Week 10 – Chapter 11

Lin 205C
Santiago Barreda

Chapter Precap

- Why Bayesian multilevel models are well suited for the analysis of complex models with large numbers of parameters, and some strategies for interpreting these models are presented.
- A model with two quantitative predictors, as well as an interaction between them, is outlined, and the model is ft and interpreted.
- We introduce Bayesian analysis of variance (BANOVA) and Bayesian ANOVA plots and use these to investigate the initial model presented in the chapter.
- A multivariate logistic regression model is ft and interpreted using a Bayesian ANOVA.
- After this, we explain how multivariate logistic models can be used to interpret classification and listener behavior.
- Finally, we discuss model selection and misspecification, and some diagnostics that can help find and diagnose problems in large models.

Problems with ‘Large’ Models

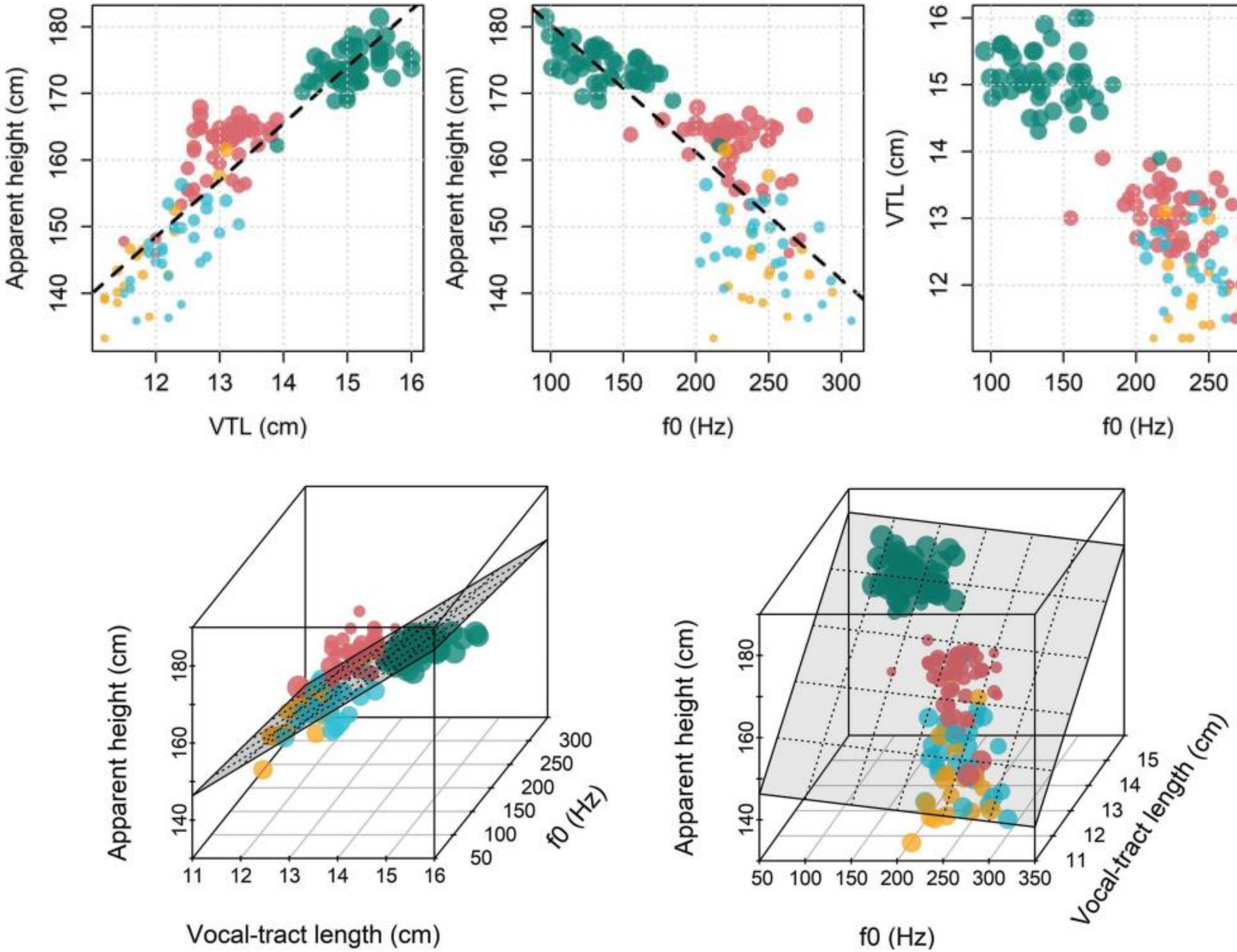
1. A model with many predictors may have ‘too many’ predictors. We will vaguely define ‘extra’ predictors as those which have no meaningful statistical association with your dependent variable, and which do not appreciably improve your model in any way. Sometimes, the ‘extra’ predictors have estimated values that are difficult to distinguish from those of the ‘real’ parameters, leading to incorrect conclusions.
2. A model with many parameters may have more difficulties with fitting/convergence, meaning you can’t get good estimates for the model coefficients. Regardless of the approach to parameter estimation, more complicated models make it more difficult to find the optimal parameter values given the data and chosen model structure.
3. It can be difficult to interpret a model with hundreds (or thousands) of parameters. With so many parameters it is important to not miss the forest for the trees, that is, not consider every parameter in isolation to the detriment of understanding what information your fitted model provides in general.

Dealing with ‘Large’ Models

In order to avoid a situation where you end up with data you can’t analyze or a model you can’t interpret, it’s worth considering the following questions *before* advancing to the data collection stage of your experiment:

- How will I analyze the data? Will I be able to carry out my planned analysis?
- What will the model structure be?
- What kind of results am I expecting?
- How will the expected results be reflected by the model parameters? How would different results manifest in the model parameters?

Multiple Quantitative Predictors

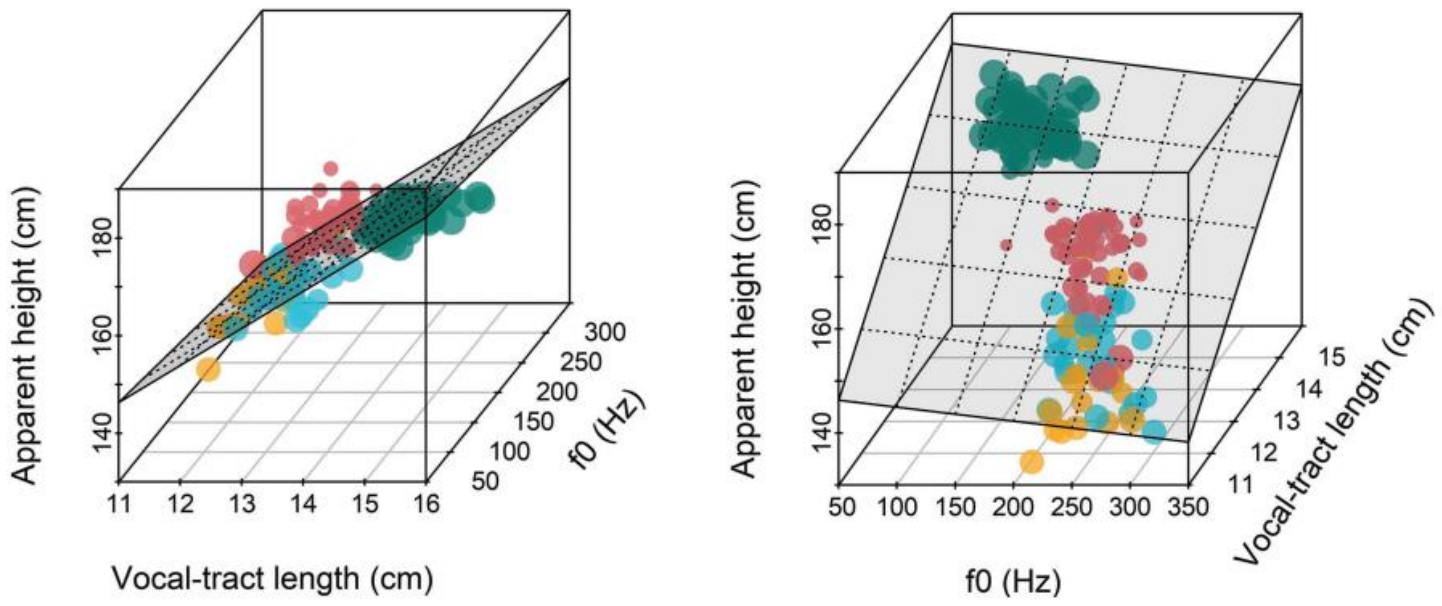


Multiple Quantitative Predictors

- Planes have one intercept and two slopes.

$$z = a + b \cdot x + c \cdot y$$

Height along 'z' (vertical) axis.



