# Portfolio

Experimental Methods 3: Multilevel models and machine learning

*E22.147201U013.A*

Submission: 15. December 2022

Supervisor: Riccardo Fusaroli

Student: Sara Kjær Kristensen (herein after SK)

Email: 202105320@post.au.dk

Student no: 202105320

Student: Maria Mujemula Olsen (herein after MO)

Email: au695881@uni.au.dk

Studentno: 202106384@uni.au.dk

School of Communication and Cognition, University of Aarhus

Nordre Ringgade 104 4 th, 8200 Aarhus N, Denmark

Parts of this assignment have been made in a group consisting of Freddy Wulf (FW), Ida Brøcker (IB), Maria Mujemula Olsen (MO), Sabrina Zaki (SZ) & Sara Kjær Kristensen (SK). As a default SK and MM are the main writers and group contributions are highlighted initial of the responsible writer.

Character count with spaces w. plots: 33.944

Link to Github repository: https://github.com/Mujemula/SK_MO_methods_3_exam_

Permission to publish assignment: yes

# 1 Portfolio

## 1.1 Introduction (SK)

This portfolio will cover handed out data[1] investigating autistic and neurotypical children's language development over a series of 6 visits with several months apart. At each visit, different measures were taken including the binary diagnostic grouping, either autistic or neurotypical, and the child's mean length of utterance during that visit. To investigate the dataset R (version 4.2.0 & 4.2.1) with RStudio (version 2022.07.1) and statistical packages for Bayesian workflow BRMS (Bürkner, 2021) were used.

## 1.2 Simulating the data at hand (SK, MO)

To better understand the data and the models, that are going to describe it, a new dataset is simulated using the literature estimates as a guideline. The goal of the simulation process is to assess whether the model we make is able to recover true values of the simulation. This way we can better trust our model to find true values when running it on real data. The dataset is structured with the most

[1] From Fusaroli, et al. (2019)

important variables and leave the others for consideration; number of visits (6); number of participants in the two diagnostic conditions (50, sample size = 100); and mean length of utterance (MLU).

To estimate the priors on the simulated data, we used the values given in the assignment as a starting point for the MLU for each diagnostic group at visit 1, average individual variability in this initial MLU, the average change in MLU across visits for the diagnostic groups, and average individual variability for the diagnostic groups. To modify the priors, we make histograms to check if the values give a realistic impression of the data, by trial and error we estimate the priors to be informed priors as they assume a difference in the development of MLU. Generally, we use a lognormal distribution to exclude the possibility of the MLU going below zero.

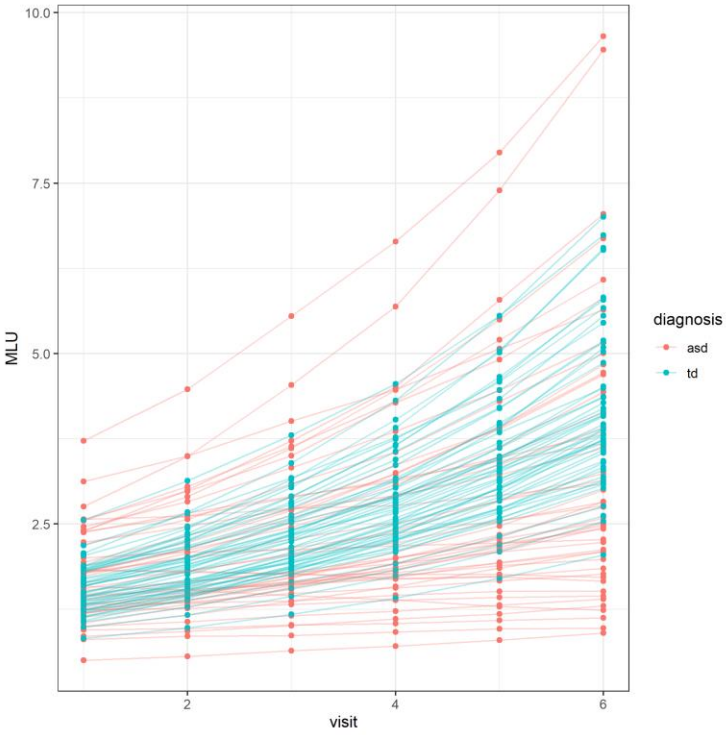



*Figure 1 - Simulated data showcasing MLU development over time*

To assess the simulated data and what predictors are best to describe the change in MLU, we formulate 3 models:

| Model no. | Equation | Description |
|---|---|---|
| 1 | $MLU \sim 0 + diagnosis$ | MLU modulated by diagnosis |
| 2 | $MLU \sim 0 + diagnosis + diagnosis{:}\,visit$ | MLU modulated by diagnosis with a difference in change per visit pending on diagnosis |
| 3 | $MLU \sim 0 + diagnosis + diagnosis{:}\,visit + (1 + visit \mid ID)$ | MLU modulated by diagnosis with a difference in change per visit pending on diagnosis and individual differences per visit |

The priors are specified as needed for the different models and fitted to their model:

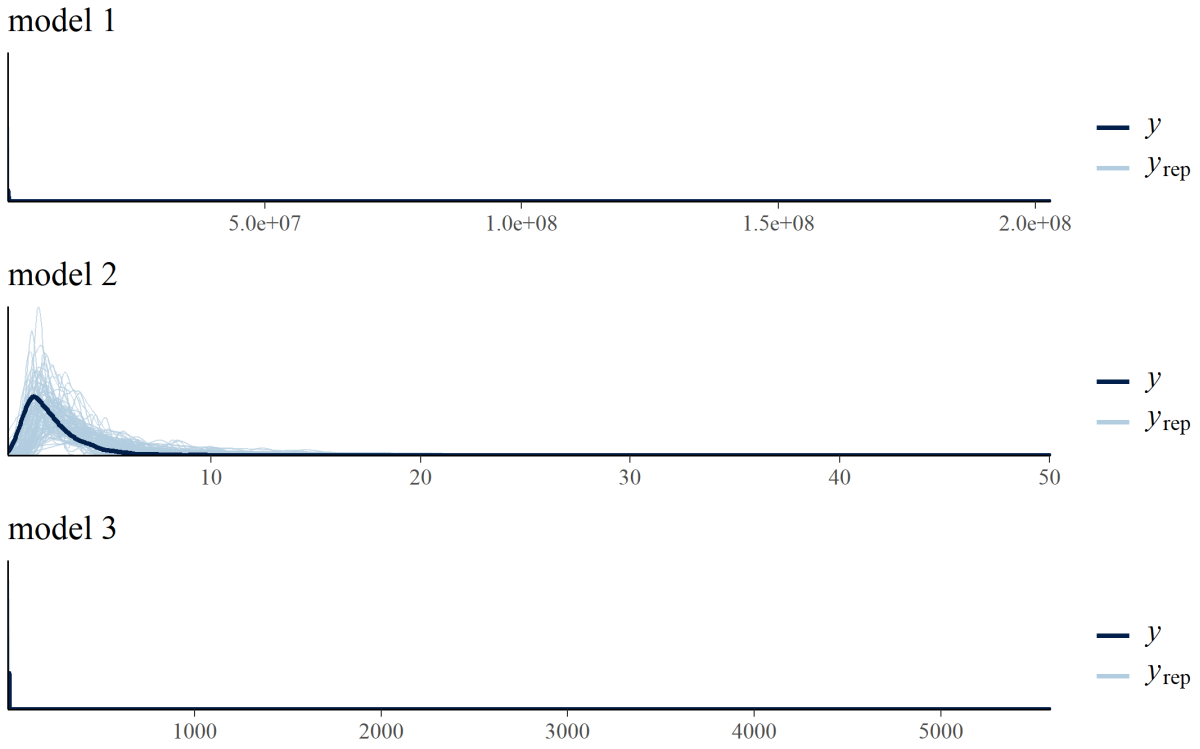



*Figure 2 - Informed priors*

Fitting the data to the posterior distributions gives these prior posterior update checks. This way the data's influence can be inspected.

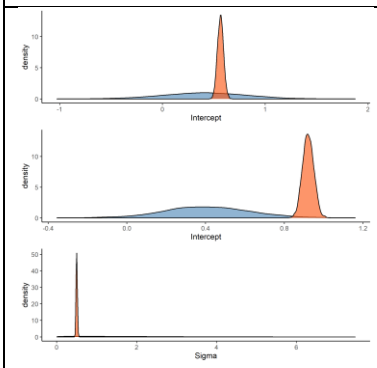
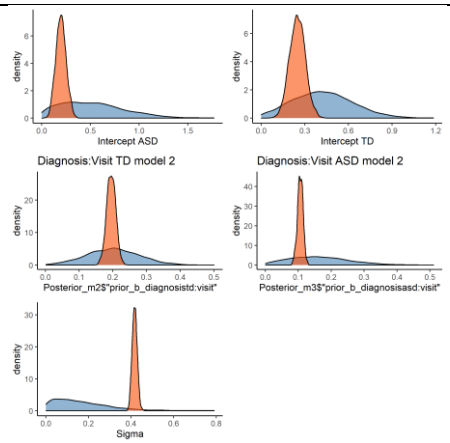
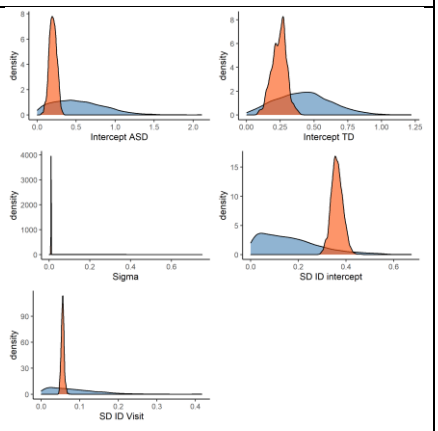
| Model 1 | Model 2 | Model 3 |
|---|---|---|
|  |  |  |

*Table 1 - blue for prior and orange for posterior*

Checking the quality of the models is assessed by looking for a ratio spread for the Markov Chain Monte Carlo (Rhat) value between 0.90 and 1 and as high values as possible for effective sample sizes for both the bulk (Bulk_ESS) and tail (Tail_ESS).

| Model no. | Summary output | Esti-mate | Est.Er-ror | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|---|
| | Population-Level Effect | | | | | | | |
| | diagno-sisasd | 0.56 | 0.03 | 0.51 | 0.62 | 1.00 | 2088 | 1534 |
| 1 | diagno-sistd | 0.92 | 0.03 | 0.8 | 0.98 | 1.00 | 1803 | 1845 |
| | Family Specific Parameters | | | | | | | |
| | Sigma | 0.50 | 0.01 | 0.47 | 0.53 | 1.00 | 1839 | 1667 |
| 2 | Population-Level Effect | | | | | | | |
| | diagno-sisasd | 0.20 | 0.05 | 0.10 | 0.30 | 1.00 | 1320 | 1262 |
| | diagno-sistd | 0.25 | 0.05 | 0.14 | 0.35 | 1.00 | 1207 | 1205 |
| | diagno-sisasd:visit | 0.11 | 0.01 | 0.08 | 0.13 | 1.00 | 1288 | 1887 |
| | diagno-sistd:visit | 0.20 | 0.01 | 0.17 | 0.22 | 1.00 | 1552 | 1374 |
| | Family Specific Parameters | | | | | | | |
| | Sigma | 0.42 | 0.01 | 0.40 | 0.44 | 1.00 | 2009 | 2251 |
| 3 | Population-Level Effects | | | | | | | |
| | diagno-sisasd | 0.20 | 0.05 | 0.11 | 0.29 | 1.01 | 103 | 239 |
| | diagno-sistd | 0.24 | 0.05 | 0.14 | 0.35 | 1.05 | 44 | 78 |
| | diagno-sisasd:visit | 0.11 | 0.01 | 0.09 | 0.12 | 1.01 | 117 | 132 |
| | diagno-sistd:visit | 0.20 | 0.01 | 0.18 | 0.21 | 1.13 | 16 | 152 |
| | Group-Level Effects (~ID) | | | | | | | |
| | sd(Inter-cept) | 0.36 | 0.02 | 0.31 | 0.41 | 1.02 | 105 | 167 |
| | sd(visit) | 0.06 | 0.00 | 0.05 | 0.06 | 1.02 | 144 | 295 |
| | cor(Inter-cept:visit) | 0.03 | 0.09 | -0.14 | 0.23 | 1.02 | 114 | 179 |
| | Family Specific Parameters | | | | | | | |
| | Sigma | 0.01 | 0.00 | 0.01 | 0.01 | 1.00 | 754 | 1206 |

*Table 2 - Comparing model summary output*

Model 1's parameters is overall okay, but the estimates are very different from our priors. Model 2's parameters are better both for estimates and mcmc specific parameters. Lastly, model 3's estimates are fine, but the mcmc specific parameters are concerning; the Rhat is above 1, Bulk_ESS is in many cases below 100 samples same goes for the Tail_ESS. Keeping these values in mind, divergency plots are consulted:

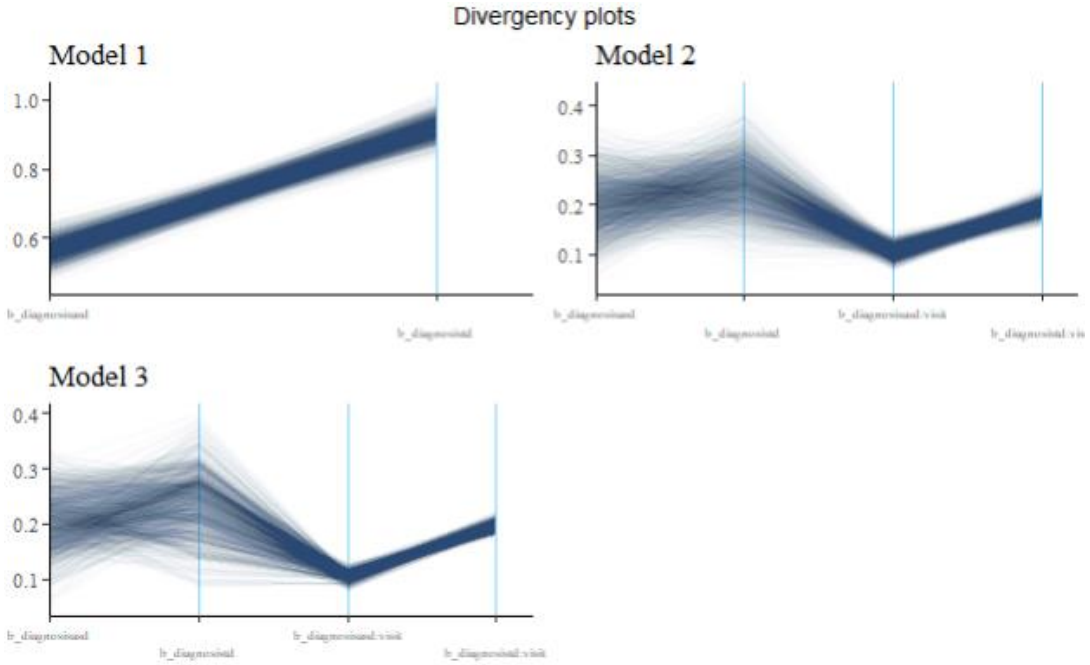



*Figure 3 - Divergency plots*

The divergency plots are not that use, since their show no divergency. Consulting different seeds, my group, model comparison with leave-one-out, suggests model 3 as the best model:

| Model no. | elpd_diff | se_diff | weight |
|---|---|---|---|
| 1 | -2270.2 | 23.8 | 0.000 |
| 2 | -2166.4 | 26.1 | 0.000 |
| 3 | 0.0 | 0.0 | 1.000 |

*Table 3 - Model comparison with LOO*

Expected log pointwise predictive density difference (elpd_diff) compares models in ascending order, and the negative numbers suggest that the function favours the previous. Logically follows that model 3 is the better one as the leave-one-out log score for predictive density (weight) is maximized by using this model solely. Since the simulated data was generated with such model, it makes good sense it would fit the best. Taking all of this into account and adding the theoretical layer, it makes sense to have children within a certain diagnostic group developing at their own pace. On top of that, the group's general view of model 3 as being the best describing model. Model 3 is chosen for further empirical analysis.

