Bayesian hierarchical models as a path for analysis of individual and group data

New Sounds 2022

Ronaldo Lima Jr.
ronaldojr@letras.ufc.br
ronaldolimajr.github.io

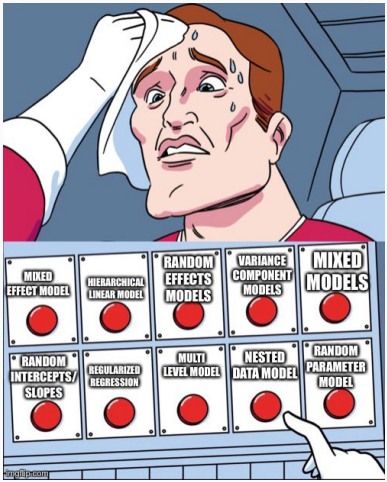
Federal University of Ceará

Outline

1. Problem

2. Suggestions
   i. Hierarchical/Multilevel/Mixed-effets models
   ii. Bayesian models

Terminology



- Mixed-effects model
- Mixed model
- Random-effects model
- Hierarchical model
- Multilevel model

https://twitter.com/chelseaparlett/status/1458461737431146500?s=21&t=6A3Ftp2BDfdT5U99k5qzBA

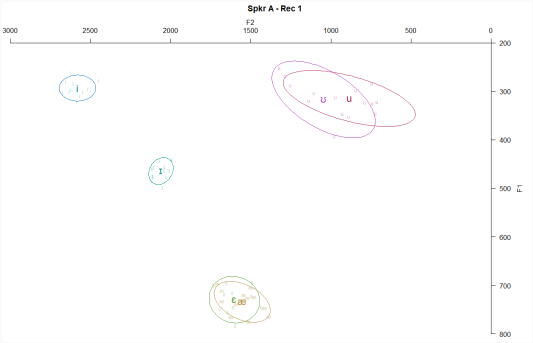
Problem

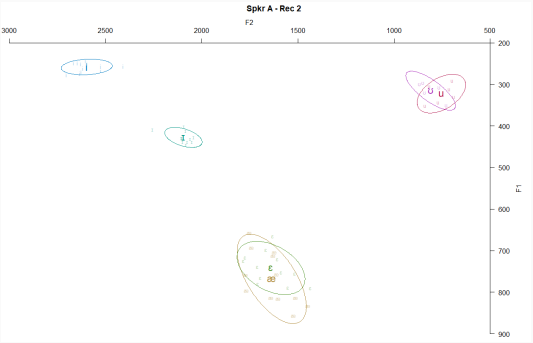
L2 speech data

- We usually collect a lot of data from the same speakers
- We want to model language development as a whole
  - The market needs generalizations (book editors, teachers, teacher trainers, proficiency tests, etc.)
- However, language development isn't the same for everyone, rather it is
  - complex, dynamic, non-linear and emergent
  - in several ways, it is idiosyncratic
- So, how do we look into L2 speech data?

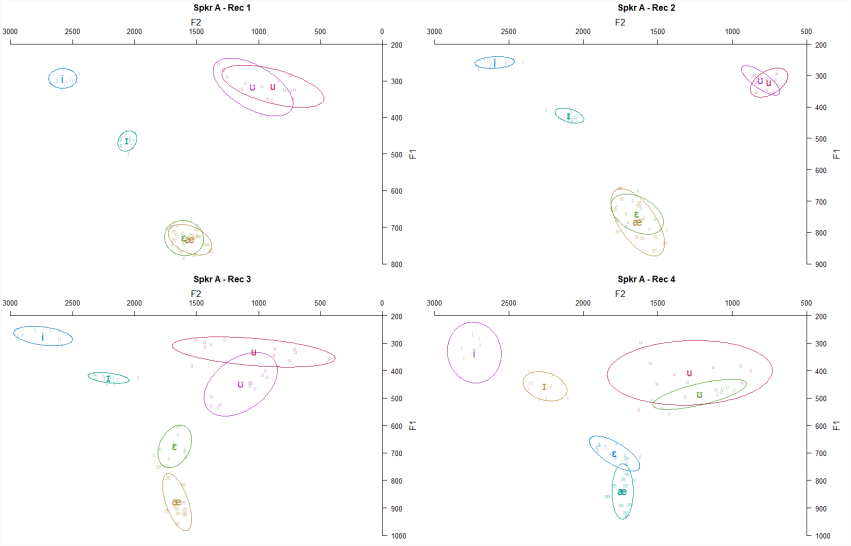
Individual data

→ Look at individual data?

