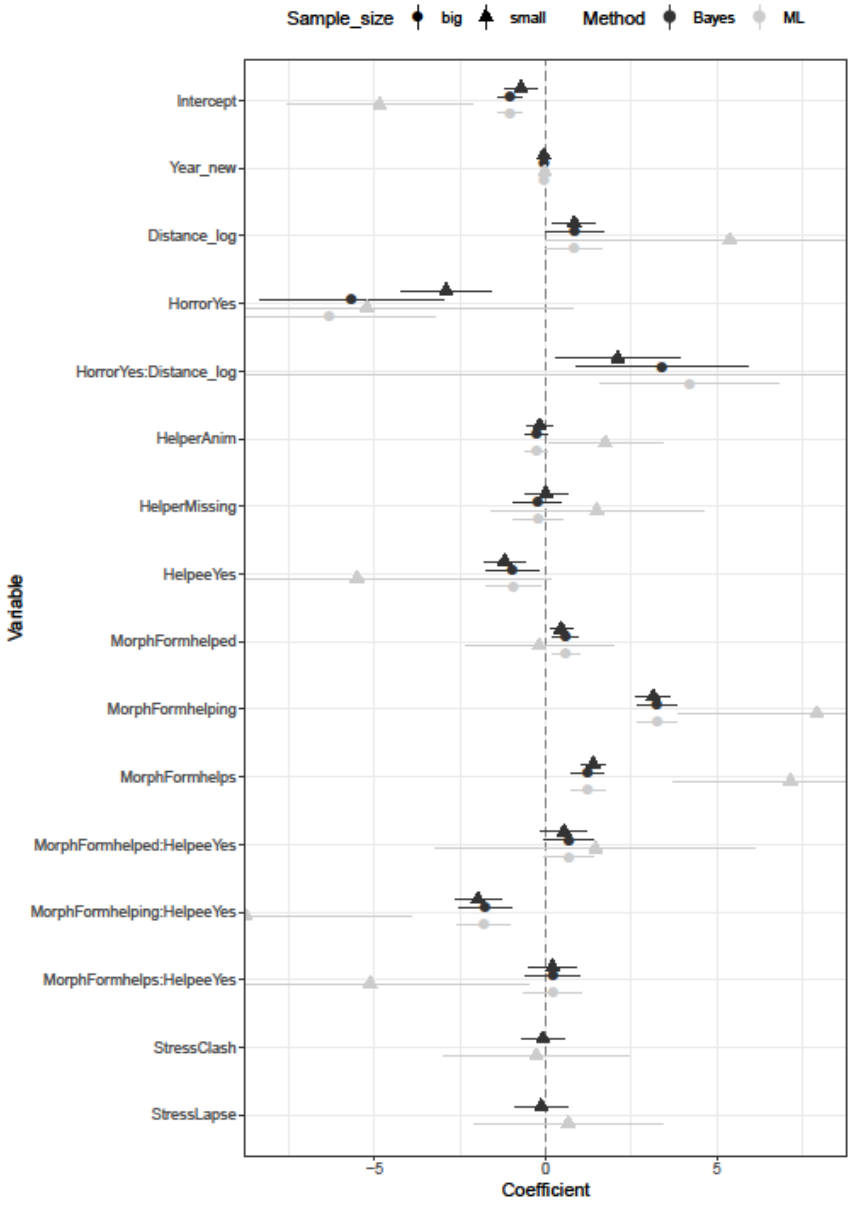
   - Large sample:
     - A maximum likelihood model
     - A Bayesian model
   - Small sample:
     - A maximum likelihood model
     - A Bayesian model with strong recycled priors

3. Conclusions

MAX PLANCK

Recycling the priors

- Normal priors with the means and the standard deviations of the posteriors from the big Bayesian model.


Sample_size
big
small
Method
Bayes
ML
Intercept
Year_new
Distance_log
HorrorYes
HorrorYes:Distance_log
HelperAnim
HelperMissing
HelpeeYes
MorphFormhelped
MorphFormhelping
MorphFormhelps
MorphFormhelped:HelpeeYes
MorphFormhelping:HelpeeYes
MorphFormhelps:HelpeeYes
StressClash
StressLapse
Variable
−5
0
5
Coefficient

Interpretation

- The small-sample ML model performs very poorly.
  - It doesn't converge.
  - Its coefficients are all over the place.
  - The confidence intervals are so broad that they even do not fit in the boundaries of the plot between -8 and 8 log odds.
  - The *C*-index is 0.99. All this indicates that the model strongly overfits the data. In other words, it fits the noise, and will be useless if we take another sample.
- The small-sample Bayesian model, in contrast, behaves similar to the large-sample models.
- No interesting effects of the stress variable.

MAX PLANCK

Outline

1. Advantages and main principles of Bayesian inference

2. A case study of help + (to) Infinitive

   - Large sample:
     - A maximum likelihood model
     - A Bayesian model
   - Small sample:
     - A maximum likelihood model
     - A Bayesian model with strong recycled priors

3. Conclusions

MAX PLANCK

Conclusions

- This example has shown how we can build on our knowledge from previous studies and recycle the results.
- The informative priors also help to keep the model reasonable in case of data sparseness.
- Given the hard, tedious work of manual annotation, this can help us test new hypotheses using smaller samples. This will speed up the process of accumulating knowledge and lead to more efficient use of resources.

MAX PLANCK