Besides comparing models, the sample size might have a substantial effect on the model's estimate. As the simulated data is already using 3 times as many data points as the empirical data for practical reasons. However, this is an unrealistic picture to paint as real-world data collection would properly not contain 50 participants per condition:

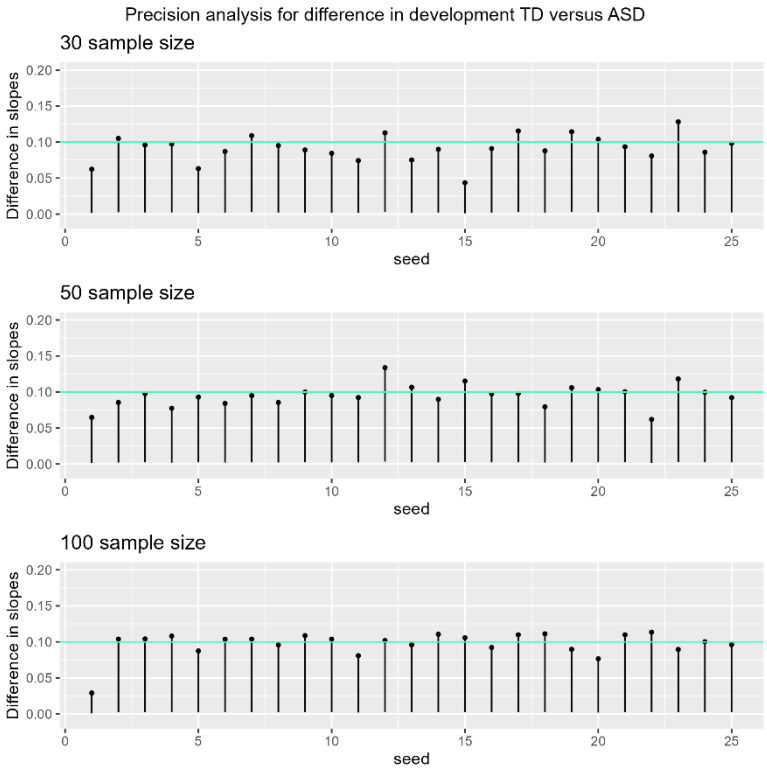



*Figure 4 - simulation of sample size effect on estimated difference in language development compared from TD to ASD and across seeds. The horizontal line is the theoretical difference. The vertical lines resemble a 95% confidence interval.*

These simulation plots give a slight insight into how different seeds for random sampling for 30, 50 and 100 participants per group gives different estimates of the difference in the children's language development across visits. Since the difference is in percentage it appears to be a bit skewed. Transforming the difference to a lognormal scale, could solve this problem. For the ease of interpretation we will keep this scale. Overall, more data points give more certainty, but 30 participants per group seem to be a sufficient place to start. Looking at the power estimates does not give further enlightenment:

| Variable | Power estimate |
|---|---|
| Diagnosis ASD | 1 |
| Diagnosis TD | 1 |
| Diagnosis ASD : visit | 1 |
| Diagnosis TD : visit | 1 |

*Table 4 - Power estimates*

Besides simulating across seeds, we changed the priors to weakly informed prior to make the data work more to make a convincing case of change and difference between visits and diagnostic groups. On that note, it is time to dive into the empirical data.

## 1.3 Empirical data (MO, SK)

Like many other cases, the simulation does not capture all the noise in the data set. Comparing the simulated and empirical data the difference is clear as is the general tendency in language development.

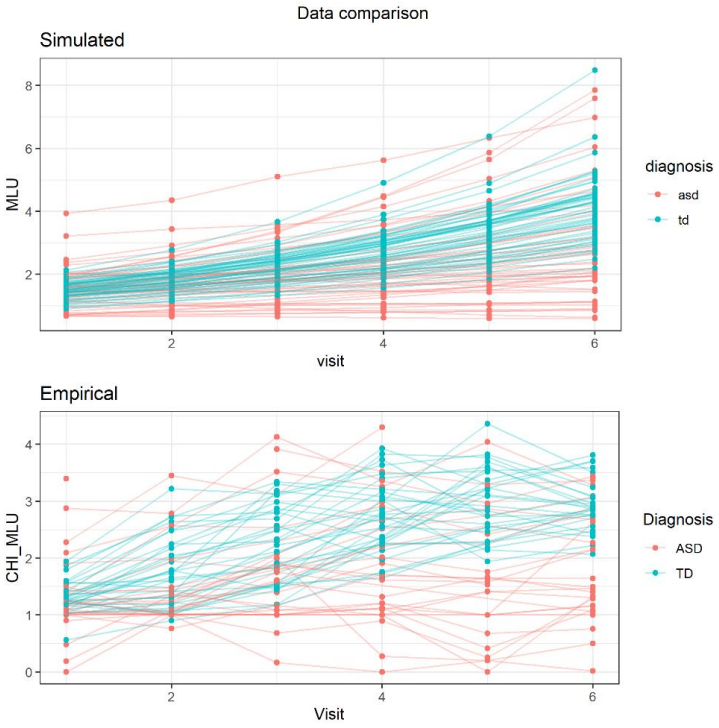



*Figure 5 - comparing data's MLU development over time*

Again, our prior belief is updated to a posterior distribution of the estimates, this time model 3's parameters are updated on the empirical data. Prior-posterior update plots are as follows:

Interestingly, there is not much overlap between the distributions of the posteriors for the slopes for ASD and TD subjects with TD subjects having a higher slope than ASD subjects. Before we dive into the exciting difference in change in MLU over time, as indicated in plot *Diagnosis:Visit*, we will take a look at the model's qualities.

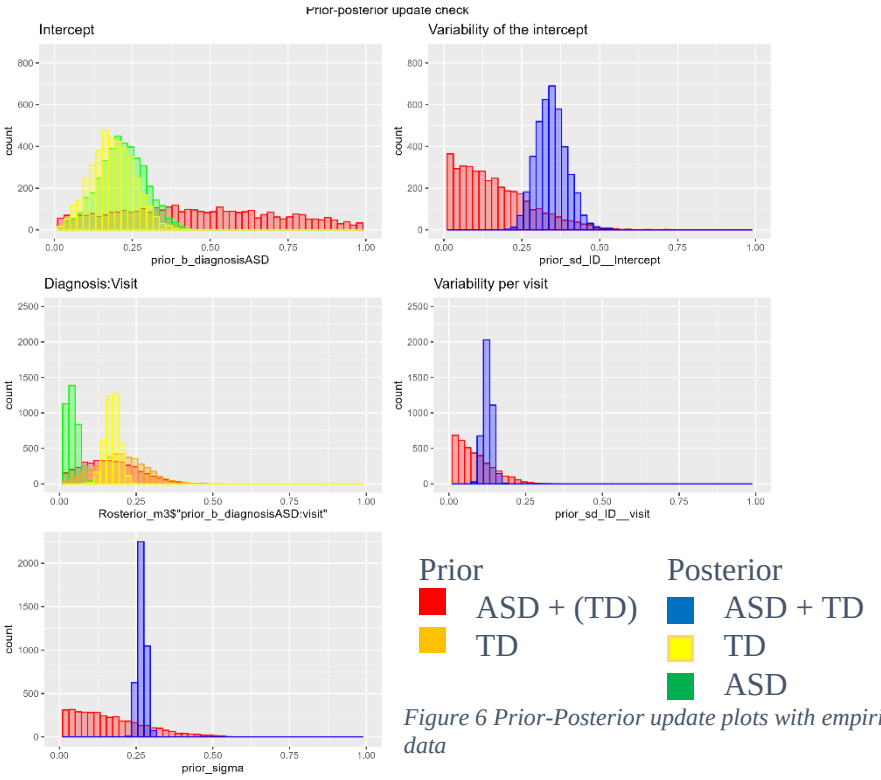



*Figure 6 Prior-Posterior update plots with empirical data*

First, the model:

$$MLU \sim 0 + diagnosis + diagnosis{:}\,visit + (1 + visit|ID)$$

predicts the mean length of utterance (MLU), calculated as the mean of morphemes in the utterance divided by the number of words in the utterance during the visit (30 min), by the binary diagnostic grouping (autism spectrum disorder (ASD) or typically developed (TD)), interaction effect of diagnostic group and visit as participants in the different groups may vary differently in acquired language over time, and lastly with individual random effects for each participant over visits. This way the participants' MLU can account for just bad days or overall smaller changes in MLU. The model's summary highlights the Rhat, Bulk_ESS and Tail_ESS as mentioned before.

| Group-Level Effects | | | |
|---|---|---|---|
| Parameters | Rhat | Bulk_ESS | Tail_ESS |
| sd(Intercept) | 1 | 1811 | 2460 |
| sd(visit) | 1.01 | 534 | 1021 |
| cor(Intercept,visit) | 1.01 | 345 | 885 |
| Population-Level Effects | | | |
| diagnosisASD | 1.00 | 1563 | 1491 |
| diagnosisTD | 1.00 | 1654 | 2022 |
| diagnosisASD:visit | 1.00 | 1112 | 1713 |
| diagnosisTD:visit | 1.00 | 975 | 1739 |
| Family Specific Parameters | | | |
| Sigma | 1.00 | 3017 | 3015 |

*Table 5 - Model summary*

It seems like the effective sampling size for both bulk and tail are relatively low, compared to the number of iterations of 4000. Which can explain the Rhat measures above 1. Due to limited resources and the fact that we have started out with 2000 iteration with same result, we will keep it at this. With this concern we consult sensitivity plots for the variability over time for the two diagnostic groups to inspect our prior's impact on the posterior estimates:

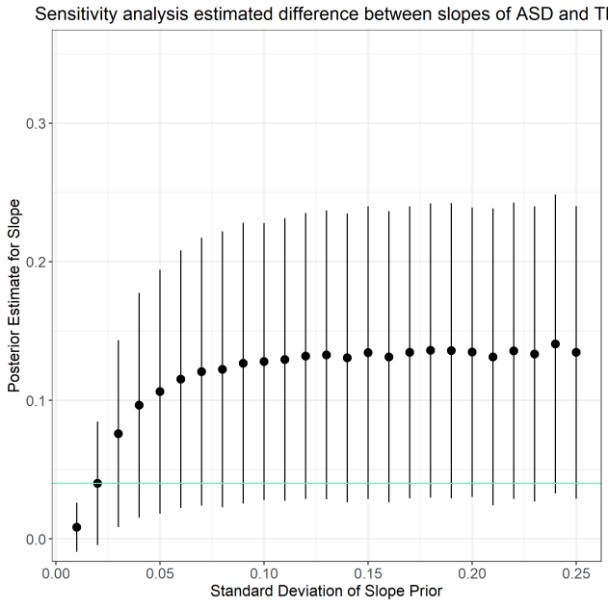



*Figure 7 - Sensitivity check of the estimated difference for variability over time for ASD and TD (MU).*

The informed prior's standard deviation specified for our model and model summary above was set to 0.1 and 0.08 for ASD and TD respectively. In figure 7 a sensitivity plot of the estimated difference of the variability of MLU over time for between ASD and TD is depicted. We see that the "true" estimated difference (marked by the green line), based on the simulation process, will be captured by the model regardless of what we set the standard deviation of the priors to (0.01-0.25). However, the estimated difference for the model gets closer to the true difference the smaller the standard deviation is. It is also worth noting that using informed priors based on this sensitivity plot, as the parameter values we then get does not come from the data, but are somewhat chosen by the experimenter.

Now returning to plot *Diagnosis:Visit* in the prior posterior update checks, the model summary is investigated.

| Prior | | | Posterior | | | |
|---|---|---|---|---|---|---|
| Estimate | Error | Variable | Estimate | Error | l-95% CI | u-95% CI |
| Population-level | | | | | | |
| 0.41 | 0.41 | diagnosisASD | 0.21 | 0.08 | 0.07 | 0.36 |
| 0.41 | 0.22 | diagnosisTD | 0.18 | 0.08 | 0.03 | 0.32 |
| 0.15 | 0.1 | diagnosisASD:visit | 0.04 | 0.02 | 0.00 | 0.09 |
| 0.2 | 0.08 | diagnosisTD:visit | 0.17 | 0.02 | 0.13 | 0.22 |
| Group level | | | | | | |
| 0 | 0.2 | sd(Intercept) | 0.34 | 0.05 | 0.25 | 0.44 |
| 0 | 0.1 | sd(visit) | 0.13 | 0.01 | 0.10 | 0.16 |
| | | cor(Intercept,visit) | -0.46 | 0.14 | -0.68 | -0.15 |

*Table 6 - Prior and posterior estimates for population and group level effects*

To test the first hypothesis about whether there is a difference in the development across diagnostic groups, the hypothesis() function is used for "diagnosisTD:visit is greater than diagnosisASD:visit". The output gives an estimate of 0.13 with a 95% credible interval of 0.08-0.18, and posterior probability of a 100%. The posterior probability reflects sampling 1 random participant from each group

and finding that the TD has a higher MLU is 100 % likely. This result is reflected in the prior-posterior update plots for diagnosis(ASD/TD):visit above and in the conditional effect plot below.