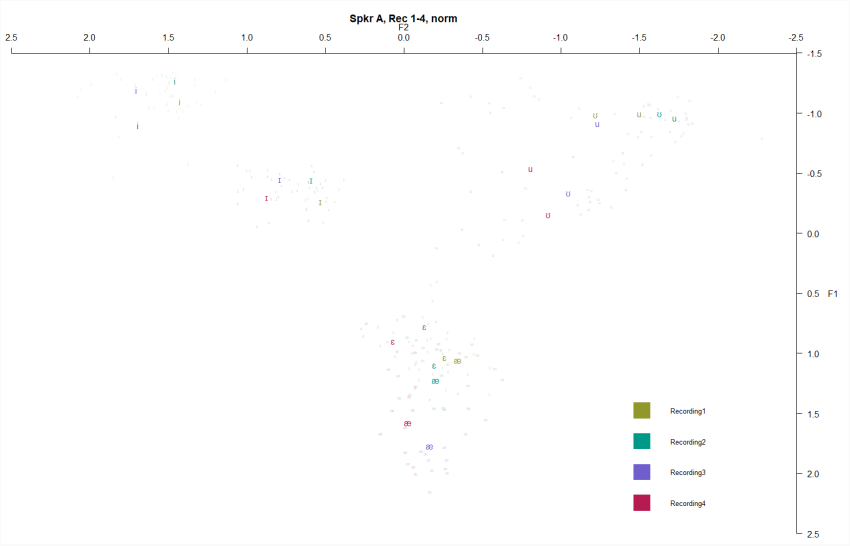


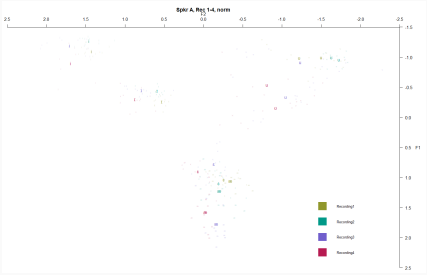


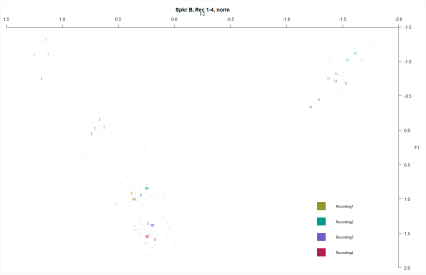
Individual data

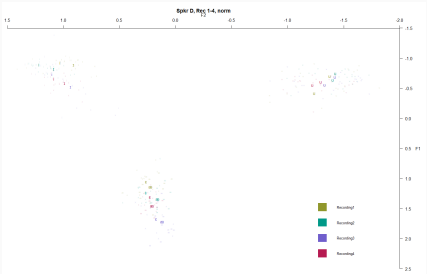


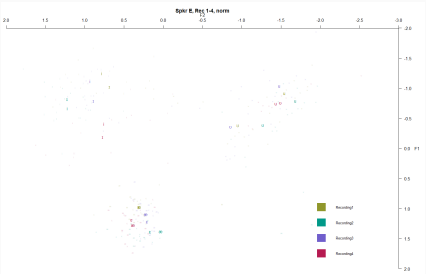
Individual data

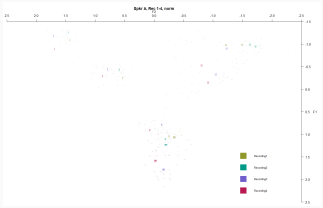


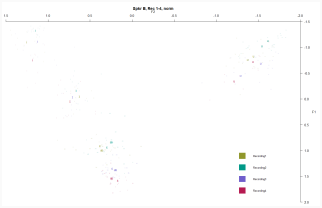

Spkr A, Rec 1-4, norm
F2
F1
Recording1
Recording2
Recording3
Recording4

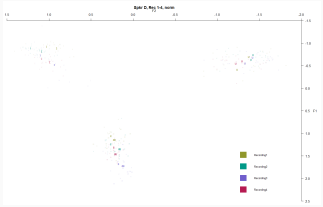

Spkr B, Rec 1-4, norm
F2
F1
Recording1
Recording2
Recording3