Interactions Between Quantitative Predictors

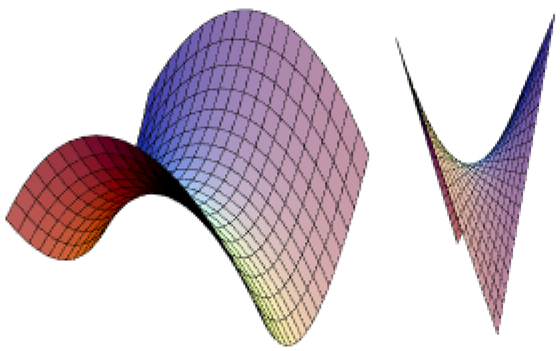
- 'Interactions' between quantitative predictors refer to the effect for the cross-product of those predictors.

$$z = a + (b \cdot x) + (c \cdot y) + (d \cdot x \cdot y)$$

- You can enter these using a model formula ($y \sim x \cdot y$) or multiply your predictors yourself, and use this as a predictor ($y \sim x \cdot y \cdot xy$).

Hyperbolic Paraboloid

Download
Wolfram Notebook



A hyperbolic paraboloid is the quadratic and doubly ruled surface given by the Cartesian equation

$$z = \frac{y^2}{b^2} - \frac{x^2}{a^2} \tag{1}$$

(left figure). An alternative form is

$$z = x\,y \tag{2}$$

(right figure; Fischer 1986), which has parametric equations

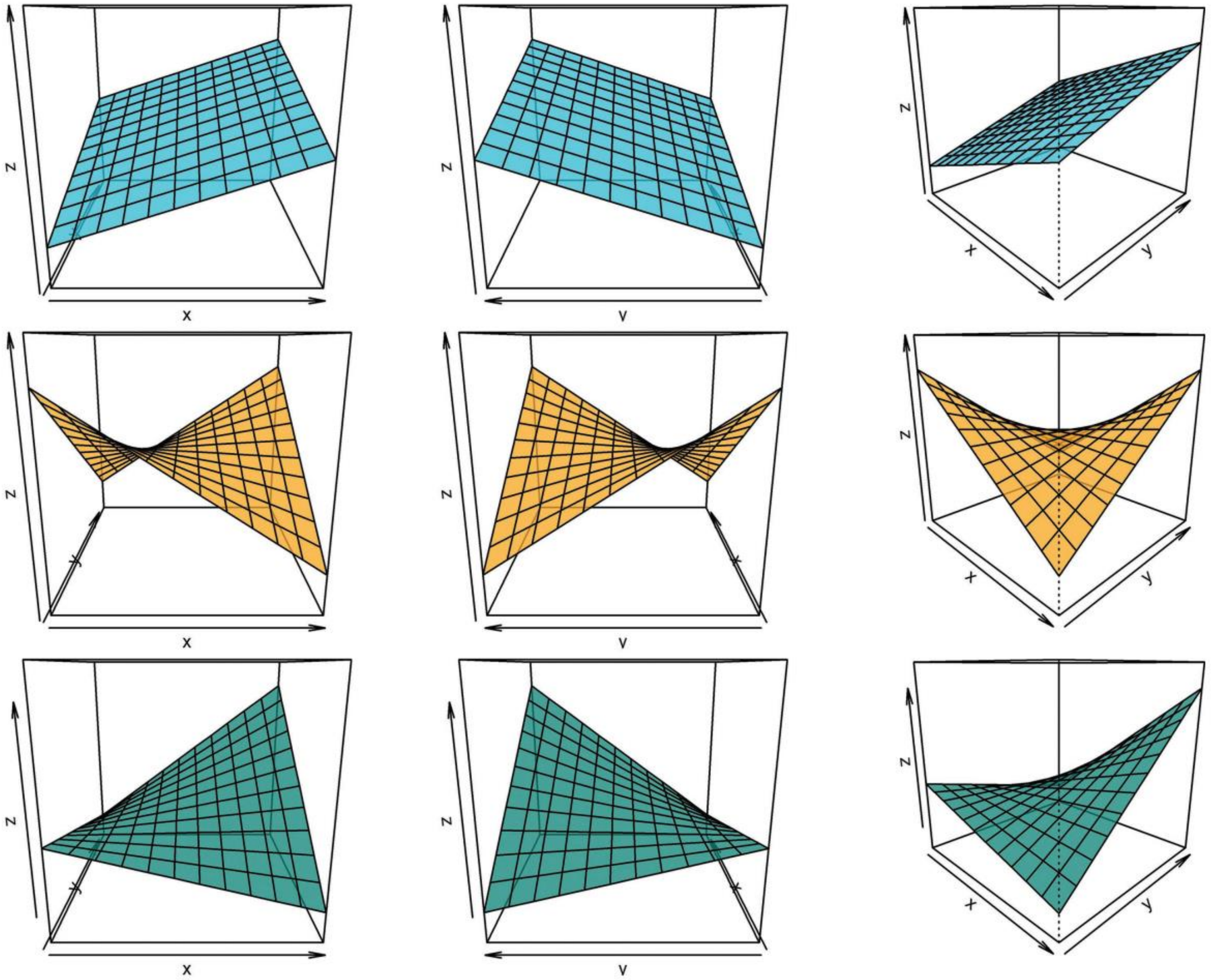
$$x(u, v) = u \tag{3}$$

$$y(u, v) = v \tag{4}$$

$$z(u, v) = u\,v \tag{5}$$

Interactions Between Quantitative Predictors

$$z = \boxed{a + (b \cdot x) + (c \cdot y)} + (d \cdot x \cdot y)$$



Linearly changing slopes

$$z = a + (b \cdot x) + \cancel{(c \cdot y)} + (d \cdot x \cdot y)$$

$$z = a + (b \cdot x) + (d \cdot x \cdot y)$$
$$z = a + (b) \cdot x + (d \cdot y) \cdot x$$
$$z = a + (b + d \cdot y) \cdot x$$

$$b = 2, d = 1, y = 1$$
$$z = a + (2 + 1 \cdot 1) \cdot x$$
$$z = a + 3 \cdot x$$

$$b = 2, d = 0$$
$$z = a + (b + d \cdot y) \cdot x$$
$$z = a + (2 + 0 \cdot y) \cdot x$$
$$z = a + 2 \cdot x$$

$$b = 2, d = 1, y = 5$$
$$z = a + (2 + 1 \cdot 5) \cdot x$$
$$z = a + 7 \cdot x$$

Center Your Predictors!

- Your x slope coefficient represents the effect of x <u>only</u> when y=0!!
- When y≠0, the effect of x may be bigger/smaller.
- Centering your predictors makes your slope for x equal to the effect at the <u>mean</u> of y.

$$b = 2,\ d = 0$$
$$z = a + (b + d \cdot y) \cdot x$$
$$z = a + (2 + 0 \cdot y) \cdot x$$
$$z = a + 2 \cdot x$$

$$b = 2,\ d = 1,\ y = 5$$
$$z = a + (2 + 1 \cdot 5) \cdot x$$
$$z = a + 7 \cdot x$$

Data and Research Questions

```
library (brms)
library (bmmb)
options (contrasts = c('contr.sum','contr.sum'))

data (exp_data)

# center VTL
exp_data$vtl_original = exp_data$vtl
exp_data$vtl = exp_data$vtl - mean (exp_data$vtl)

# center and scale f0
exp_data$f0_original = exp_data$f0
exp_data$f0 = exp_data$f0 - mean(exp_data$f0)
exp_data$f0 = exp_data$f0 / 100
```

(Q1) What do we do with all these parameters? How do we know what to focus on, and where to begin?

Description of the Model

```
height ~ vtl*f0*A*G + (vtl*f0*A*G|L) + (1|S)
```

```
height ~ Intercept + vtl + f0 + A + G +
         A:G1 + vtl:f0 + vtl:A + vtl:G + f0:A + f0:G +
         vtl:f0:A + vtl:f0:G + vtl:A:G + f0:A:G + vtl:f0:A:G
```

```
a = Intercept + A          + G          + A:G
b = vtl       + vtl:A      + vtl:G      + vtl:A:G
c = f0        + f0:A       + f0:G       + f0:A:G
d = vtl:f0    + vtl:f0:A + vtl:f0:G + vtl:f0:A:G
```

Description of the Model

```
height ~ vtl*f0*A*G + (vtl*f0*A*G|L) + (1|S)
```

$$\text{height}_{[i]} \sim \text{t}\left(\nu, \mu_{[i]}, \sigma\right)$$

$$\mu_{[i]} = a_{[i]} + \left(b_{[i]} \cdot \text{vtl}_{[i]}\right) + \left(c_{[i]} \cdot \text{f0}_{[i]}\right) + \left(d_{[i]} \cdot \text{f0}_{[i]} \cdot \text{vtl}_{[i]}\right)$$

...

```
a = Intercept + A         + G         + A:G
b = vtl       + vtl:A     + vtl:G     + vtl:A:G
c = f0        + f0:A      + f0:G      + f0:A:G
d = vtl:f0    + vtl:f0:A + vtl:f0:G + vtl:f0:A:G
```

Fitting the Model

```
# Fit the model yourself
priors = c(brms::set_prior("student_t(3,0, 12)", class = "Intercept"),
           brms::set_prior("student_t(3,0, 12)", class = "b"),
           brms::set_prior("student_t(3,0, 12)", class = "sd"),
           brms::set_prior("lkj_corr_cholesky (2)", class = "cor"),
           brms::set_prior("gamma(2, 0.1)", class = "nu"),
           brms::set_prior("student_t(3,0, 12)", class = "sigma"))

model_height_vtl_f0 =
  brms::brm (height ~ vtl*f0*A*G + (vtl*f0*A*G|L) + (1|S), data = exp_data,
             chains = 4, cores = 4, warmup = 1000, iter = 5000, thin = 4,
             prior = priors, family = "student")
```

Advantages of Bayesian Multilevel Models for ‘Large’ Models

- Three problems with large models are:

  1. More likely to have ‘too many’ predictors.

  2. More difficult to find the solution.

  3. More difficult to interpret the model.

multilevel models help with these.