To test the second hypothesis for the starting value across diagnostic groups, same function for "diagnosisASD < diagnosisTD". The output of 0.03 with a 95% credible interval of -0.13-0.20, and posterior probability of 40%. Formulating last the hypothesis as the individual development for ASD is smaller than the development for TD. The output gives an estimate of 0.13 with a 95% credible interval of 0.08-0.18, and posterior probability of a 100% across all children. Which funny enough is the same as for the population-level. However, all in all this indicates that there is a high probability that the diagnostic slopes are statistically different at the population-level:

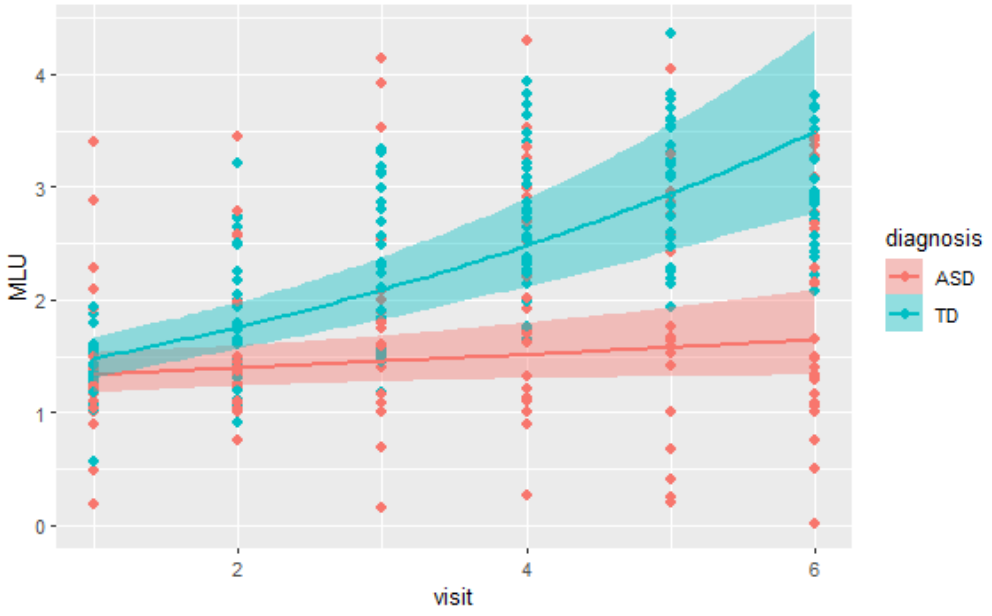



*Figure 8 - conditional effects of the development of MLU over time depending on the diagnostic group*

Some additional factors would be interesting to add to the model, as the child's own effect on her language can hardly account for her linguistic development. E.g., the mother's MLU would be interesting to consider as well as it we know that children learn from their environment and if the environment or environment controller (in this case the mother) decides what the child is exposed to and therefore decides the syllabus on the upbringings linguistic course. If the child is inhibited from exposure to more diverse linguistic environments, the child could quickly reach a plateau in the

linguistic acquirement or make up her own language. The latter part is probably more likely if a 3-year-old could make up her own language and syntax. On the other hand, if the child's environment is stimulating and has a lot to offer, then the development in MLU would reflect more on the child's ability to acquire linguistic concepts rather than it being a lack of resources.

# 2 Portfolio

In this assignment a meta-analysis on the acoustic traits of patients with schizophrenia has been conducted. A Bayesian analysis pipeline is set up and conducted on a simulated dataset, whereafter the meta-analysis data from Parola et al (2020) is run through the pipeline. This is done in order to examine the current evidence for distinctive vocal patterns in schizophrenia. Furthermore, the role of publication bias will be discussed throughout the report.

## 2.1 Bayesian analysis on simulated data

### 2.1.1 Simulated effect sizes of pitch difference for schizophrenic and control participants (MO)

We start by simulating some data. We simulate a data set of 100 studies, making sure that the participants follow a normal distribution with a mean of 20 and a standard deviation of 10 with a minimum of 10 participants. The effect size of the simulated data is set to 0.4 with an average deviation of 0.4 and a measurement error of 8. A column with a publication bias is also simulated in the dataset, so that only papers with a positive effect size greater than two standard deviations are noted as published. This is done because it is expected that papers with high effect sizes in the positive direction are more likely to be published, as. To account for p-hacking, three outliers with a high effect size is added to the dataset.

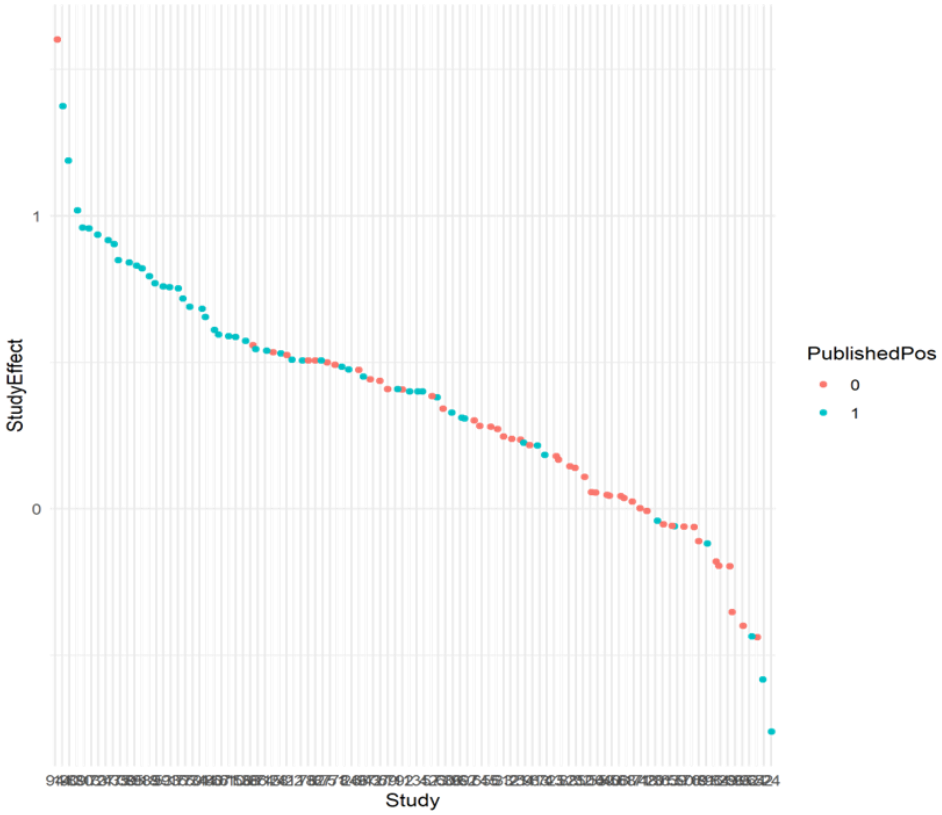



*Figure 9 - Plot of the study effect of each of the 100 simulated papers. The dots colored pink indicate non-published papers, and the dots colored blue indicate published papers.*

Through visual inspection of figure 9 it is seen that fewer papers are categorized as published the lower the study's effect size get. It is also seen that the majority of papers with an effect size bigger than 0.5 are marked as published. This tendency also becomes clear though visual inspection of figure 2. Here it is seen that the density of the histogram of the mean effect sizes for the published papers is lower than the histogram of the mean effect sizes for all the studies, and studies with lower effect sizes are not present in the former histogram.

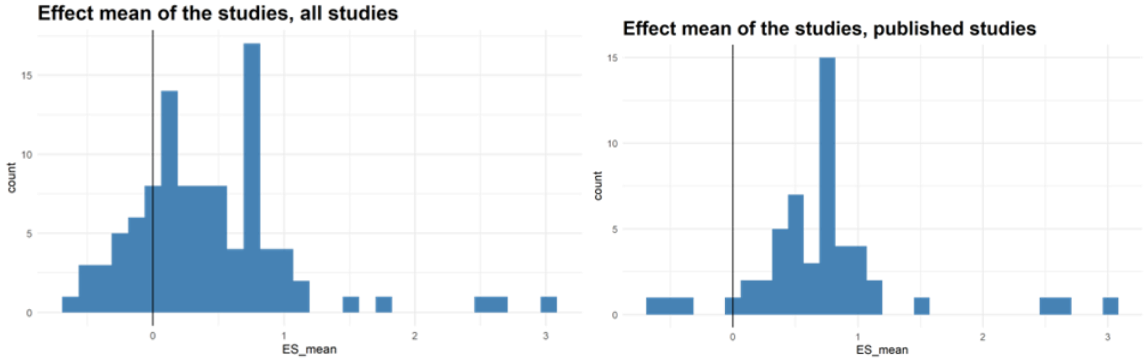



*Figure 10 - Histograms over the mean effect size of the simulated papers*

### 2.1.2 Setting up a Bayesian pipeline (MO, SK)

A Bayesian pipeline is set up using the simulated data in order to get insight into the problem at hand. Based on given literature, the following model is made to explain the data.

*Effect Size | se(Standard Error) ~ 1 + (1|Study_ID)*

The model tells us that the effect size of a study is predicted by the standard error, and that the intercept varies by study ID. We then iteratively plotted and adjusted the weekly informed priors.
The model was first run on all the simulated data and afterwards the model was run in a subset of the simulated data that only contained the papers categorized as published according to the publication bias. Predictive checks were then made as well as prior-posterior update plots.

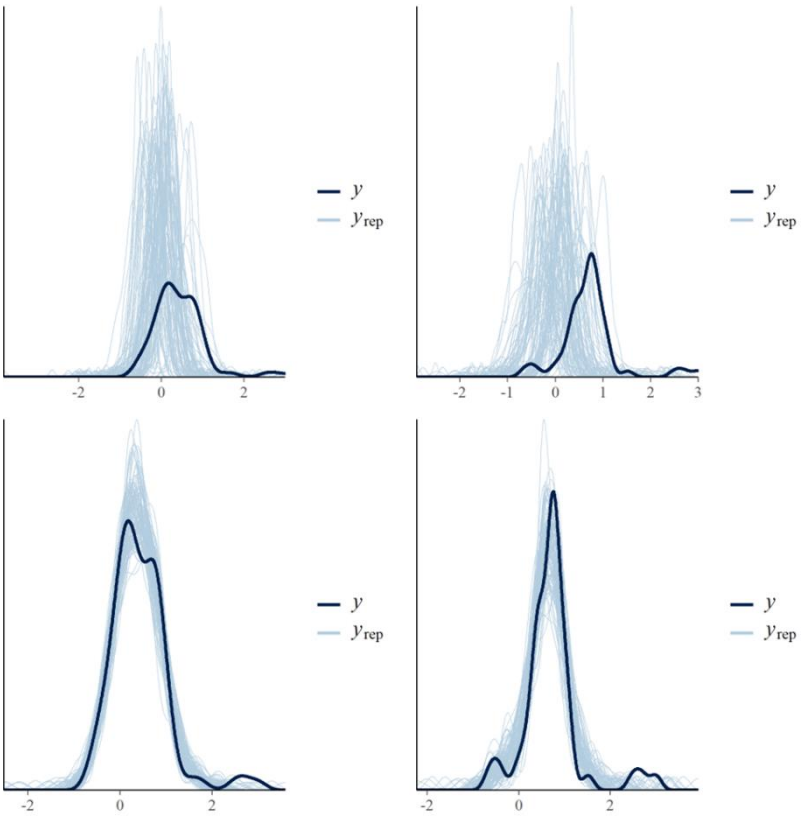



*Figure 11 - pp-checks for all simulated data and for simulated data marked as published. Top left: prior predictive check for all simulated data. Top right: prior predictive check for simulated data marked as published. Bottom left: posterior predictive check f*

The prior predictive checks are accepted, as are the posterior predictive checks.

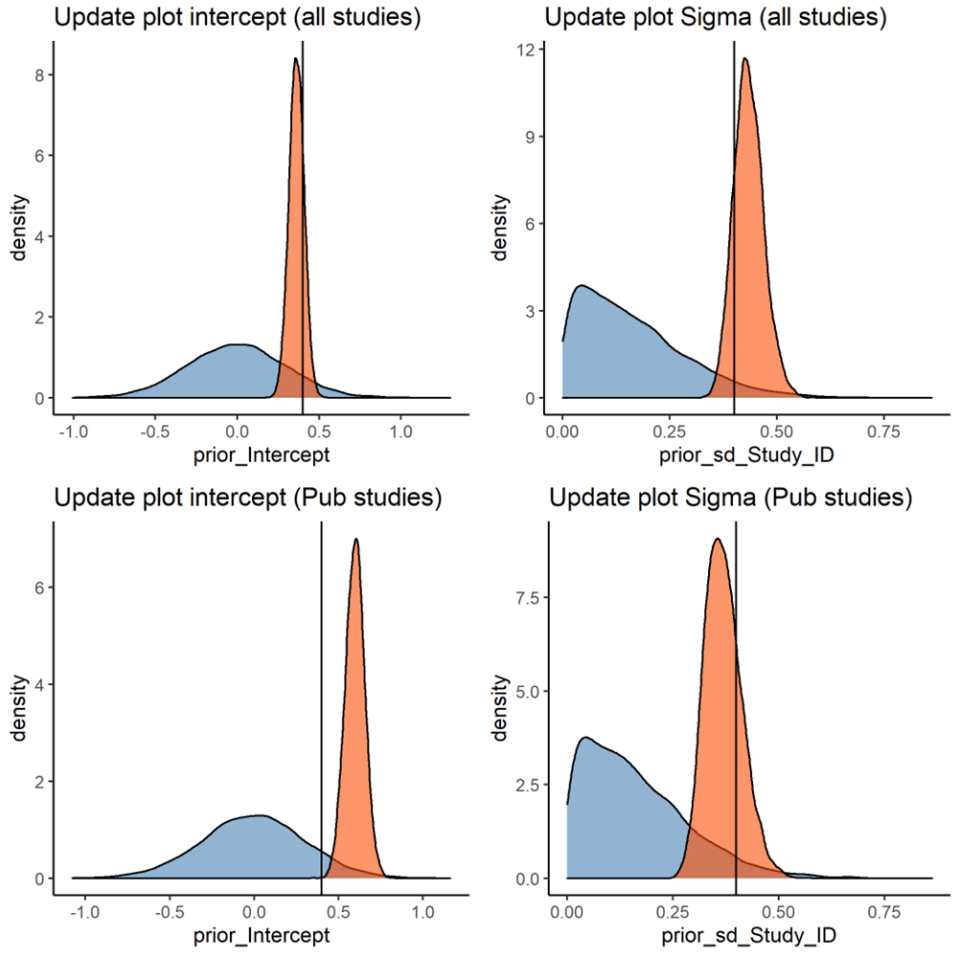



*Figure 12 - Update plots for fitted models. The blue distributions are prior distributions, and the orange distributions are posterior distributions. Top left: update plot of the intercept (effect size) for all studies. Top right: update plot of sigma for all studies*

Looking at the update plots in figure 12, the fitted model for all studies and for published studies have learned from the model as the posterior distribution for both sigma and the intercept for each model is relatively peaked. The black vertical line indicates the theoretical effect size (intercept) and standard deviation used to simulate the data. It is worth noting that the posterior distribution for the effect size of published studies does not capture the theoretical effect size in contrast to the posterior distribution for the effect size of all the studies.

Looking at the summery outputs for the two fitted models in figure 5 and 6, it is seen that the estimate of the effect size (intercept) for the published data is higher much higher for the published studies (0.60) than for all the studies (0.36). As the "true" effect size of the simulated data is set to 0.4, the fitted model for all the studies comes much closer to the "true" effect size, where the fitted model for the published studies varies quite a bit. Both models converged fully as they have a Rhat of 1.00, and the bulk as well as the tail values are accepted for both models.