Recording4

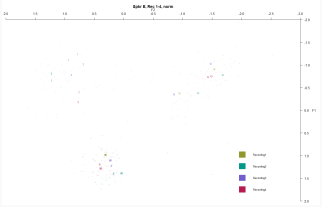

Spkr D, Rec 1-4, norm
F2
F1
Recording1
Recording2
Recording3
Recording4


Spkr E, Rec 1-4, norm
F2
F1
Recording1
Recording2
Recording3
Recording4







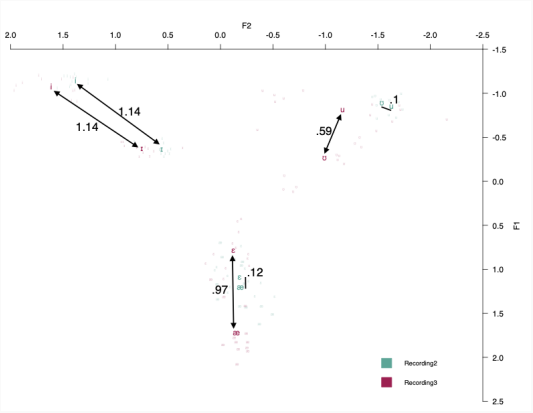


→ 10 speakers!

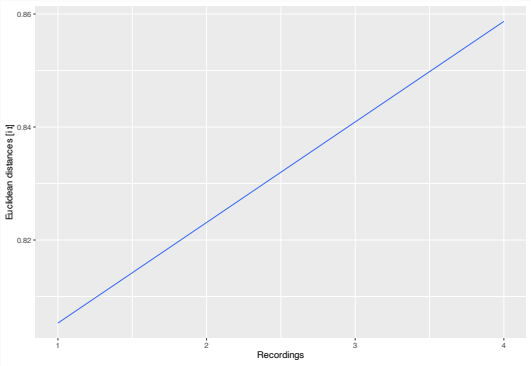
→ Impossible to make generalizations

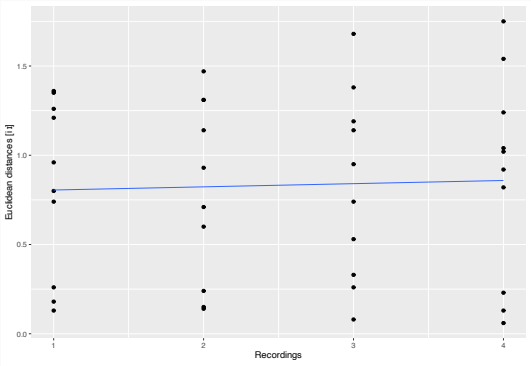
Group data

→ Calculate (Euclidean) distances between (means of) vowels in each pair, for each speaker, at each recording