Bayesian ANOVA helps with this.

Comparison with lmer

- lmer fits multilevel modes similar to those in brms, even using adaptive partial pooling to estimate ‘random’ effects.

- There are some important differences:

    - lmer does not use priors for any parameter estimates.

    - Shrinkage is only applied to random effects, not to any other parameters (i.e., random effect variances or correlations).

    - Intervals are only provided for ‘fixed effects’.

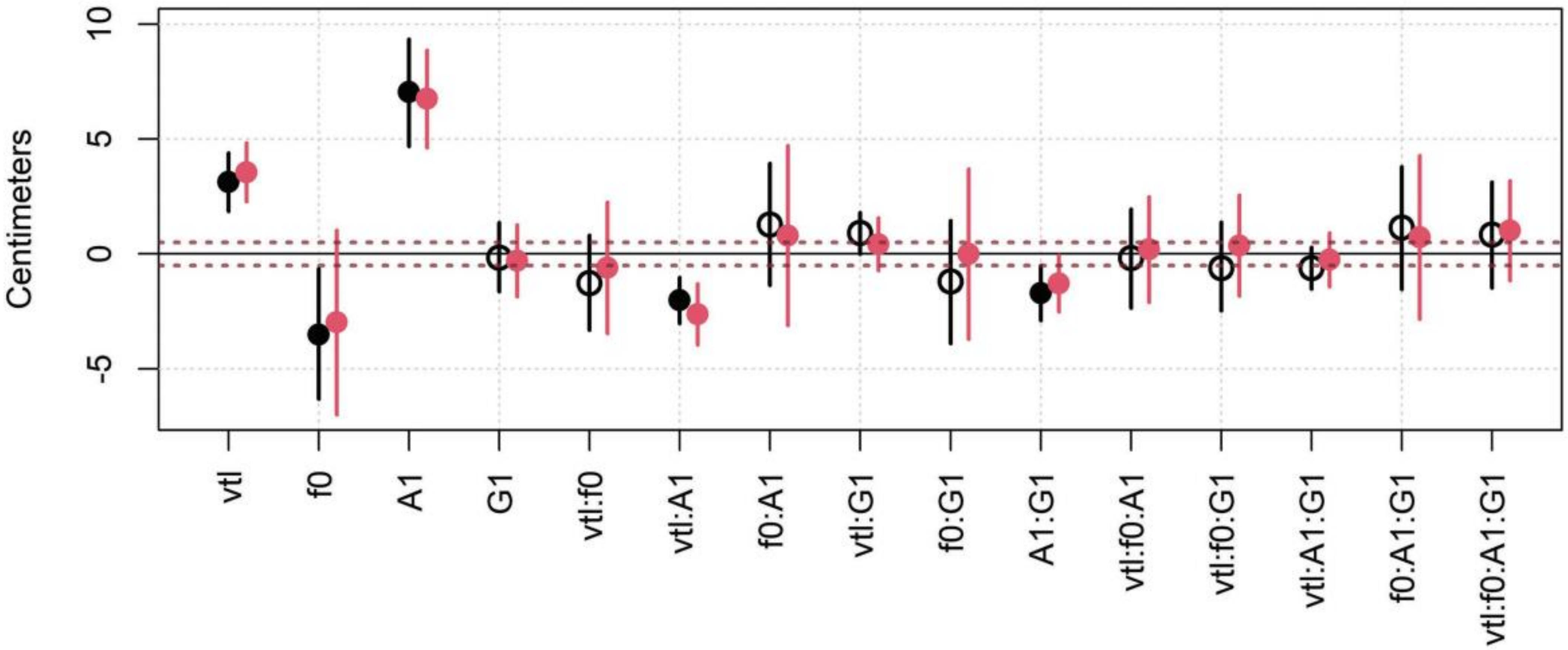
Comparison with lmer

```
# Fit the model yourself
lme_model_height_vtl_f0 =
  lme4::lmer (height ~ vtl*f0*A*G + (vtl*f0*A*G|L) + (1|S),
              data=exp_data,verbose = TRUE,
              control=lme4::lmerControl(optCtrl=list(maxfun= 20000),
              optimizer="bobyqa"))
```

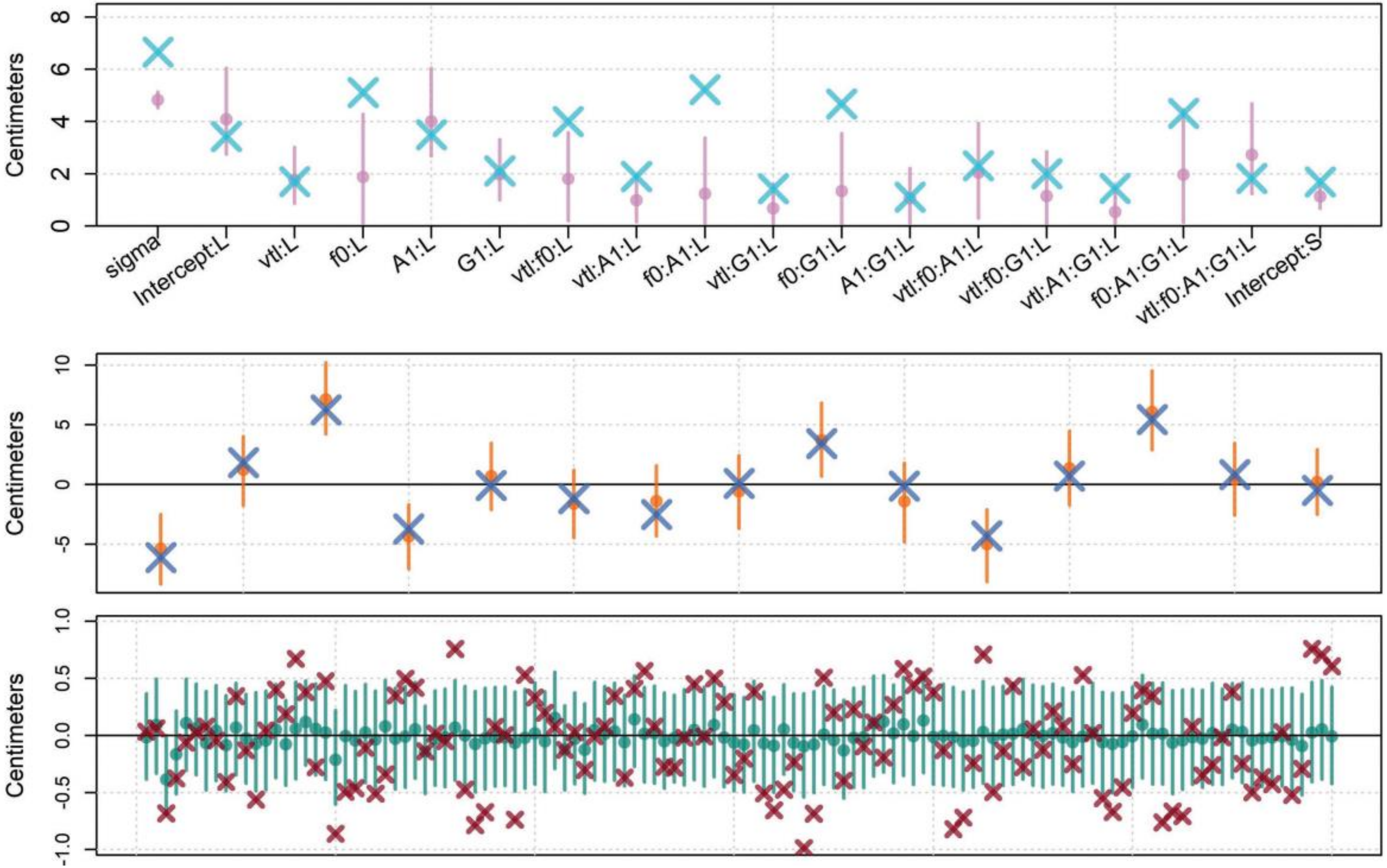
```
boundary (singular) fit: see help('isSingular')
```

Comparison: Fixed Effects

- Fixed effects about the same between brms (red) and lmer (black).