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: ES_mean | se(ES_SE) ~ 1 + (1 | Study_ID)
   Data: d (Number of observations: 100)
  Draws: 2 chains, each with iter = 5000; warmup = 1000; thin = 1;
         total post-warmup draws = 8000

Group-Level Effects:
~Study_ID (Number of levels: 100)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.43      0.03     0.37     0.51 1.00     1856     2844

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     0.36      0.05     0.27     0.45 1.00     1067     2597

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.00      0.00     0.00     0.00   NA       NA       NA
```

*Figure 13 - Summary output of the fitted model for all studies*

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: ES_mean | se(ES_SE) ~ 1 + (1 | Study_ID)
   Data: d_pub (Number of observations: 52)
  Draws: 2 chains, each with iter = 5000; warmup = 1000; thin = 1;
         total post-warmup draws = 8000

Group-Level Effects:
~Study_ID (Number of levels: 52)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.37      0.04     0.29     0.46 1.00     2041     3644

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     0.60      0.06     0.49     0.71 1.00     1103     2005

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.00      0.00     0.00     0.00   NA       NA       NA
```

*Figure 14 - Summary output of fitted model for published studies*

Based on this analysis, when taking publication bias into account, the estimated effect size is more likely to reflect a desired effect size, more specifically a high effect size in a positive direction. But the published studies are less likely to contain the true effect size. This means, that based on the meta-analysis on our simulated data, we could expect that studies finding high effect sizes in pitch variability between schizophrenic patients and neurotypical controls are more likely to be published. However, these studies are less likely to reflect the true underlying effect size.

## 2.2 Bayesian analysis on real data

Now that we have some theoretical background and that we have gained insight on the implication of publication bias, we will run the data from Parola et al (2020) through the Bayesian analysis pipeline set up in the previous sections.

### 2.2.1 Describing the data (SK)

The dataset from Parola et al (2020) contains 57 different articles published between 1977 and 2018 which have measured the pitch of healthy controls (HC) and schizophrenic patients (SZ). Only those studies with a sample of both HC and SZ are included in the following analysis in order to calculate effect size (Cohen's D) between the two groups. 12 studies were eligible. Sample sizes for HC had a mean of 33.5 participants and standard deviation of 23.5; SZ with a mean of 43.21 participants and standard deviation of 20.2. Lastly the pitch for HC with a mean of 24.2 and standard deviation of 12.9; SZ with a mean of 20.6 and standard deviation of 12.9 as well.

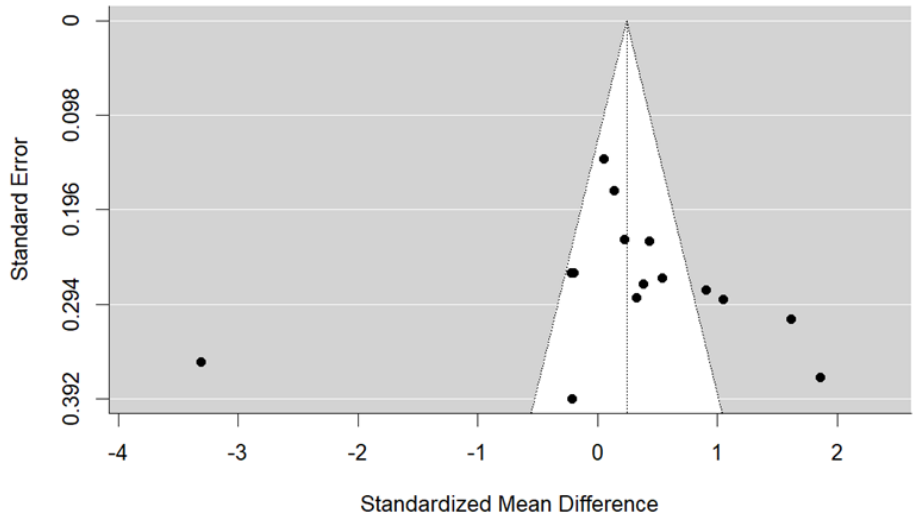



*Figure 15 - Funnel plot*

A funnel plot is made to visually inspect the possibility of the existence of a publication bias (see figure 15). In figure 15 we see effect size (standardized mean difference) from each study against the precision of the estimate. From the plot we shall be aware of a possible publication bias, as the effect size estimates are asymmetrically distributed. However, the following Bayesian analysis will spread more light upon this issue.

### 2.2.2 Bayesian analysis on real data (MO, SK)

After slightly adjusting the previously used priors, the data is fitted to the same model made for the simulated data. Predictive checks are made on the model, and both the prior predictive check and the posterior predictive check are accepted.

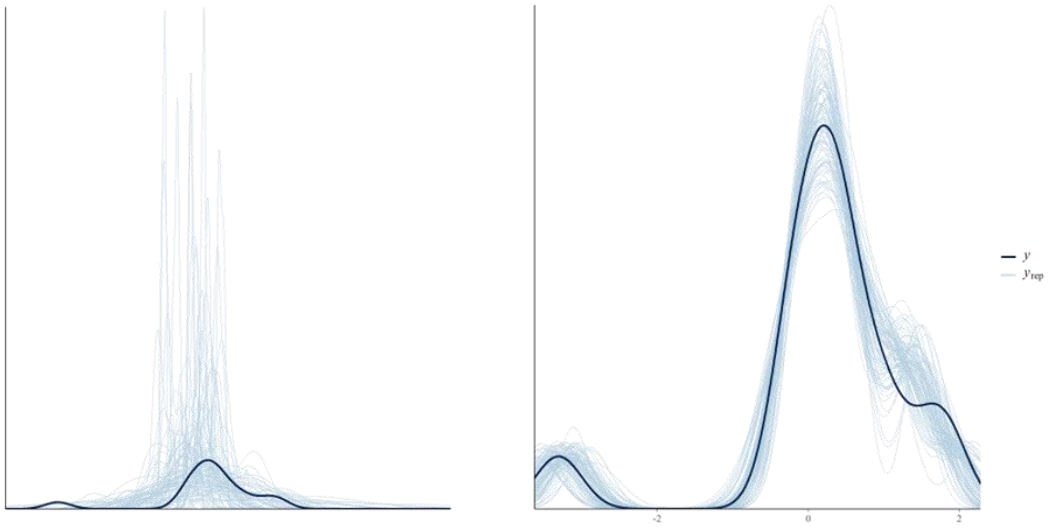


*Figure 16 - Predictive checks for the fitted data. Left: prior predictive check. Right: posterior predictive check*

.

Update plots for the effect size (intercept) and sigma is then made which can be seen in figure 16. The model has learned from the priors, though the posteriors for the intercept and sigma are not that peaked.

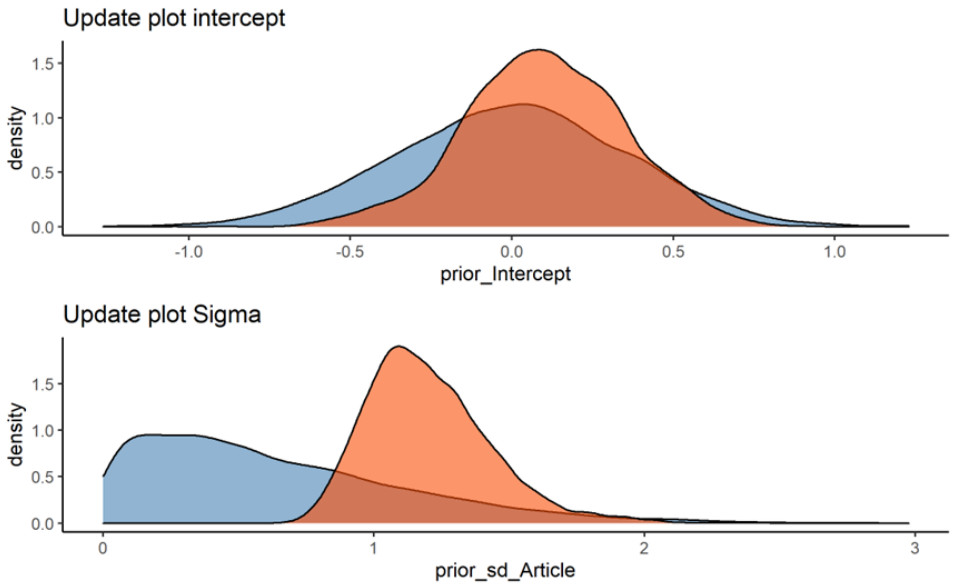



*Figure 17 - Update plots for the effect size (intercept) and sigma. Top: update plot for effect size. Bottom: update plot for sigma.*

Looking at the summary output for the fitted model, the estimated effect size (intercept) is 0.10 with an estimated error of 0.25. 95% of the effect sizes will lie between -0.41 and 0.58. The tail values are above 1000 and are acceptable given the 2 chains of 5000 iterations. However, the bulk values are a bit low (1000<), meaning that the model has not explored that much before converging. The Rhat values are accepted as they are either 1.00 or close enough to that.

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: yi | se(vi) ~ 1 + (1 | Article)
   Data: pitch_varience (Number of observations: 15)
  Draws: 2 chains, each with iter = 5000; warmup = 1000; thin = 1;
         total post-warmup draws = 8000

Group-Level Effects:
~Article (Number of levels: 12)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     1.20      0.23     0.84     1.75 1.01      991     1593

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     0.10      0.25    -0.41     0.58 1.00      686     1118

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.00      0.00     0.00     0.00   NA       NA       NA

Draws were sampled using sample(hmc). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

*Figure 18 - Summary output of the fitted model*

This meta-analysis is visualized in the forest-plot depicted in figure 19 of the mean standardized difference.

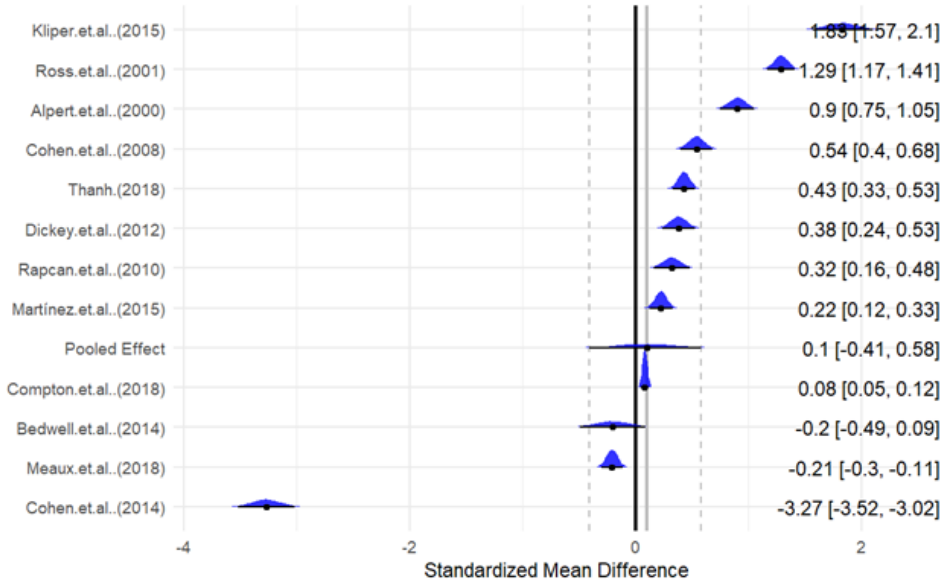



*Figure 19 - Forest plot of the meta-analysis*

The pooled effect shows the population-level effect size with a grey vertical line going through the mean (0.10), and grey dotted vertical lines indicating the upper- and lower 95% confidence intervals (-0.41, 0.58). The individual mean and sigma for each of the studies is depicted. Only looking at the pooled effect, it would not seem that there is much evidence for a distinctive vocal pattern in schizophrenia patients. However, looking at each study, 11 out of 12 studies' confidence intervals do not contain 0, meaning that these studies all find a difference in pitch between SZ and HC. But taking into account that this plot is heavily right skewed as well as the asymmetry in the funnel plot from figure 15, it is likely that publication bias may be influencing the results of our model. From the analysis on our simulated data, one should be cautious when making inferences about there being a distinctive vocal pattern in schizophrenia patients based on this meta-analysis, as we saw that there was a low probability of the true effect size being precent in the simulated published studies.

# 3 Portfolio

### The assignment (SK)

The Machine Learning assignment has 3 main parts: First we create a skeptical and an informed simulation, based on the meta-analysis. Second, we build and test our machine learning pipeline on the simulated data. Third, we apply the pipeline to the empirical data.

In this assignment we will establish a machine learning pipeline to investigate the relationship between several vocal variables and people diagnosed with schizophrenia. We use data from Parola et al. (2020). See section 2.2.1 for further description of the data, or see the paper for further variable description and general interest.

## 3.1 Simulating data (SK, MO)

First step for us is to determine what model to use. Therefore, we simulate two datasets; one using Parola et al. (2020) estimates and standard deviations; one using sceptical parameters. Both sets have 100 matched pairs of schizophrenia and controls, each participant producing 10 repeated measures (10 trials with their speech recorded). The informed dataset has 6 acoustic measures from the meta-analysis and 4 measures representing noise: pitch mode (PitchMode), pitch variability (PitchVar), proportion of spoken time (ProSpotime), speech rate (SpeechRate), number (NumPause) and length (LenPause) of pauses. The sceptical simulation only contains noise measures.