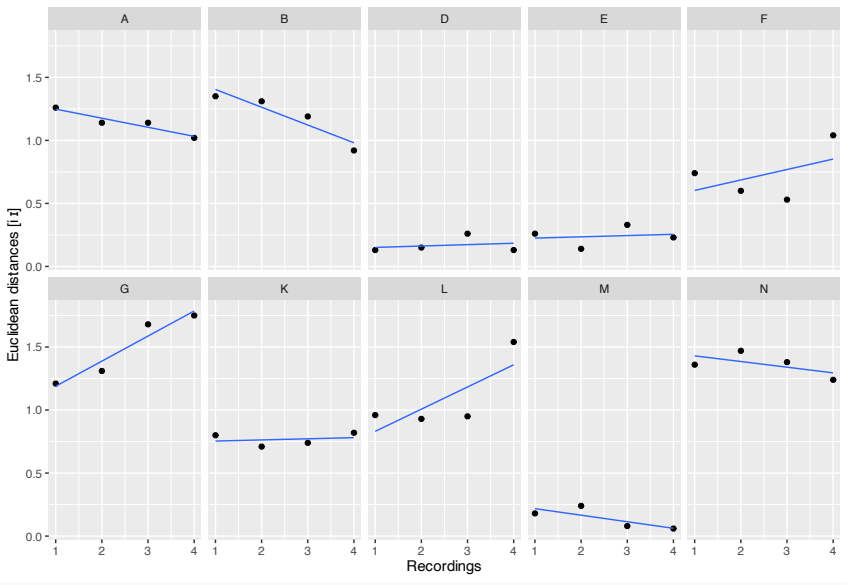
Group data for [i ɪ]





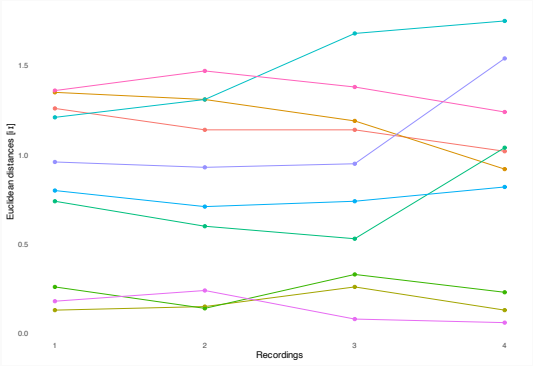
→ $f(3) = 0.035$; $p = 0.991$

But...

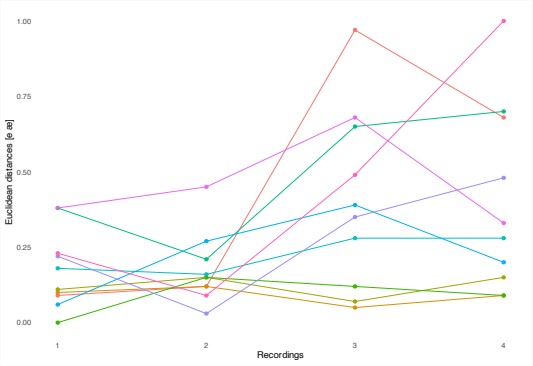

A
B
D
E
F
G
K
L
M
N
Euclidean distances [i,i]
0.0
0.5
1.0
1.5
1
2
3
4
Recordings

Straight line?

[i ɪ]



[ɛ ae]



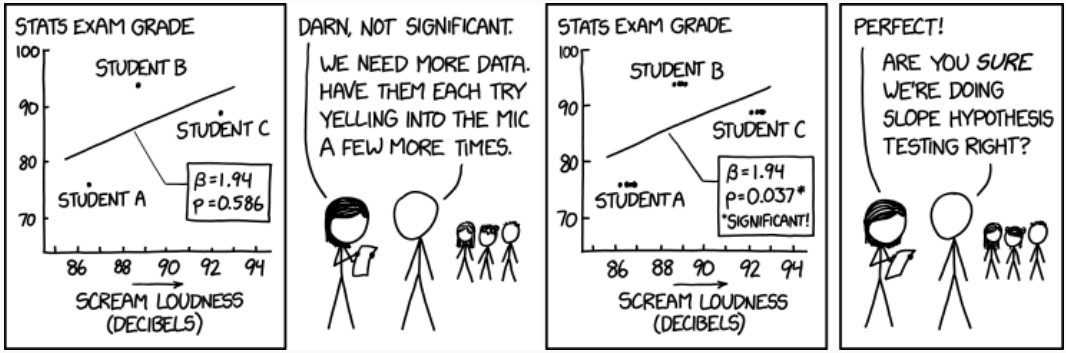
Individual or group data?