Comparison: Random Effects and Correlations



- No intervals:
  - Uncertain uncertainty.
  - What is actually different from zero?

ANalysis Of VAriance

- ANOVA sees variation in the dependent variable as the sum of a set of **components of variation** related to the predictors in our model. For example:

```
height ~ C + (C|L) + (1|S)
```

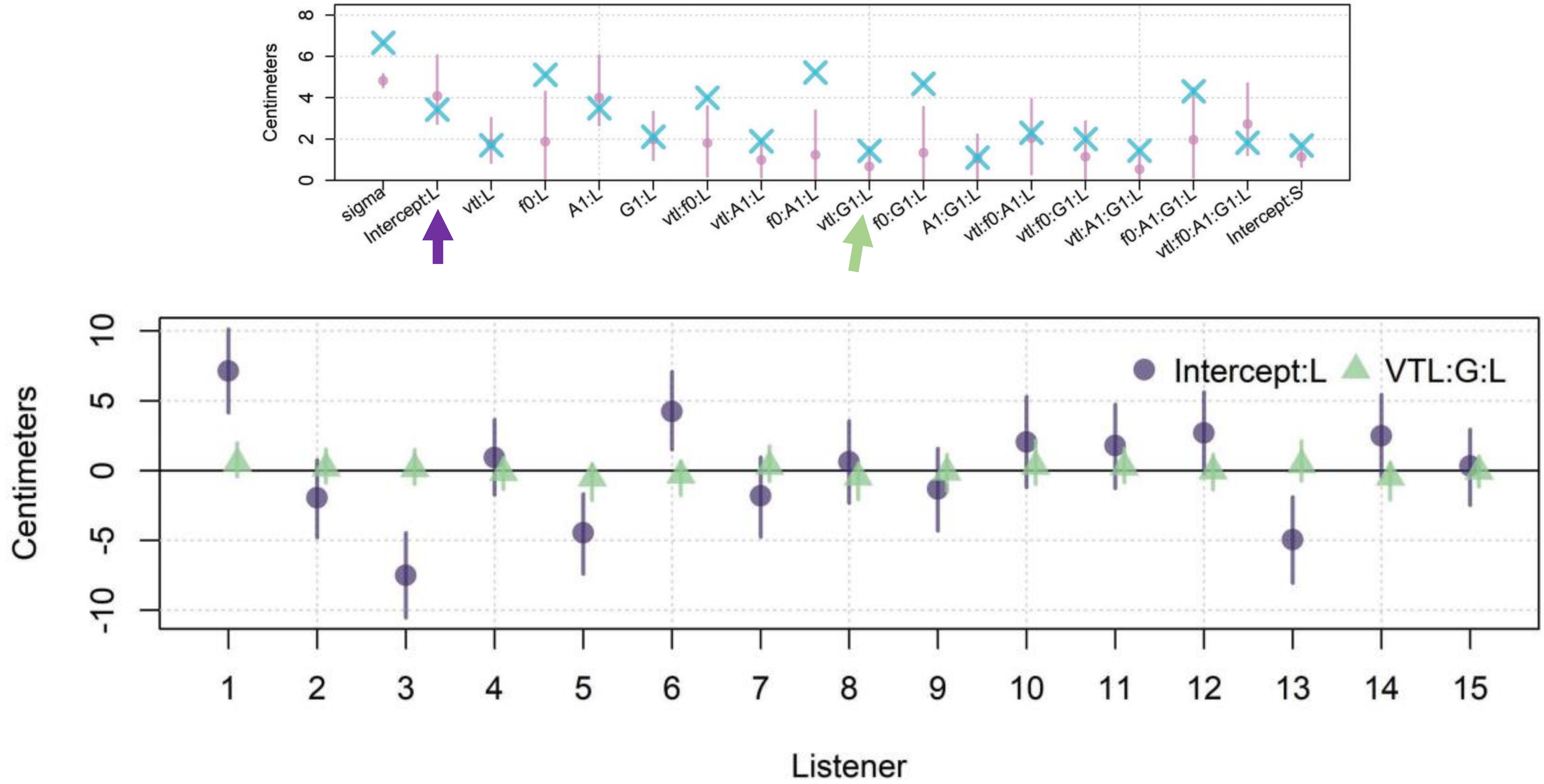
- 3 parameters representing the fixed effects for apparent speaker category `C`: `C1, C2, C3`.
- 15 parameters representing the listener-dependent intercepts.
- 45 ($15 \cdot 3$) parameters representing the listener-dependent effects for parameters `C1`, `C2`, and `C3`.
- 139 parameters representing the speaker-dependent intercepts.

ANalysis Of VAriance

- Batches are semantically/thematically linked, these are not just parameters grouped at random.
- When we break up variation into thematically-grouped parameters, we are doing an *ANOVA-like decomposition* of variation of the dependent variable.

"When moving to multilevel modeling, the key idea we want to take from the analysis of variance is the estimation of the importance of different batches of predictors ("components of variation" in ANOVA terminology). As usual, we focus on estimation rather than testing: instead of testing the null hypothesis that a variance component is zero, we estimate the standard deviation of the corresponding batch of coefficients. If this standard deviation is estimated to be small, then the source of variation is minor—we do not worry about whether it is exactly zero. In the social science and public health examples that we focus on, it can be a useful research goal to identify important sources of variation, but it is rare that anything is truly zero." (Gelman and Hill 2006, p. 490)

Batches of Parameters



Parameter Standard Deviations

We've been modeling
these (i.e. σ_L).

*"The **superpopulation** standard deviation, which represents the variation among the modeled probability distribution from which they were drawn, is relevant for determining the uncertainty about the value of a new group not in the original set of J."*

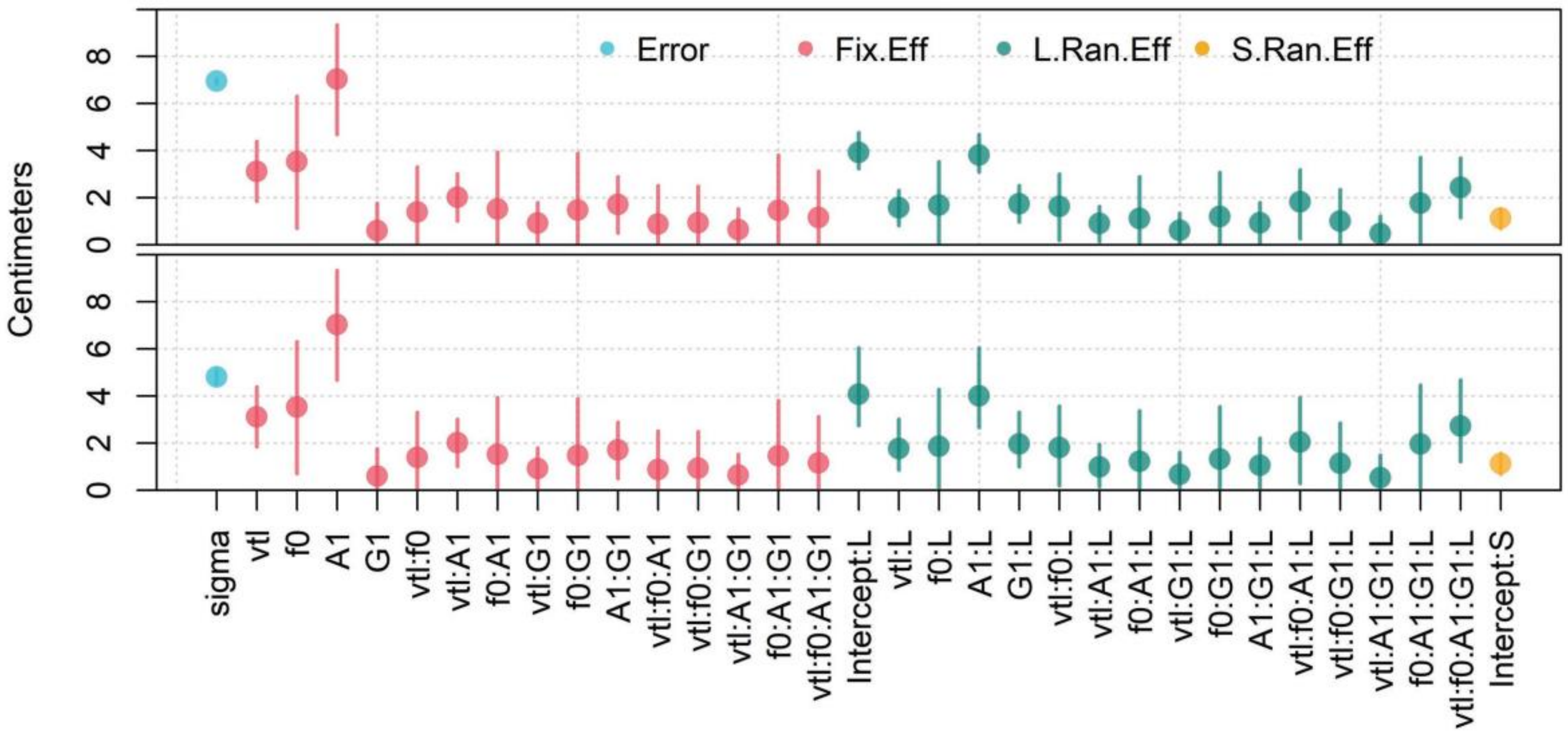
*"The **finite-population** standard deviation of the particular J values of [some coefficient] describes variation within the existing data."*

This is the actual standard deviation
of the L parameters (for example).