After finding the variables, we set the error (0.2), standard deviation for trial (0.5) and individual variance (1). Looping through all parameters and variables gave us a data frame for the informed and sceptic datasets. Figure 20 visualizes the differences from data frames, diagnosis group and variable:

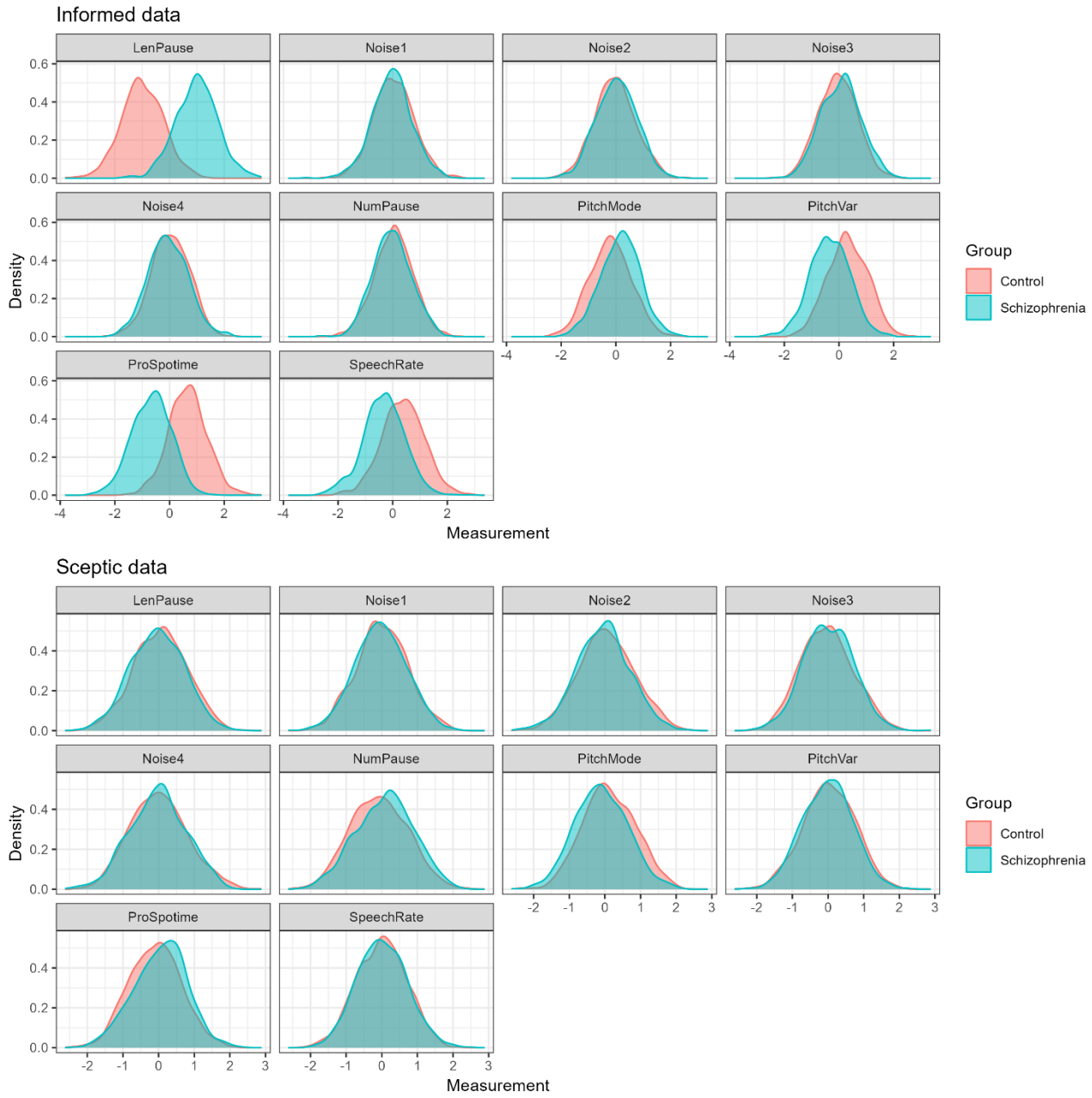



*Figure 20 - Simulated data with comparing participant group estimates for informed and sceptic datasets.*

## 3.2 Setting up machine learning pipeline on simulated data (SK, MO)

We setup our machine learning pipeline as depicted in the figure below and separately ran the sceptic and informed datasets through the pipeline. We will use R (version 4.2.1) and the packages brms, cmdstanr, tidymodels and tidyverse (see references) for a Bayesian approach to this pipeline.

Workflow

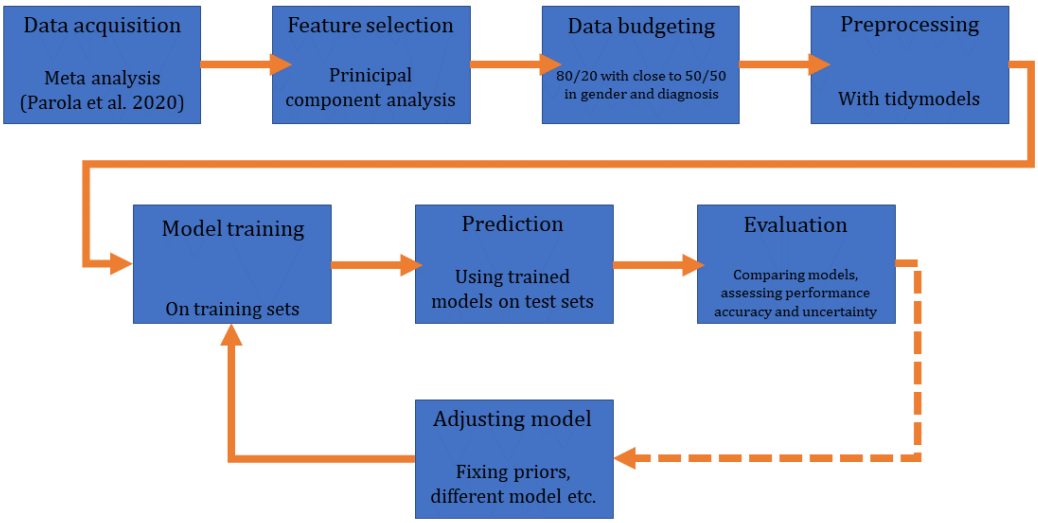



*Figure 21 - Workflow for machine learning. Stipples line suggests this step might be needed.*

We start of by splitting our data with an 80/20 ratio for training and test sets for both datasets. Next, we pre-process all four datasets by standardizing all measures. Then we set up three different logistic models: a baseline model with fixed effects, a model with varying intercept per participant, and lastly a model with varying intercepts for participants and slopes per measure.

| Model | Description |
|---|---|
| Varying intercepts/slopes | Diagnosis ~ 1 + PitchVar+SpeechRate+ProSpoTime+NumPause+LenPause+Noise1+Noise2+Noise3+Noise4 + (1 + PitchMode+PitchVar + SpeechRate + ProSpoTime + NumPause + LenPause + Noise1 + Noise2 + Noise3 + Noise4 \| ID) |
| Varying intercepts | Diagnosis ~ 1 + PitchVar+SpeechRate+ProSpoTime+NumPause+LenPause+Noise1+Noise2+Noise3+Noise4 + (1 \| ID) |
| Baseline | Diagnosis ~ 1 + PitchVar+SpeechRate+ProSpoTime+NumPause+LenPause+Noise1+Noise2+Noise3+Noise4 |

To all our models, we fix normally distributed priors. Following, we fit our models using the Bernoulli family. After the fit we asses prior posterior update plots for all relevant variables. The common green colour is for posterior, and the common red colour is for the prior distribution for all intercept plots.

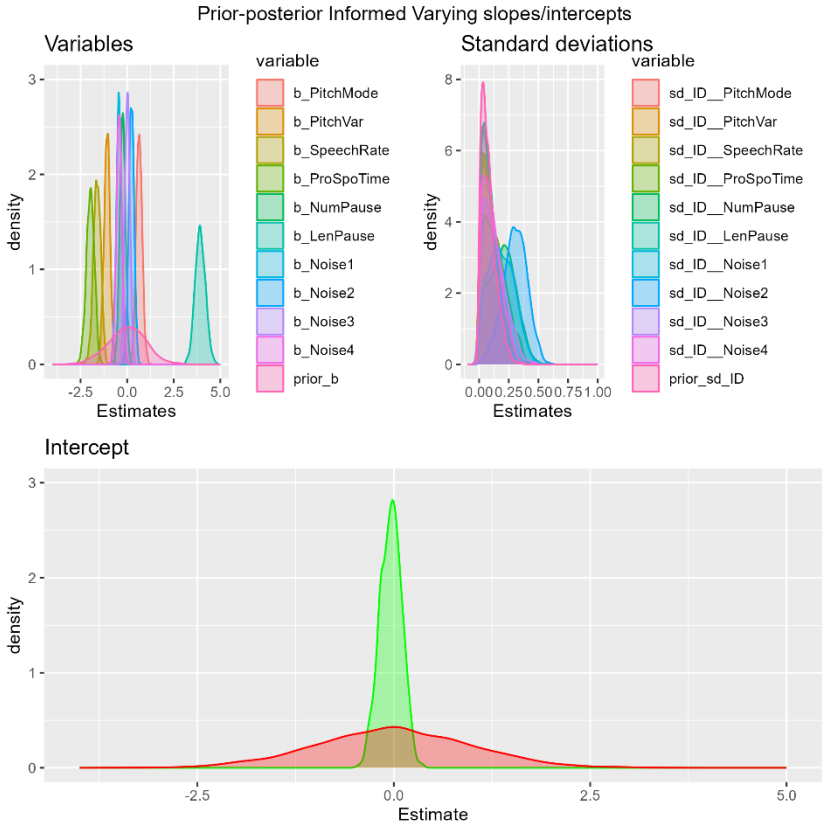



*Figure 22 - Informed training dataset fitted to model with varying intercepts and slopes*

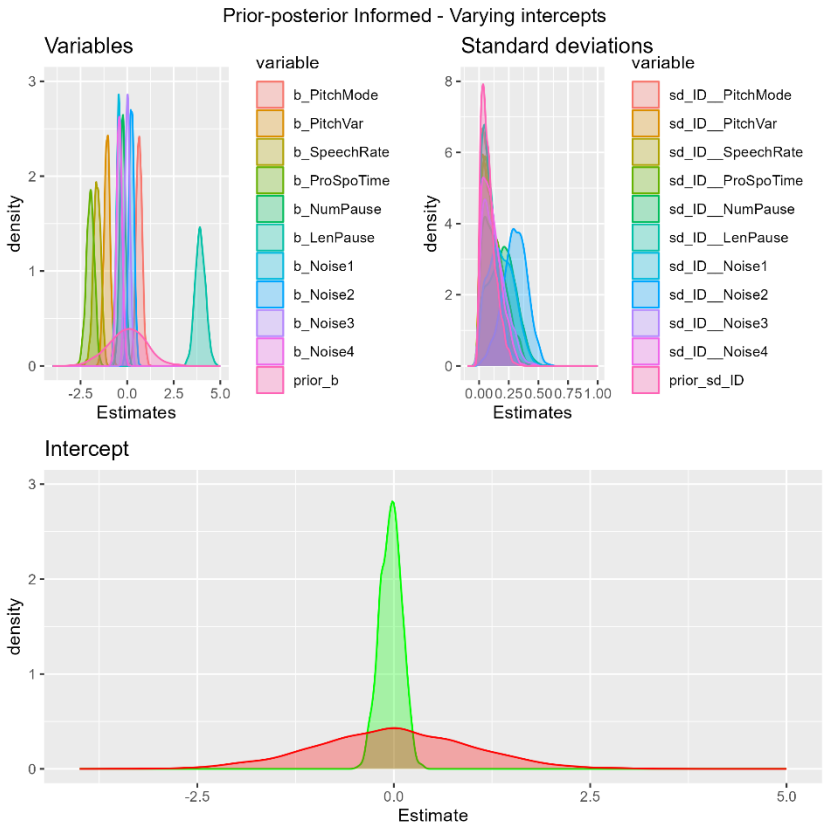



*Figure 23 - Informed training dataset fitted to model with varying intercepts*

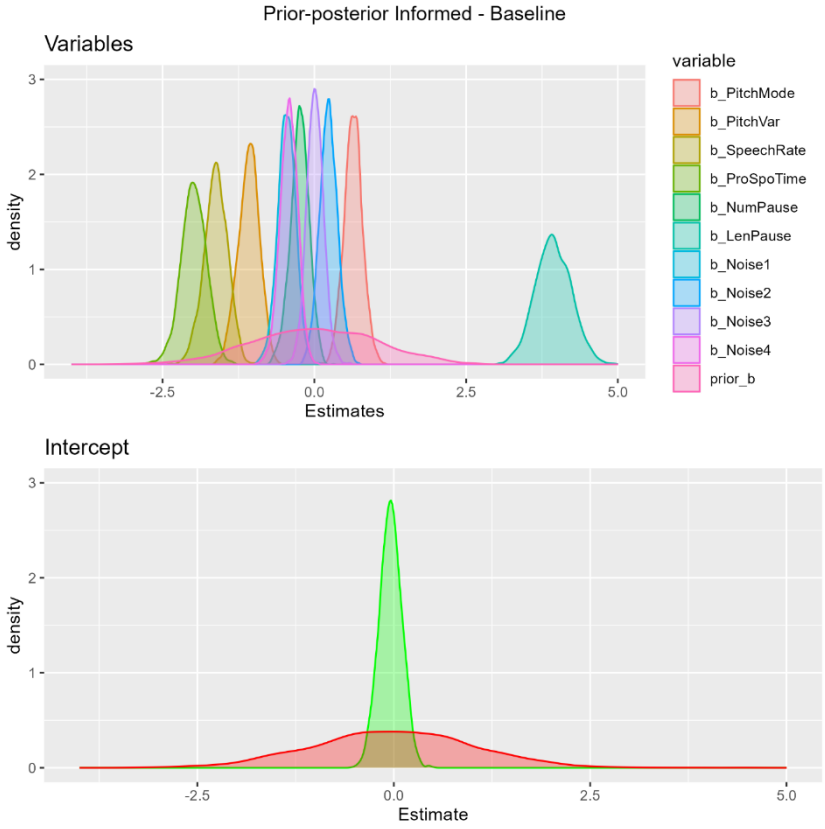



*Figure 24 - Informed training dataset fitted to baseline model*

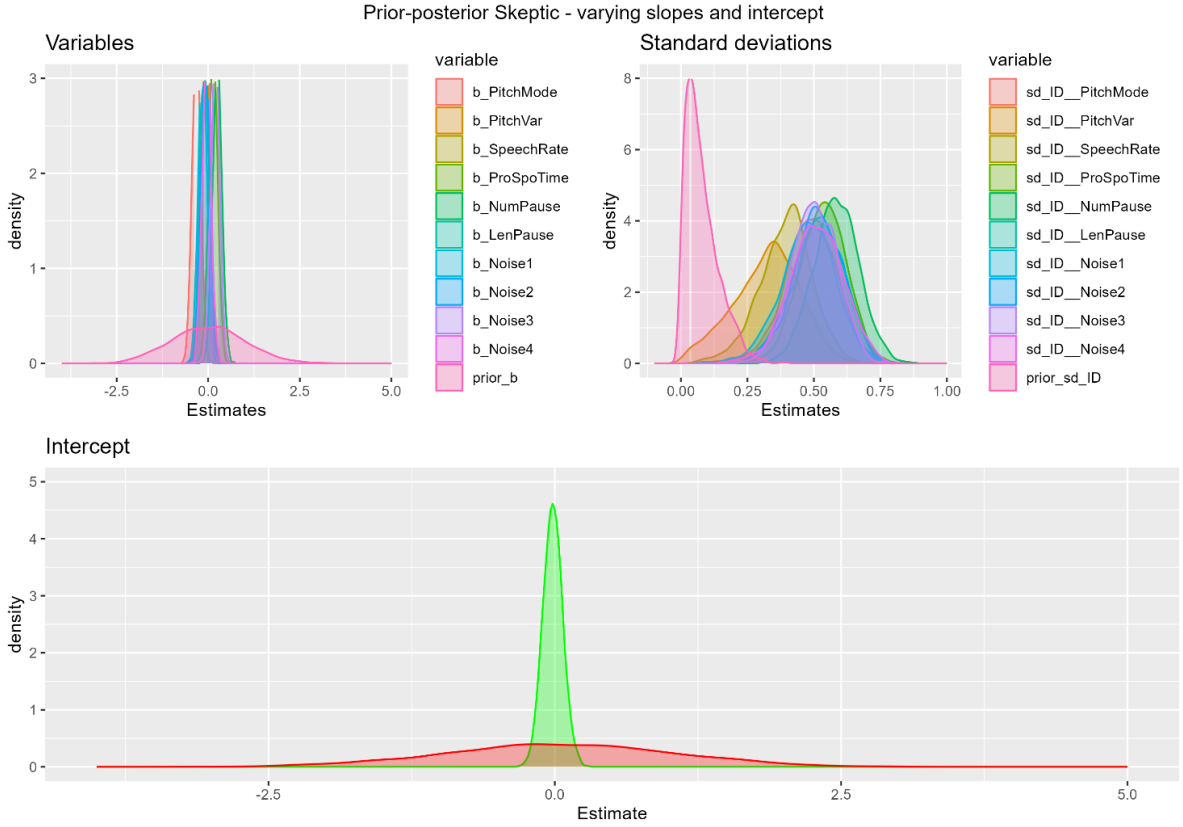



*Figure 25 - Sceptic training dataset fitted to model with varying slopes and intercepts*