- We can't model L2 (speech) development according to individual trends (overfitting)
- We can't model L2 (speech) development according to ground tendency alone (underfitting)

→ **Suggestion**: Look into both

Suggestions

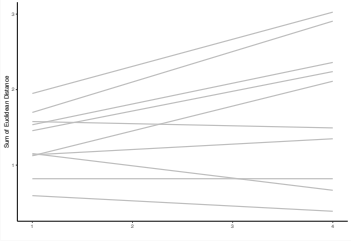
Hierarchical/Multilevel/Mixed-effets model



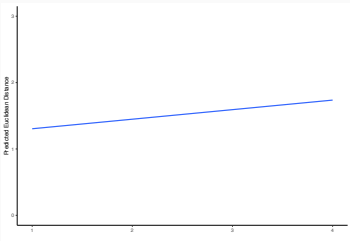
https://xkcd.com/2533/

Sum of Euclidean Distances

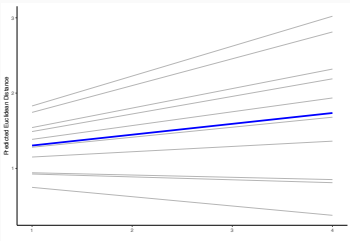
Individual trends



Linear model (no mixed effects)



Varying intercepts/slopes



→ Did adding the varying terms change the line?

Not really, but it changed the confidence of model about the line:

Sum of Euclidean Distances

→ Did adding the varying terms change the line?

Not really, but it changed the confidence of model about the line:

```
> fit 1 = lm(sum ~ recording)
> summary(fit1)

Coefficients:
             Estimate Std. Error t value
(Intercept)    1.1605     0.2767   4.195
recording      0.1439     0.1010   1.424
```

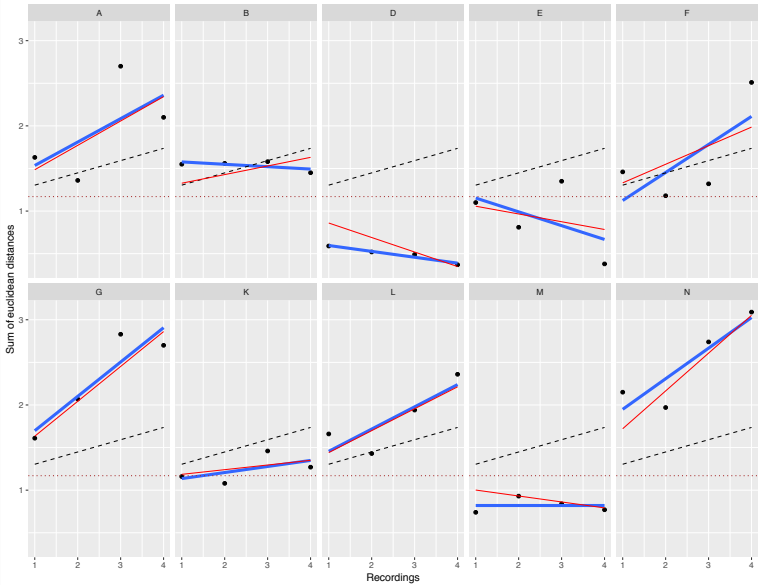
```
> fit2 = lmer(sum ~ recording + (recording|speaker))
> summary(fit2)

Fixed effects:
             Estimate Std. Error       df t value
(Intercept)  1.16050    0.12993 15.49896   8.932
recording    0.14390    0.07027  9.81783   2.048
```

→ As a result:

Predictors	Estimates	CI	p	Estimates	CI	p
Intercept	1.16	0.60 – 1.72	<0.001	1.16	0.90 – 1.42	<0.001
recording	0.14	-0.06 – 0.35	0.162	0.14	0.00 – 0.29	**0.048**

Varying intercepts and slopes

→Also, a mixed-effects model can predict different lines for each subject



Why Bayesian?