Bayesian ANOVA (BANOVA)

1. Fit the model with the structure you think is required to capture the variation in the data and answer your research questions.
2. Calculate the finite-population (and/or superpopulation) standard deviations for predictors or batches of predictors.
3. Make a plot comparing the magnitudes of different batches of predictors, and the uncertainty in the estimates. The authors refer to this as an **ANOVA plot**.
4. Use the ANOVA plot to make inferences about the relative importance of your predictors, and to guide your analysis.

Bayesian ANOVA (BANOVA)

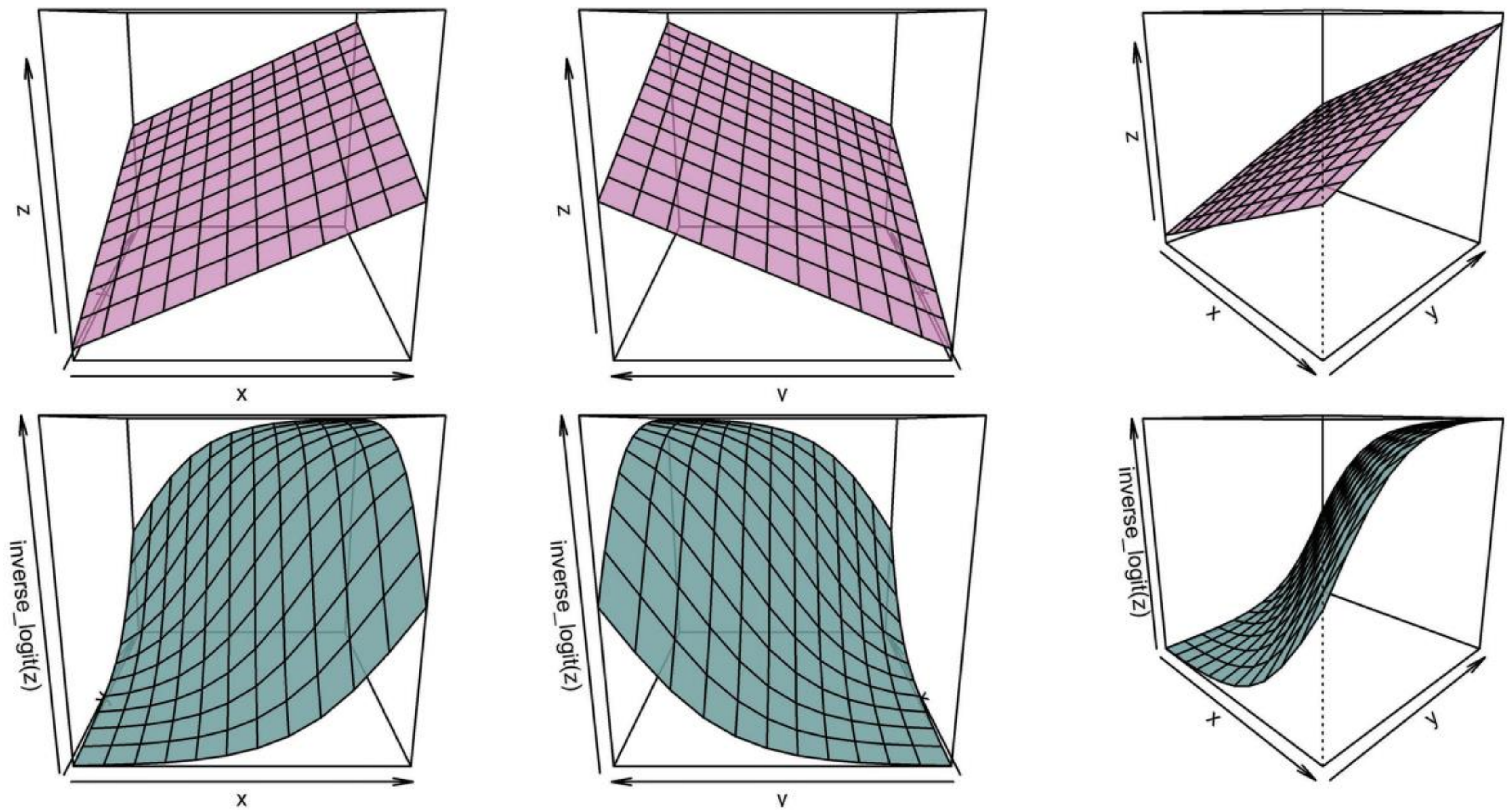


Logistic Regression with Multiple Quantitative Predictors

$$y \sim a + b \cdot x + c \cdot y$$

logits

probabilities

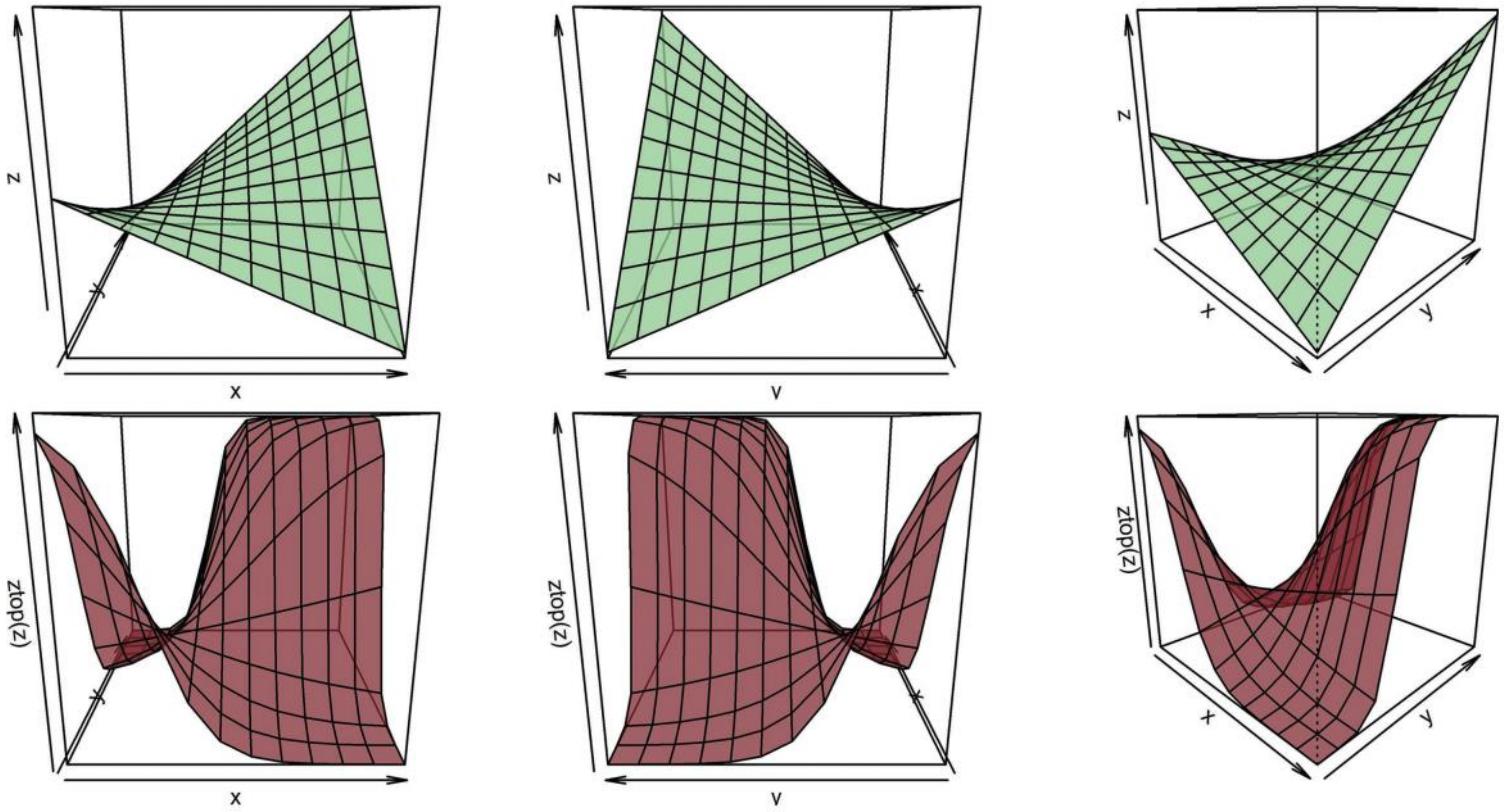


Logistic Regression with Multiple Quantitative Predictors

$$y \sim a + b \cdot x + c \cdot y + d \cdot x \cdot y$$

logits

probabilities



Data and Research Questions

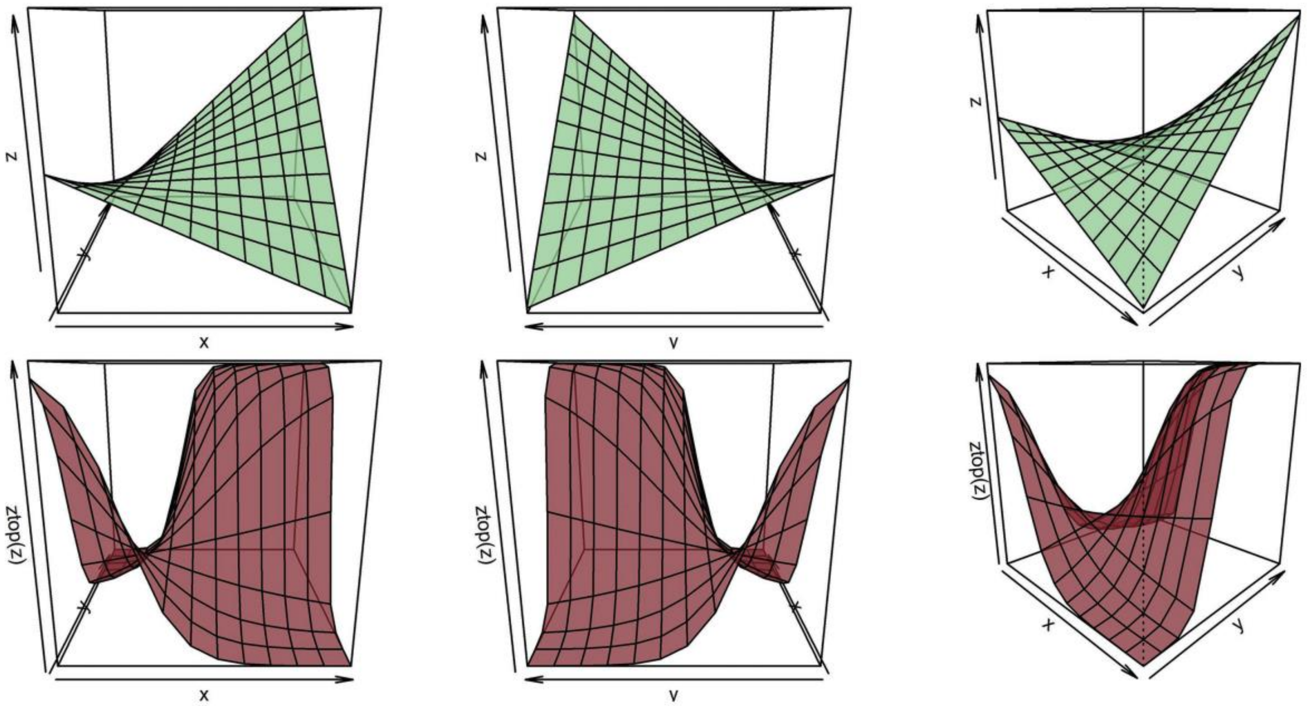
```
library (brms)
library (bmmb)
options (contrasts = c('contr.sum','contr.sum'))

data (exp_data)

# create our dependent variable
exp_data$Female = ifelse(exp_data$G == 'f', 1, 0)