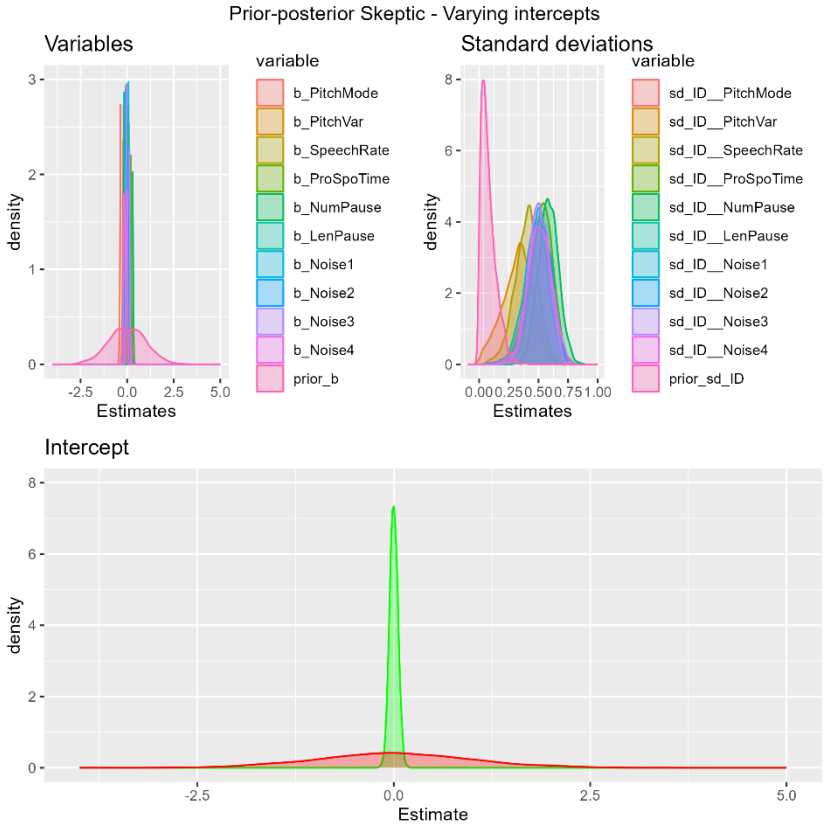



*Figure 26 - sceptic training dataset fitted to model with varying intercepts*

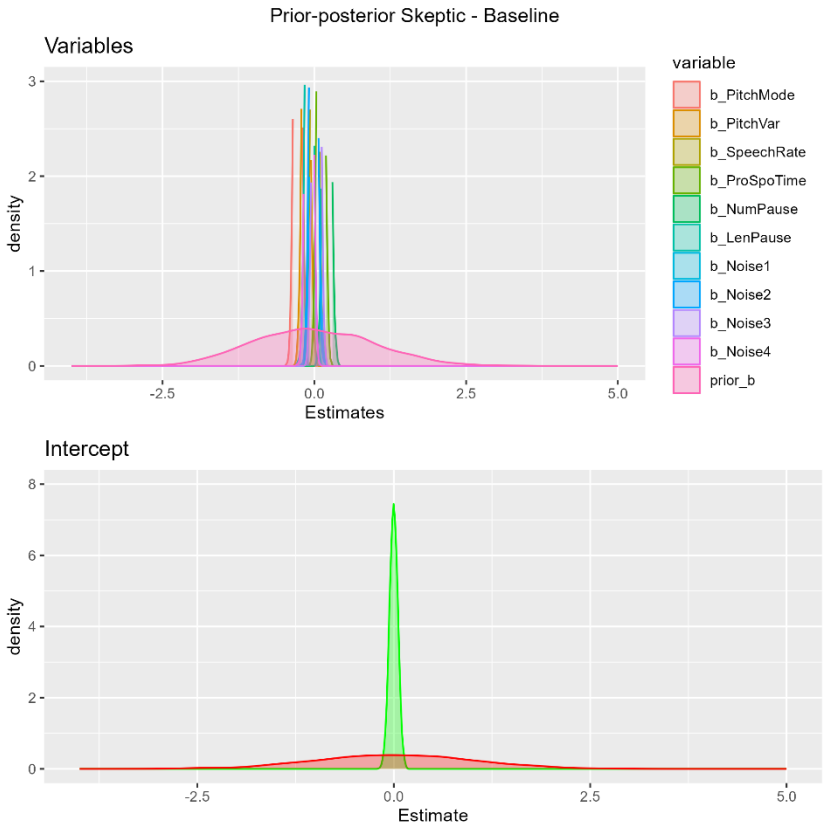



*Figure 27 - Sceptic training set fitted to baseline model*

As our workflow suggest it would be an idea to finetune the priors before continuing as some variables are pushing the tails of our prior distribution. However, most of the posterior distribution look acceptable, we therefore accept our priors.

Moving on, we compare the different models using leave one out (LOO) weighted method.

| Informed | | | | Sceptic | | |
|---|---|---|---|---|---|---|
| Elpd_diff | Se_diff | Weight | Model | Elpd_diff | Se_diff | Weight |
| 0.0 | 0.0 | 1.0 | Varying slopes and intercepts | 0.0 | 0.0 | 1.0 |
| -32.1 | 4.0 | 0.0 | Varying intercepts | -785.2 | 8.8 | 0.0 |
| -32.0 | 4.0 | 0.0 | Baseline | -784.1 | 8.8 | 0.0 |

Based on the LOO comparison, the varying slopes and intercept model seem to be performing the best, secondly the baseline and lastly the varying intercepts model. As the first model contains all the information possible, we are curious to see if really does perform better than the baseline model.

To further compare the models, we assess the summary outputs:

***Informed***

*Varying intercepts and slopes model summery output with population-level effects and an excerpt of group-level effects.*

```
Group-Level Effects:
~ID (Number of levels: 80)
                      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)             0.07      0.05     0.00     0.20 1.00     1103      630
sd(PitchMode)             0.09      0.07     0.00     0.26 1.00      911      626
sd(PitchVar)              0.10      0.08     0.00     0.29 1.00      796      932
sd(SpeechRate)            0.10      0.08     0.00     0.27 1.00      746      562
sd(ProSpoTime)            0.14      0.10     0.01     0.35 1.00      572      602
sd(NumPause)              0.19      0.11     0.01     0.40 1.00      548      482
sd(LenPause)              0.09      0.07     0.00     0.25 1.00     1300      833
sd(Noise1)                0.19      0.11     0.02     0.40 1.01      537      524
sd(Noise2)                0.30      0.10     0.09     0.49 1.01      880      701
sd(Noise3)                0.13      0.09     0.01     0.33 1.00      722      634
sd(Noise4)                0.11      0.08     0.00     0.28 1.00     1031      864
cor(Intercept,PitchMode)  0.01      0.29    -0.54     0.54 1.00     1426     1205
cor(Intercept,PitchVar)   0.01      0.30    -0.56     0.58 1.00     2236     1391
cor(PitchMode,PitchVar)   0.02      0.29    -0.55     0.59 1.00     1910     1490
cor(Intercept,SpeechRate) -0.01     0.29    -0.54     0.54 1.00     1799     1527
cor(PitchMode,SpeechRate) 0.00      0.29    -0.55     0.56 1.00     1593     1335
cor(PitchVar,SpeechRate)  -0.02     0.28    -0.55     0.51 1.00     1782     1494
cor(Intercept,ProSpoTime) -0.01     0.28    -0.57     0.51 1.00     1786     1437
cor(PitchMode,ProSpoTime) -0.00     0.28    -0.54     0.53 1.00     1942     1586
```

```
Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    -0.03      0.14    -0.32     0.26 1.00     2019     1608
PitchMode     0.59      0.17     0.25     0.93 1.00     1629     1256
PitchVar     -1.04      0.18    -1.39    -0.70 1.00     1541     1163
SpeechRate   -1.53      0.21    -1.93    -1.14 1.00     1649     1375
ProSpoTime   -2.09      0.24    -2.57    -1.63 1.00     1746     1574
NumPause     -0.23      0.16    -0.55     0.09 1.00     1741     1729
LenPause      3.99      0.30     3.43     4.60 1.00     1432     1637
Noise1       -0.41      0.17    -0.74    -0.08 1.00     1760     1381
Noise2        0.20      0.17    -0.13     0.51 1.00     1924     1628
Noise3        0.04      0.15    -0.27     0.34 1.00     1605     1411
Noise4       -0.43      0.16    -0.75    -0.13 1.00     1786     1414
```

*Varying intercept model*

```
Group-Level Effects:
~ID (Number of levels: 80)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.07      0.05     0.00     0.20 1.00     1672      948

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    -0.04      0.14    -0.31     0.24 1.00     3042     1621
PitchMode     0.65      0.16     0.34     0.96 1.00     1833     1454
PitchVar     -1.08      0.16    -1.42    -0.78 1.00     2179     1555
SpeechRate   -1.62      0.19    -2.00    -1.27 1.00     1769     1610
ProSpoTime   -1.98      0.21    -2.39    -1.59 1.00     2646     1536
NumPause     -0.24      0.15    -0.52     0.04 1.00     3160     1576
LenPause      3.93      0.28     3.39     4.50 1.00     1815     1653
Noise1       -0.45      0.15    -0.75    -0.17 1.00     2337     1560
Noise2        0.23      0.14    -0.05     0.51 1.00     2292     1633
Noise3        0.01      0.14    -0.26     0.29 1.00     4515     1599
Noise4       -0.42      0.14    -0.69    -0.14 1.00     2436     1730
```

*Baseline model*

```
Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    -0.04      0.14    -0.32     0.23 1.00     2634     1271
PitchMode     0.64      0.15     0.35     0.94 1.00     2516     1612
PitchVar     -1.08      0.17    -1.43    -0.76 1.00     2222     1637
SpeechRate   -1.62      0.19    -2.02    -1.26 1.00     1841     1438
ProSpoTime   -1.99      0.21    -2.42    -1.59 1.00     2091     1454
NumPause     -0.24      0.15    -0.56     0.04 1.00     2940     1294
LenPause      3.94      0.29     3.37     4.52 1.00     1964     1417
Noise1       -0.46      0.15    -0.75    -0.17 1.00     2053     1296
Noise2        0.23      0.15    -0.06     0.54 1.00     2153     1358
Noise3        0.01      0.14    -0.26     0.28 1.00     3008     1171
Noise4       -0.42      0.14    -0.70    -0.15 1.00     2398     1604
```

***Sceptic***

*Varying intercepts and slopes model summery output with population-level effects and an excerpt of group-level effects.*

```
Group-Level Effects:
~ID (Number of levels: 80)
                     Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)            0.05      0.04     0.00     0.14 1.00     1908      875
sd(PitchMode)            0.49      0.10     0.29     0.67 1.00      639      602
sd(PitchVar)             0.32      0.12     0.06     0.55 1.01      336      402
sd(SpeechRate)           0.39      0.10     0.18     0.58 1.01      521      447
sd(ProSpoTime)           0.53      0.09     0.36     0.70 1.00     1210     1194
sd(NumPause)             0.58      0.08     0.42     0.75 1.00     1286     1206
sd(LenPause)             0.51      0.10     0.30     0.69 1.02      565      464
sd(Noise1)               0.48      0.10     0.27     0.66 1.00      449      493
sd(Noise2)               0.52      0.09     0.33     0.69 1.00      815      597
sd(Noise3)               0.50      0.08     0.33     0.66 1.00     1216      977
sd(Noise4)               0.51      0.10     0.30     0.70 1.00      875      850
cor(Intercept,PitchMode)    -0.05      0.29    -0.61     0.53 1.01      177      269
cor(Intercept,PitchVar)      0.00      0.29    -0.56     0.56 1.01      552      919
cor(PitchMode,PitchVar)     -0.12      0.24    -0.55     0.37 1.00      760     1022
cor(Intercept,SpeechRate)    0.00      0.29    -0.54     0.53 1.02      251      806
cor(PitchMode,SpeechRate)   -0.09      0.21    -0.49     0.32 1.00      773     1266
cor(PitchVar,SpeechRate)    -0.04      0.25    -0.50     0.45 1.01      534      964
cor(Intercept,ProSpoTime)    0.06      0.28    -0.49     0.57 1.01      216      468
cor(PitchMode,ProSpoTime)   -0.29      0.18    -0.64     0.08 1.00      606      983
```

```
Population-Level Effects:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     -0.02      0.08    -0.18     0.15 1.00     2770     1621
PitchMode     -0.32      0.12    -0.55    -0.09 1.00     1338     1427
PitchVar      -0.07      0.11    -0.30     0.15 1.00     2329     1485
SpeechRate    -0.05      0.11    -0.26     0.16 1.00     1942     1771
ProSpoTime     0.13      0.12    -0.12     0.37 1.00     1979     1531
NumPause       0.25      0.13     0.00     0.49 1.00     1610     1532
LenPause      -0.11      0.12    -0.35     0.14 1.00     1814     1734
Noise1        -0.15      0.12    -0.38     0.08 1.00     1743     1581
Noise2        -0.07      0.12    -0.31     0.16 1.00     1831     1781
Noise3         0.19      0.12    -0.04     0.42 1.00     1664     1674
Noise4        -0.04      0.12    -0.28     0.19 1.00     2102     1696
```

*Varying intercept model*

```
Group-Level Effects:
~ID (Number of levels: 80)
             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.04      0.03     0.00     0.11 1.00     1866     1174

Population-Level Effects:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     -0.00      0.05    -0.10     0.10 1.00     6476     1288
PitchMode     -0.27      0.05    -0.38    -0.17 1.00     6602     1510
PitchVar      -0.14      0.05    -0.24    -0.04 1.01     6520     1516
SpeechRate     0.00      0.05    -0.11     0.11 1.00     4729     1413
ProSpoTime     0.11      0.05     0.00     0.22 1.00     5004     1436
NumPause       0.21      0.05     0.11     0.31 1.00     5822     1653
LenPause      -0.09      0.05    -0.20     0.02 1.00     6327     1285
Noise1        -0.00      0.05    -0.10     0.10 1.00     3770     1435
Noise2        -0.02      0.05    -0.12     0.09 1.00     5401     1193
Noise3         0.04      0.05    -0.07     0.14 1.00     5512     1576
Noise4        -0.09      0.05    -0.19     0.01 1.00     5400     1582
```

*Baseline*

```
Population-Level Effects:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept     -0.00      0.05    -0.11     0.10 1.00     1787     1379
PitchMode     -0.28      0.05    -0.37    -0.17 1.00     2367     1395
PitchVar      -0.14      0.05    -0.24    -0.03 1.00     2156     1514
SpeechRate     0.00      0.05    -0.10     0.10 1.00     2098     1514
ProSpoTime     0.11      0.05     0.01     0.21 1.00     2389     1518
NumPause       0.21      0.05     0.11     0.32 1.00     2185     1147
LenPause      -0.09      0.05    -0.19     0.01 1.00     2012     1157
Noise1        -0.00      0.05    -0.10     0.10 1.00     2009     1193
Noise2        -0.01      0.05    -0.12     0.09 1.00     1973     1325
Noise3         0.03      0.05    -0.07     0.14 1.00     1882     1627
Noise4        -0.09      0.05    -0.18     0.01 1.00     2307     1684
```

All models seem to be estimating quite well. Most convergence diagnostic (Rhat) are about 1, and all effective sample size (ESS) values are high relative to our iterations and number of chains.