→ Probability of the parameters (hypotheses) given the data
(instead of probability of the data given the H_0)

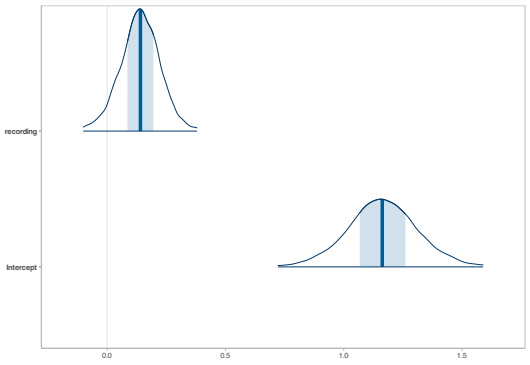
→ Probability distributions for coefficients
(instead of point estimates)

→ Credible intervals
(instead of confidence intervals)

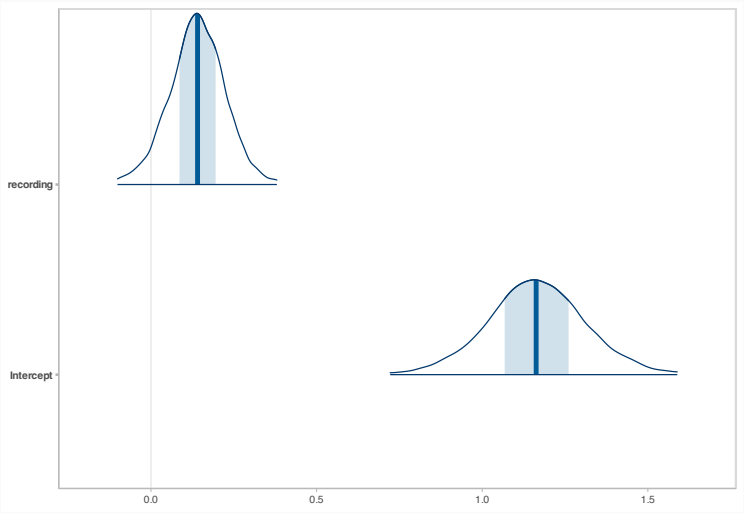
→ Add prior information/knowledge to the model
(instead of all outcomes having equal probability *a priori*)

Bayesian mixed-effects

Predictors	Estimates	50% CI	95% CI
Intercept	1.16	1.07 – 1.26	0.86 – 1.47
Recording	0.14	0.09 – 0.20	-0.04 – 0.31