# center vtl
exp_data$vtl_original = exp_data$vtl
exp_data$vtl = exp_data$vtl - mean (exp_data$vtl)

# center and scale f0
exp_data$f0_original = exp_data$f0
exp_data$f0 = exp_data$f0 - mean(exp_data$f0)
exp_data$f0 = exp_data$f0 / 100
```

(Q1) Can we use what we've learned so far to fit and evaluate a logistic regression model with two quantitative predictors?

(Q2) Can we extend the concepts from Chapter 10 to classify speakers along two stimulus dimensions (f0 and VTL) using our logistic models?

Description of the Model

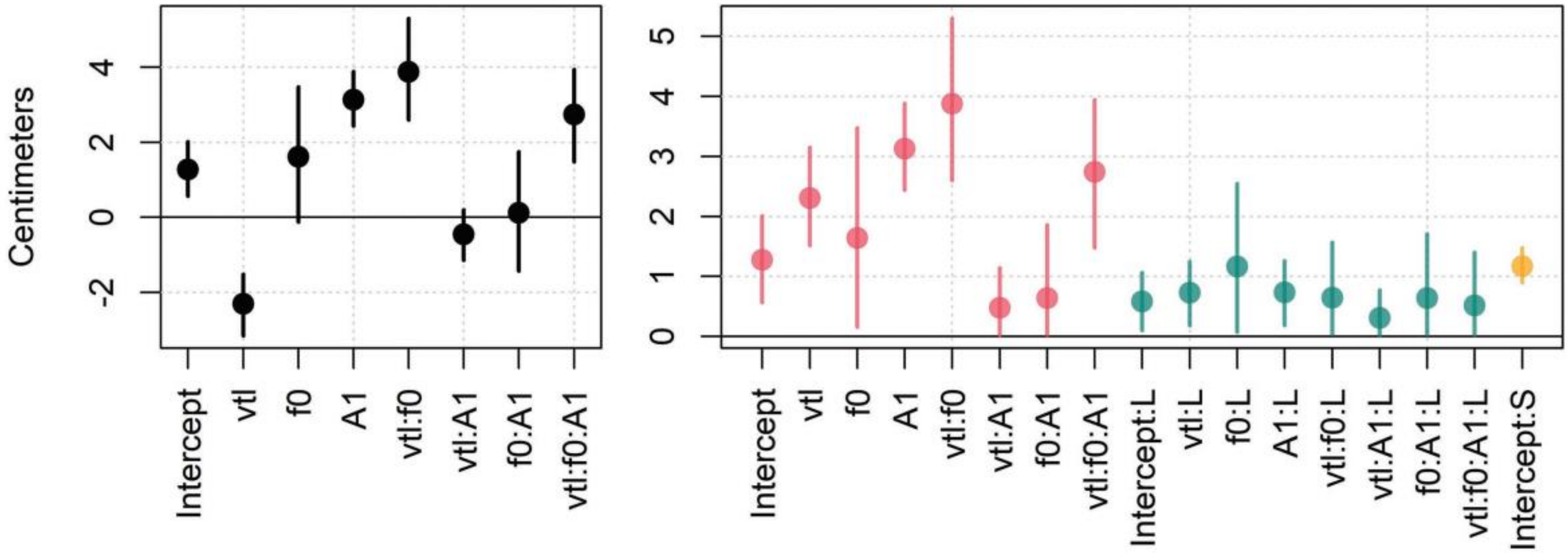
```
Female ~ vtl * f0 * A + (vtl * f0 * A|L) + (1|S)
```



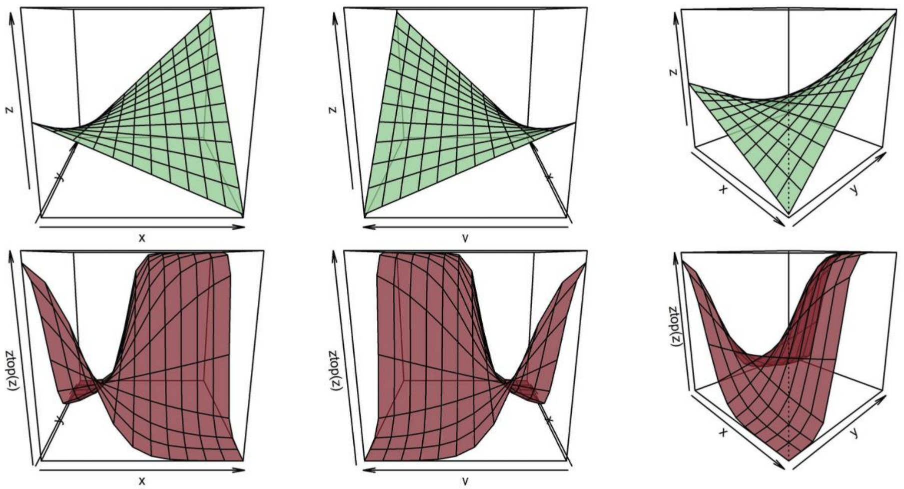
Fitting the Model

```
# Fit the model yourself
model_gender_vtl_f0 =
  brm (Female ~ vtl*f0*A + (vtl*f0*A|L) + (1|S), data=exp_data,
       chains=4, cores=4, family="bernoulli",
       warmup=1000, iter = 5000, thin = 4,
       prior = c(set_prior("student_t(3, 0, 3)", class = "Intercept"),
                 set_prior("student_t(3, 0, 3)", class = "b"),
                 set_prior("student_t(3, 0, 3)", class = "sd"),
                 set_prior("lkj_corr_cholesky (2)", class = "cor")))
```