Next step in our workflow is to assess the accuracy of our model's prediction both using confusion matrices and uncertainty plots. We do this across sceptic/informed dataset and the three models both for test and train parts:

| ***Informed*** | | | | | | |
|---|---|---|---|---|---|---|
| ***Train*** | Varying intercepts/slopes | | Varying intercept | | Baseline | |
| Prediction\Truth | Control | Schizophrenic | Control | Schizophrenic | Control | Schizophrenic |
| Control | 780 | 19 | 771 | 29 | 771 | 29 |
| Schizophrenic | 20 | 781 | 29 | 771 | 29 | 771 |
| ***Test*** | Varying intercepts/slopes | | Varying intercept | | Baseline | |
| Prediction\Truth | Control | Schizophrenic | Control | Schizophrenic | Control | Schizophrenic |
| Control | 190 | 7 | 189 | 8 | 189 | 8 |
| Schizophrenic | 10 | 193 | 11 | 192 | 11 | 192 |

| ***Sceptic*** | | | | | | |
|---|---|---|---|---|---|---|
| ***Train*** | Varying intercepts/slopes | | Varying intercept | | Baseline | |
| Prediction\Truth | Control | Schizophrenic | Control | Schizophrenic | Control | Schizophrenic |
| Control | 795 | 5 | 475 | 343 | 481 | 336 |
| Schizophrenic | 5 | 795 | 325 | 457 | 319 | 464 |
| ***Test*** | Varying intercepts/slopes | | Varying intercept | | Baseline | |
| Prediction\Truth | Control | Schizophrenic | Control | Schizophrenic | Control | Schizophrenic |
| Control | 128 | 72 | 116 | 71 | 119 | 69 |
| Schizophrenic | 72 | 128 | 84 | 129 | 81 | 131 |

From the matrices we observe a difference in performance for the training datasets, where the varying intercepts and slopes model perform much better, and the two other models are relatively close to each other's performance level. As for the test data set all three models seem to be performing at

somewhat the same level. Lastly, comparing between the sceptic and informed datasets, we see a general trend of the informed model is better at prediction as expected. We plot our model's prediction uncertainty to better grasp which model performs better.

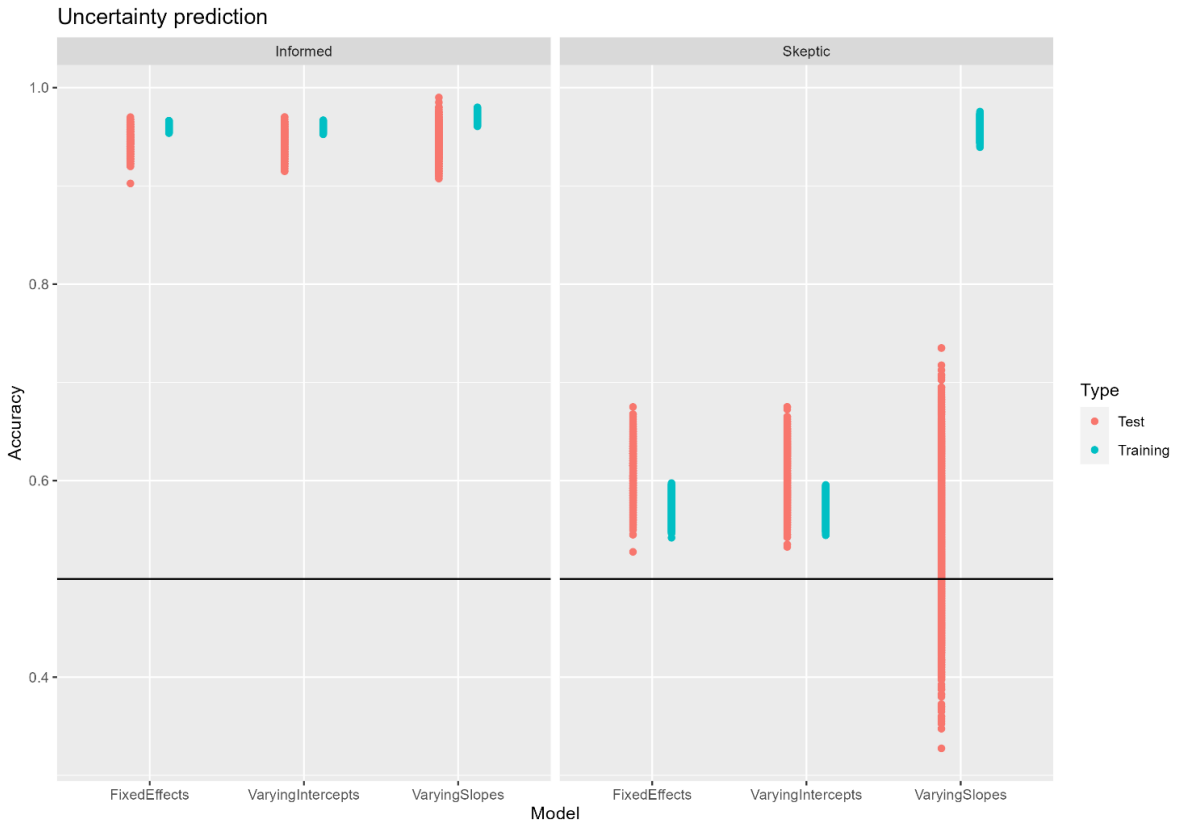



*Figure 28 - Prediction with uncertainty. FixedEffects is the label for baseline model. The black line resembles chance level performance.*

Here the models are in the order of baseline, varying intercepts, and varying slopes and intercepts. The plot paints a rather clear picture of the varying slopes model over fitting likewise all the informed models. As expected, the sceptic models show more uncertainty and general lower accuracy of prediction. We may for future reference use a sceptic set of priors and baseline, as this was our second-best model after the overfitting varying slopes and intercept model.

## 3.3 Applying the pipeline to empirical data (SK, MO)

Looking at the empirical data from Parola et al. (2020) we have 1889 observations over 398 variables. In total 122 participants with an equal number of females and males in the two diagnostic groups:

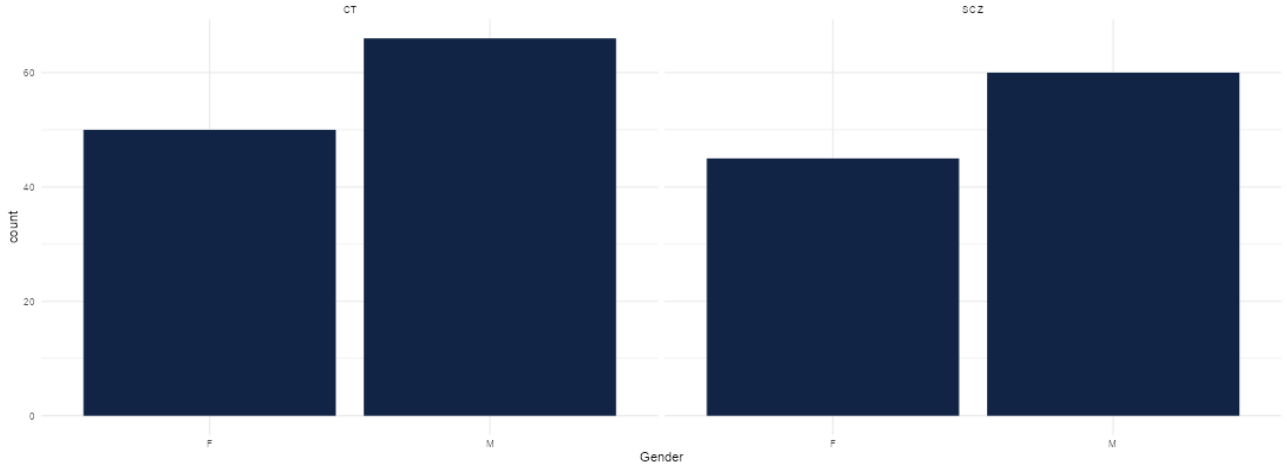



*Figure 29 - Gender distribution in diagnostic groups controls (CT) and schizophrenic (SCZ) in empirical data*

Next plot displays a glimpse of diagnostic differences in duration of speech, number of pauses, percentage of spoken time and gender differences in mean pitch. For a more in-depth description of measurements see Parola et al. (2020).

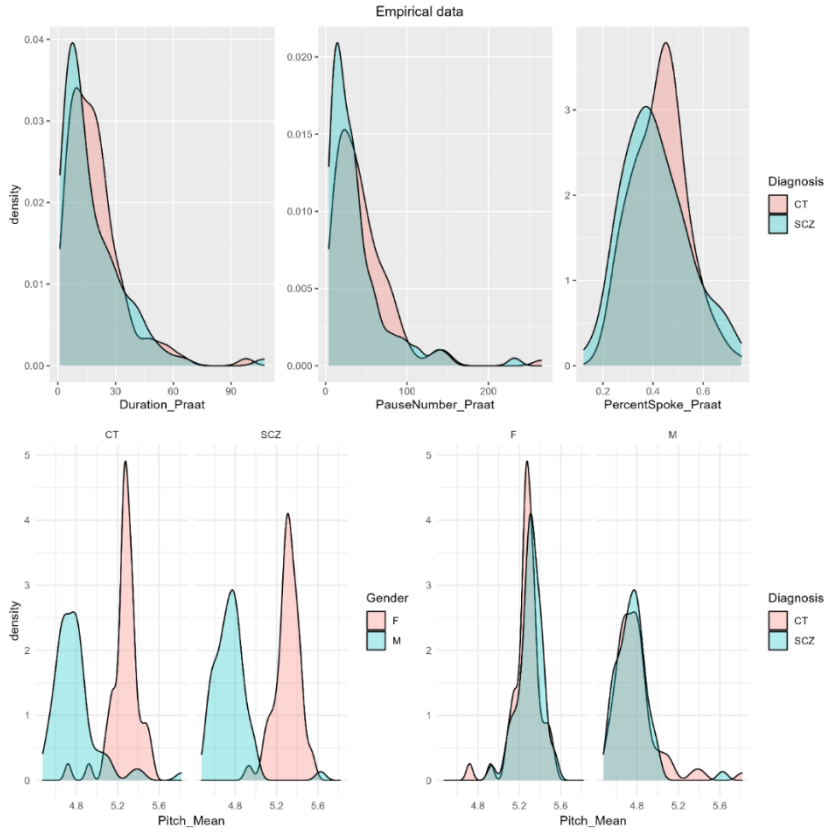



*Figure 30 - A excerpt of diagnostic and gender differences in empirical data*

Running a model on many highly correlated measures, can be done in a better way. Therefore, we pre-process the data by standardizing all measurements and run a parallel analysis based on a principal component analysis (PCA):

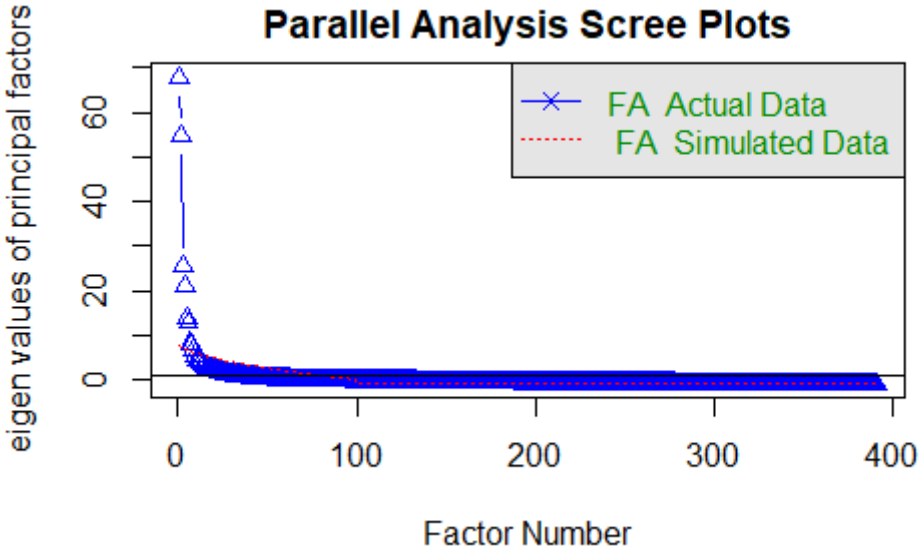



*Figure 31 - Scree plot from parallel analysis*

The PCA suggest using nine factors for further investigations. The loadings of these factors can be seen in the appendix. The nine factors', MR1 through MR9, proportion and cumulative variance with a grand total of only 52% of the variance is explained below:

| | MR2 | MR5 | MR9 | MR6 | MR3 | MR4 | MR1 | MR7 | MR8 |
|---|---|---|---|---|---|---|---|---|---|
| SS loadings | 37.742 | 36.525 | 23.285 | 21.205 | 23.523 | 19.933 | 16.610 | 12.615 | 10.419 |
| Proportion Var | 0.097 | 0.093 | 0.060 | 0.054 | 0.060 | 0.051 | 0.042 | 0.032 | 0.027 |
| Cumulative Var | 0.097 | 0.190 | 0.249 | 0.304 | 0.364 | 0.415 | 0.457 | 0.490 | 0.516 |

For further data pre-processing, we again data budget by splitting the data into a training and test set with a ratio of 80/20 while keeping an equal distribution of gender and diagnosis in both sets.