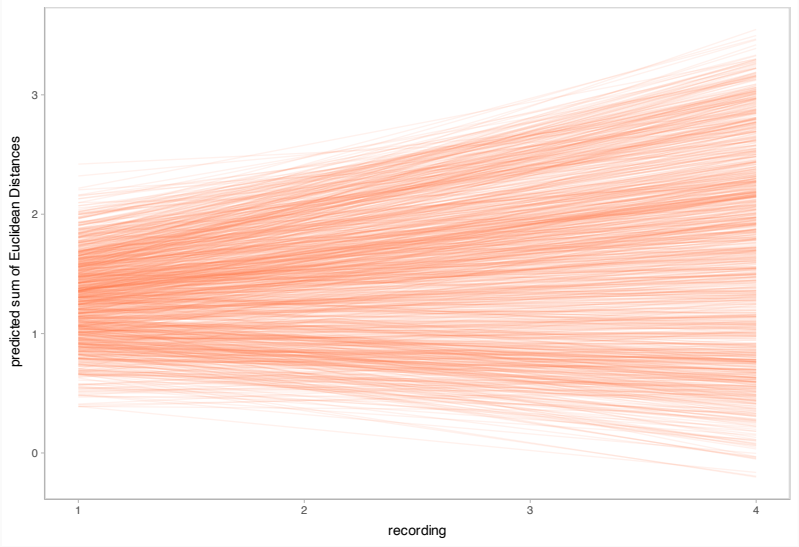


Bayesian mixed-effects



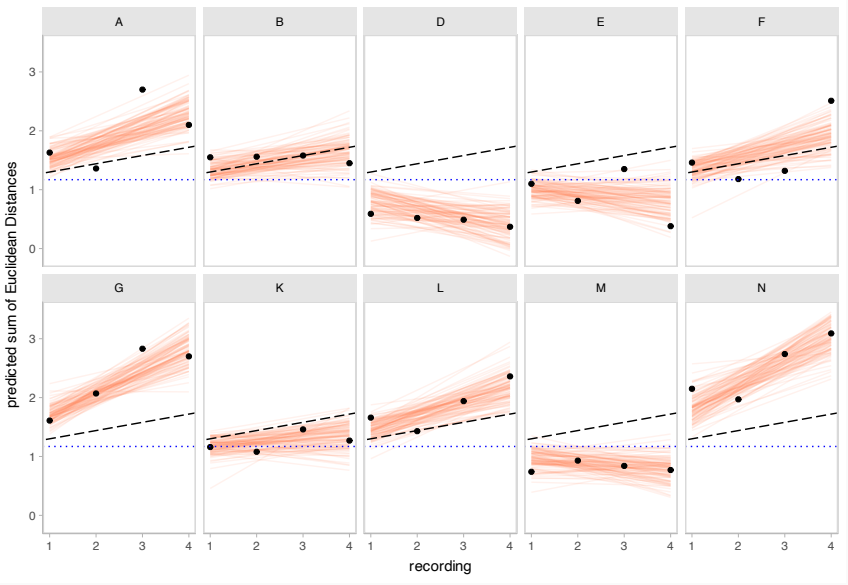
- → 6% of AUC (area under the curve) below 0
- → This analysis adds the uncertainty needed when inferring population values from limited samples

Bayesian mixed-effects: predicted values



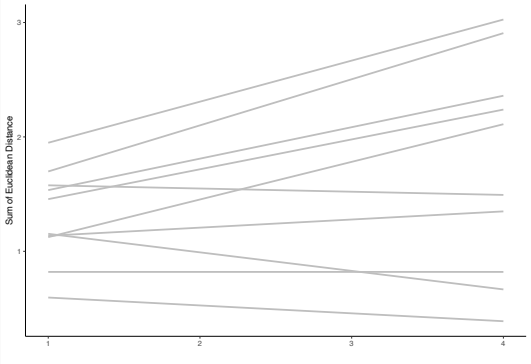
Bayesian mixed-effects

→Several (in this case, 100) probable lines predicted by the model sampled from the posterior distribution (instead of a single line)

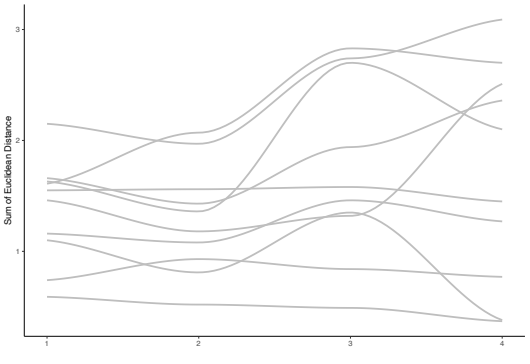


It doesn't have to be lines

```
1 | geom_smooth(method = lm)
```




```
1 | geom_smooth(method = loess)
```



It doesn't have to be lines

- "linear" in math does not mean a 1:1 relationship, nor does it mean a straight line

→ It means addition of terms

- There are (linear) regression models that predict curves by adding specific terms to the regression formula. E.g.:
  - Polynomial regressions (quadratic, cubic, etc.)
  - Splines
  - Generalized Additive Models (GAMs)

Questions?