Fixed Effects and BANOVA Plot

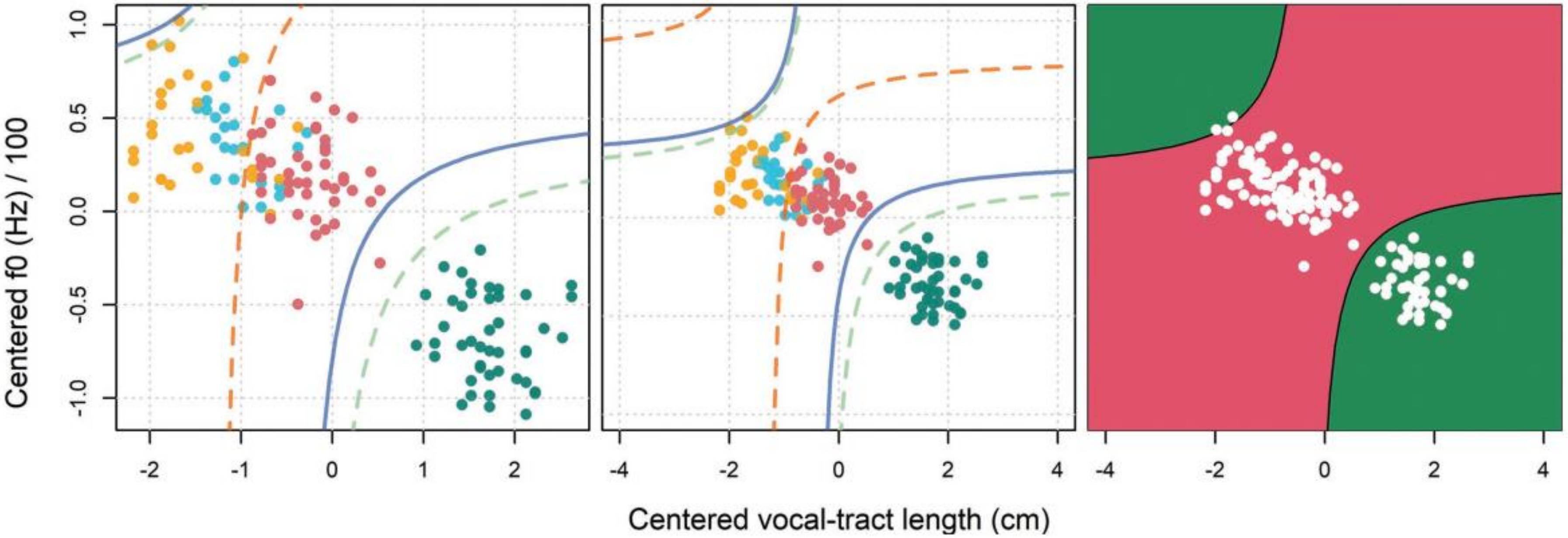


Categorization in Two Dimensions



$$0 = \mathrm{a} + (\mathrm{b} \cdot x) + (\mathrm{c} \cdot y) + (d \cdot x \cdot y)$$

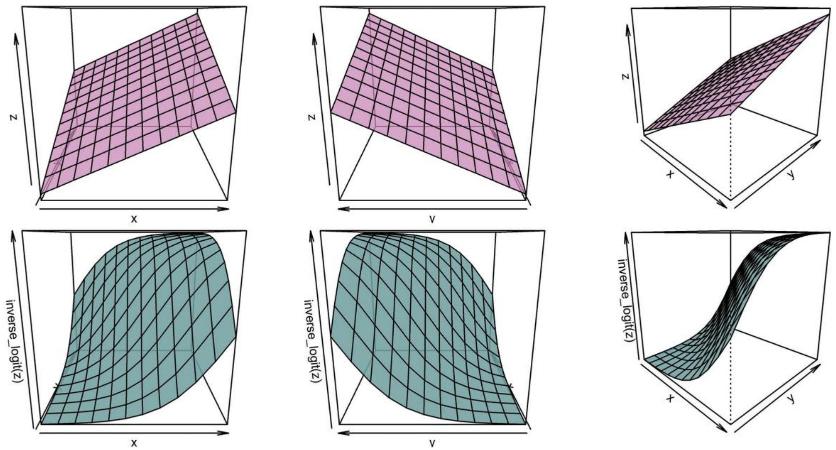
$$y = (-\mathrm{b} \cdot x - \mathrm{a}) / (d \cdot x + c)$$



A Simpler Model

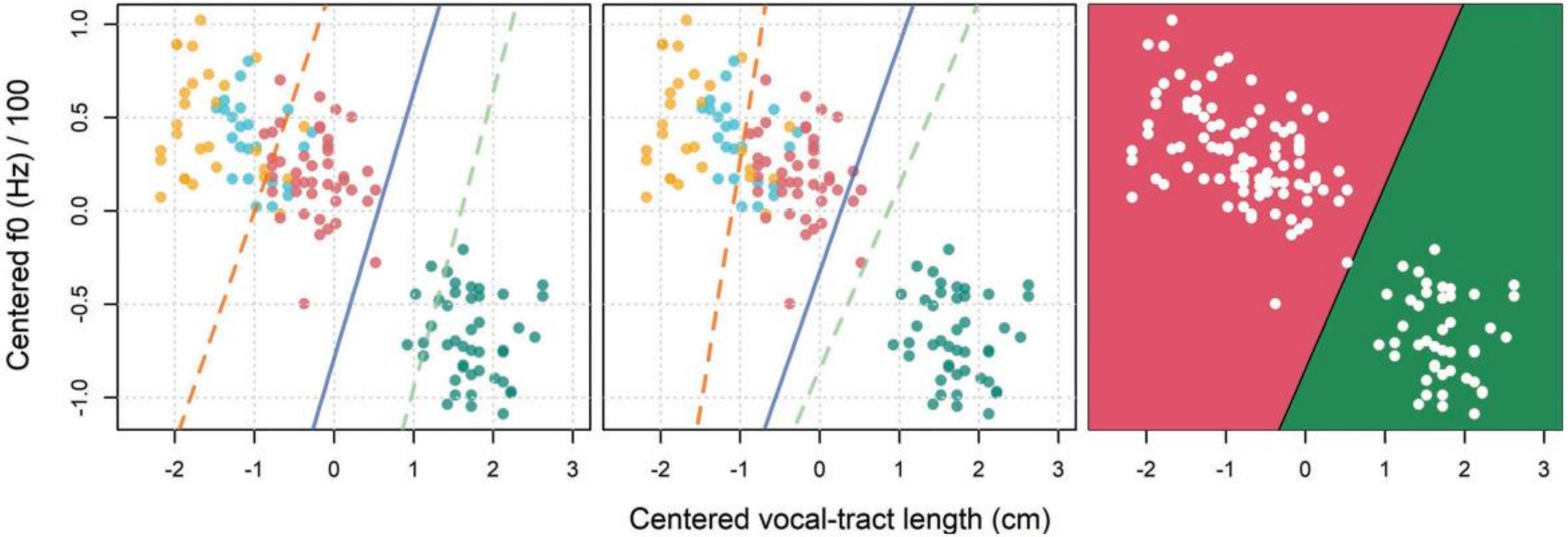
```
# Fit the model yourself
model_gender_vtl_f0_reduced =
  brm (Female ~ (vtl+f0)*A + ((vtl+f0)*A|L) + (1|S), data=exp_data,
       chains=4, cores=4, family="bernoulli",
       warmup=1000, iter = 5000, thin = 4,
       prior = c(set_prior("student_t(3, 0, 3)", class = "Intercept"),
                 set_prior("student_t(3, 0, 3)", class = "b"),
                 set_prior("student_t(3, 0, 3)", class = "sd"),
                 set_prior("lkj_corr_cholesky (2)", class = "cor")))
```

Categorization in Two Dimensions



$$0 = \mathrm{a} + (\mathrm{b} \cdot x) + (\mathrm{c} \cdot y)$$

$$y = -(\mathrm{b} \cdot x - \mathrm{a}) / \mathrm{c}$$



Model Misspecification

```
brms::loo_compare (model_gender_vtl_f0, model_gender_vtl_f0_
reduced)
##                               elpd_diff se_diff
## model_gender_vtl_f0             0.0       0.0
## model_gender_vtl_f0_reduced -18.2       9.9
```

```
## Warning: Found 205 observations with a pareto_k > 0.7 in model
## 'model_gender_vtl_f0'. It is recommended to set 'moment_match = TRUE'
## in order to perform moment matching for problematic observations.
```

Pareto k Values

- A diagnostic of the method used to estimate LOO ELPD, that measures the ‘unexpectedness’ of an observation.

- Values of $\hat{k}$ less than 0.5 are ‘good’, values between 0.5 and 0.7 are ‘ok’ (not so great but not bad either), values greater than 0.7 are ‘bad’, and values greater than 1.0 are ‘very bad’.

- High $\hat{k}$ values indicate a large mismatch between your data and your model.

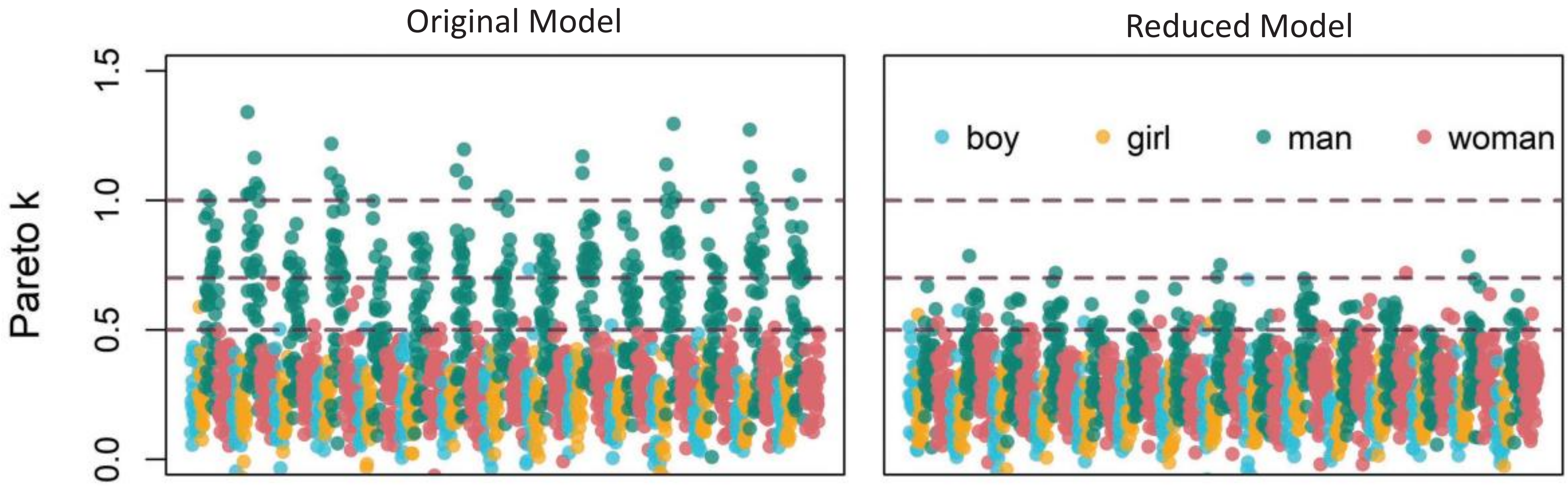
Model Misspecification

1 p_loo is *smaller* than the number of estimated parameters: If the estimated number of parameters (p) is relatively large relative to the number of observations (n), e.g. p > n/5, then the model may be too flexible, or your priors may be too weak.
2 p_loo is *much smaller* than the number of estimated parameters: The model is likely to be **misspecified**. A misspecified model is one whose structure contains important differences compared to the processes being modeled.
3 p_loo is *greater than* the number of parameters: The model is likely to be '*badly*' misspecified.

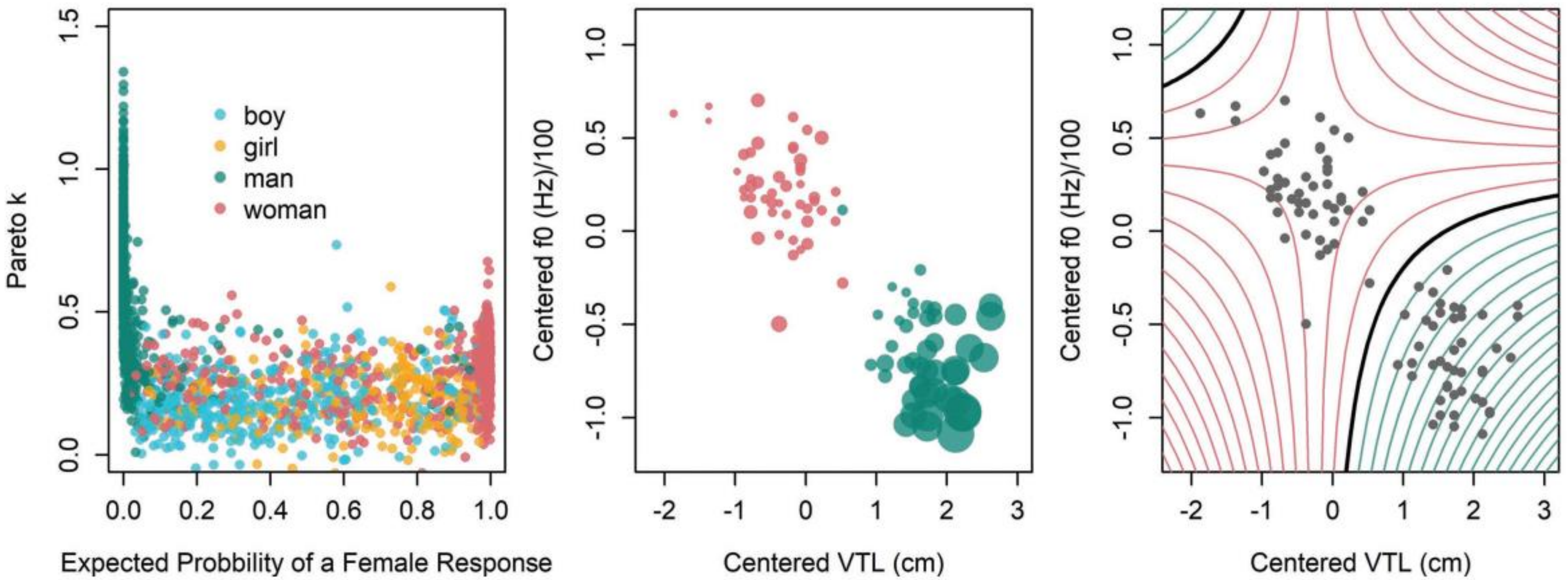
```
# actual number of estimated parameters
ncol (bmmb::get_samples(model_gender_vtl_f0))-2
## [1] 304

# number of observations
nrow (model_gender_vtl_f0$data)
## [1] 2085

# information related to loo creiterion
model_gender_vtl_f0$criteria$loo$estimates
##              Estimate     SE
## elpd_loo       -541.5 23.405
## p_loo           105.8  6.637
## looic          1083.0 46.810
```

Pareto k Values

```
pareto_k = model_gender_vtl_f0$criteria$loo$diagnostics$
pareto_k
```

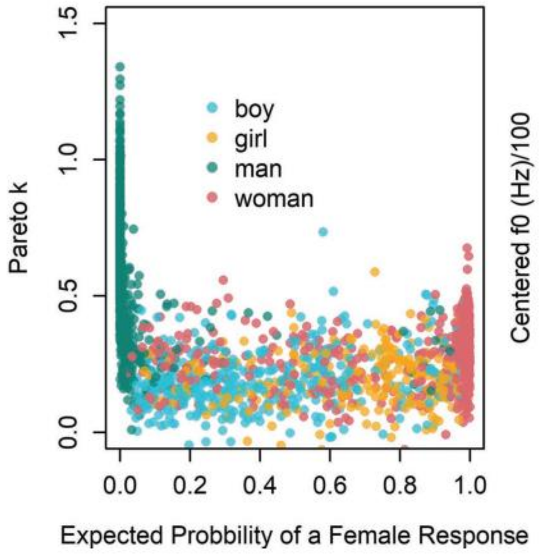


Pareto k Values



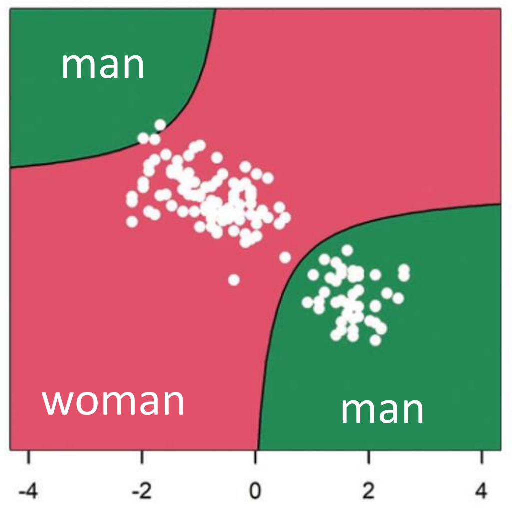
Model Selection: Statistical

- An overly powerful model with weak priors was badly behaved.
- Problems only became clear when adding LOO criterion.
- This is a good way to diagnose ‘invisible’ problems with your (large) models.



Model Selection: Common Sense

- There's no chance that speakers with the highest f0 and shortest vocal tracts are most likely to be perceived as adult males.



- This alone is a reason to be highly suspicious of this model.

- Our simpler model aligns with our domain knowledge and prior expectations, and that alone is a good reason to prefer it.