Next, we model the data using the baseline model from the previous section:

$$Diagnosis \sim 1 + MR1 + MR2 + MR3 + MR4 + MR5 + MR6 + MR7 + MR8 + MR9$$

After fitting the model to the priors and data, we check the prior posterior distribution plots:

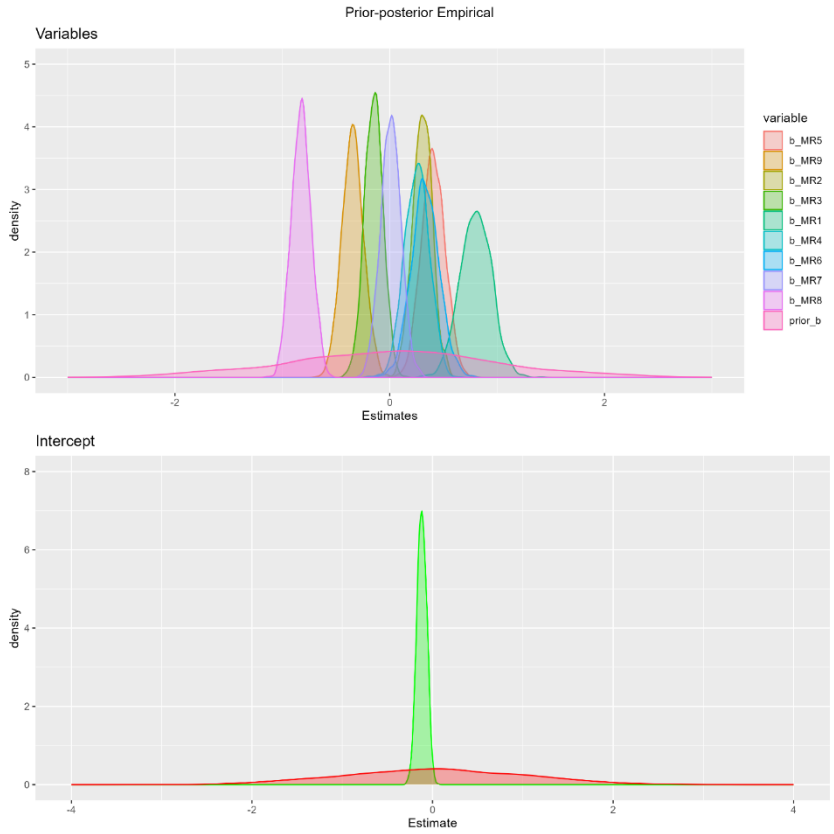



*Figure 32 - Prior posterior distributions of empirical training set*

Looking at the models output we see the model has a both nice Rhat, bulk and tail values. Suggesting this was a reliable sampling.

```
Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    -0.12      0.05    -0.22    -0.02 1.00     2145     1486
MR2           0.30      0.09     0.12     0.49 1.00     1465     1425
MR5           0.41      0.11     0.20     0.62 1.00     1239     1146
MR9          -0.34      0.10    -0.54    -0.14 1.00     1411     1161
MR6           0.33      0.13     0.08     0.60 1.00     1294     1086
MR3          -0.15      0.09    -0.32     0.01 1.00     1362     1377
MR4           0.25      0.11     0.03     0.46 1.00     1229     1393
MR1           0.80      0.14     0.51     1.08 1.00     1197     1028
MR7           0.01      0.09    -0.17     0.19 1.00     1490     1496
MR8          -0.82      0.09    -1.00    -0.65 1.00     1370     1208
```

Next, we look at predicting performance both average and with uncertainty.

| ***Empirical*** | | | | |
|---|---|---|---|---|
| Baseline | | | | |
| | ***Train*** | | ***Test*** | |
| Prediction\Truth | Control | Schizophrenic | Control (68) | Schizophrenic (62) |

| | | | | |
|---|---|---|---|---|
| Control | 547 | 288 | 133 (correct rejection) | 69 (false positive) |
| Schizophrenic | 244 | 432 | 65 (false negative) | 111 (true positive) |

As displaying in the above matrix, the model predicts the majority of the participants diagnostic group correctly. However, in the case of diagnosing people with schizophrenic symptoms different approaches should be considered. Would one rather diagnose too many as false positives and have more people in health system taking the space from those who need the help the most. Or would one rather take a conservative approach and diagnose a majority false negatives. Calculating the percentage of true cases compared to the prediction of control and schizophrenic diagnosed people; we get that false positives are predicted 32% of the time compared to false negatives 38% of the time. Giving an indication that our model is on the tipping point of a conservative diagnostic approach. We look further into our model's uncertainty:

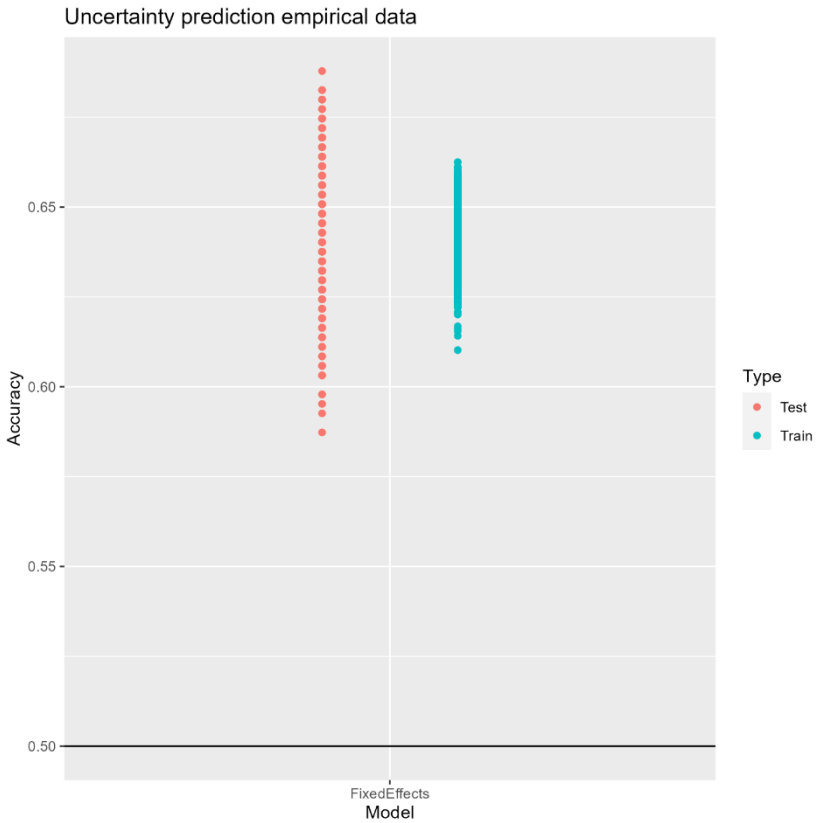



*Figure 33 - Model prediction with black resembling chance level*

The uncertainty is quite high, though not overfittingly high. The uncertainty of the model's predictions stays nicely over chance level for both test and train sets. Giving the circumstances of sample

size, pre-processing and publication bias from the meta-analysis, this approximation is in accordance with what is expected of what a model would look like to diagnose schizophrenia from vocal markers.

Going forward we can recommend getting a larger sample size with measures like those from the meta-analysis, since it would be easier to cumulatively assess vocal markers in the control and schizophrenic groups. Another suggestion is to do a feature importance analysis, to see which variables can explain the most variance in the dataset. Following, we remove highly correlated variables, as minimizing the number of variables need to diagnose a new patient would highly benefit the diagnostic process. Lastly, repeating the workflow for different algorithms (herein logistic regression) and addition iterations. Time is a precious resource and prioritizing it is a crucial decision in diagnostics and machine learning.

# 4 References

## 4.1 Literature

Parola, A., Arndis, S., Vibeke, B., & Riccardo, F. (2020). Voice patterns in schizophrenia: A systematic review and Bayesian meta-analysis—ScienceDirect. *Schizophrenia Research*, *216*, 24–40. https://doi.org/doi.org/10.1016/j.schres.2019.11.031

Fusaroli, R., Weed, E., Fein, D., & Naigles, L. (2019). Hearing me hearing you: Reciprocal effects between child and parent language in autism and typical development. *Cognition*, *183*, 1–18. https://doi.org/10.1016/j.cognition.2018.10.022

## 4.2 Materials

RStudio Team (2022). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Bürkner P (2021). "Bayesian Item Response Modeling in R with brms and Stan." *Journal of Statistical Software*, **100**(5), 1–54. doi:10.18637/jss.v100.i05.

Paul-Christian Bürkner (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. Journal of Statistical Software, 80(1), 1-28. doi:10.18637/jss.v080.i01

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." _Journal of Open Source Software_, *4*(43), 1686. doi:10.21105/joss.01686

Kuhn M, Wickham H (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*. https://www.tidymodels.org.

Gabry J, Češnovar R (2022). _cmdstanr: R Interface to 'CmdStan'_. https://mc-stan.org/cmdstanr/, https://discourse.mc-stan.org.

# 5 Appendix

***Factor analysis loadings of empirical data after pre-processing for portfolio 3***

Loadings:

| | MR2 | MR5 | MR9 | MR6 | MR3 | MR4 | MR1 | MR7 | MR8 |
|---|---|---|---|---|---|---|---|---|---|
| Duration_Praat | | | 0.127 | | | 0.102 | | 0.889 | |
| F0_Mean_Praat | 0.114 | -0.115 | 0.188 | 0.228 | 0.121 | | 0.224 | -0.126 | 0.422 |
| F0_SD_Praat | | | 0.143 | -0.341 | 0.171 | | | | 0.451 |
| Intensity_Mean_Praat | 0.134 | 0.251 | | 0.172 | | -0.118 | 0.127 | -0.280 | -0.180 |
| Intensity_SD_Praat | | | -0.103 | 0.153 | | | 0.111 | -0.156 | -0.214 |
| PauseDuration_Praat | | | 0.175 | | | 0.131 | | 0.799 | |
| TurnDuration_Praat | 0.102 | | | 0.265 | | | | 0.861 | |
| TurnNumber_Praat | | | | | | | | 0.869 | 0.109 |
| PauseNumber_Praat | | | | | | | | 0.868 | 0.110 |
| PercentSpoke_Praat | 0.191 | 0.125 | -0.238 | 0.641 | | -0.132 | | | |
| PercentSilence_Praat | -0.191 | -0.125 | 0.238 | -0.641 | | 0.132 | | | |
| NHR_mean | | | | | | | | 0.114 | |
| NHR_std | | | | | | | | 0.112 | |
| Duration_Cova | | | 0.127 | | | 0.102 | | 0.889 | |
| PauseDuration_Cova | | | 0.182 | -0.132 | | 0.130 | | 0.780 | |
| TurnDuration_Cova | | | | 0.372 | | | | 0.870 | |
| TurnNumber_Cova | | | | | | | | 0.888 | |
| PauseNumber_Cova | | | | | | | | 0.888 | |
| PercentSpoke_Cova | | | -0.318 | 0.798 | | -0.131 | | 0.203 | |
| PercentSilence_Cova | | | 0.318 | -0.798 | | 0.131 | | -0.203 | |
| Pitch_Mean | 0.127 | -0.211 | 0.280 | 0.507 | | 0.100 | 0.213 | -0.138 | 0.259 |
| Pitch_Median | 0.127 | -0.206 | 0.316 | 0.509 | | | 0.200 | -0.146 | 0.233 |
| Pitch_SD | 0.224 | | 0.656 | -0.507 | | | | | 0.331 |
| Pitch_IQR | 0.199 | 0.125 | 0.302 | -0.388 | | | | | 0.414 |
| Pitch_MAD | 0.202 | 0.123 | 0.279 | -0.334 | | | | 0.161 | 0.347 |
| F0_Mean | 0.141 | -0.207 | 0.295 | 0.446 | | | 0.253 | -0.131 | 0.269 |
| F0_Median | 0.136 | -0.204 | 0.316 | 0.465 | | | 0.244 | -0.141 | 0.230 |
| F0_SD | 0.223 | | 0.584 | -0.140 | | 0.124 | 0.158 | | 0.421 |
| F0_IQR | 0.205 | | 0.402 | -0.168 | | | 0.102 | | 0.484 |
| F0_MAD | 0.205 | | 0.416 | -0.106 | | | 0.116 | | 0.429 |
| F1_Mean | 0.764 | | | | | | | | 0.351 |
| F1_Median | 0.384 | | | | | | | | 0.477 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| F1_SD | 0.917 | | | | | | | | |
| F1_IQR | 0.718 | | | | | | | | 0.249 |
| F1_MAD | 0.706 | | | | | | | | 0.288 |
| F2_Mean | 0.927 | | | | | | | | |
| F2_Median | 0.896 | | | | | | | | 0.138 |
| F2_SD | 0.911 | | | | | | | | |
| F2_IQR | 0.813 | | | | | | 0.126 | | |
| F2_MAD | 0.797 | | | | | | 0.115 | | |
| F3_Mean | 0.938 | | | | | | | | |
| F3_Median | 0.926 | | | | | | | | |
| F3_SD | 0.904 | | | | | | | | |
| F3_IQR | 0.863 | -0.110 | | | | | | 0.114 | |
| F3_MAD | 0.850 | -0.108 | | | | | | 0.114 | |
| F4_Mean | 0.940 | | | | | | | | |
| F4_Median | 0.938 | | | | | | | | |
| F4_SD | 0.890 | | | | | | | 0.101 | |
| F4_IQR | 0.871 | | | | | | 0.111 | | |
| F4_MAD | 0.855 | | | | | | 0.103 | | |
| F5_Mean | 0.938 | | | | | | | | |
| F5_Median | 0.936 | | | | | | | | |
| F5_SD | 0.876 | -0.109 | | | | | | | |
| F5_IQR | 0.803 | | | -0.100 | | | 0.117 | 0.101 | |
| F5_MAD | 0.779 | -0.102 | | | | | 0.109 | 0.117 | |
| NAQ_Mean | 0.220 | -0.106 | 0.304 | 0.347 | 0.402 | | | -0.228 | |
| NAQ_Median | 0.225 | -0.105 | 0.312 | 0.352 | 0.374 | | | -0.232 | |
| NAQ_SD | 0.124 | | 0.261 | 0.270 | 0.485 | | 0.108 | -0.191 | |
| NAQ_IQR | 0.114 | | 0.260 | 0.314 | 0.458 | | | -0.208 | |
| NAQ_MAD | 0.118 | | 0.273 | 0.326 | 0.452 | | | -0.206 | |
| QOQ_Mean | 0.182 | | 0.280 | 0.340 | 0.545 | | | -0.190 | |
| QOQ_Median | 0.183 | | 0.302 | 0.364 | 0.503 | | | -0.203 | |
| QOQ_SD | | | 0.110 | | 0.349 | | | | |
| QOQ_IQR | | | 0.129 | 0.193 | 0.644 | 0.133 | | | 0.115 |
| QOQ_MAD | | | 0.145 | 0.207 | 0.627 | 0.128 | | | 0.120 |
| H1H2_Mean | | -0.206 | 0.368 | 0.493 | | 0.186 | | -0.245 | |
| H1H2_Median | | -0.211 | 0.361 | 0.485 | | 0.193 | | -0.244 | |
| H1H2_SD | | 0.123 | -0.218 | -0.412 | | | | 0.242 | 0.281 |
| H1H2_IQR | | 0.122 | -0.176 | -0.311 | | | | 0.227 | 0.245 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| H1H2_MAD | | 0.118 | -0.169 | -0.317 | | | | 0.220 | 0.246 |
| PSP_Mean | -0.614 | | -0.155 | -0.188 | -0.365 | | | 0.167 | |
| PSP_Median | -0.588 | | -0.154 | -0.202 | -0.306 | | -0.131 | 0.163 | |
| PSP_SD | -0.481 | | -0.162 | -0.180 | -0.452 | | | 0.161 | |
| PSP_IQR | -0.536 | | -0.167 | -0.132 | -0.412 | | | 0.170 | |
| PSP_MAD | -0.535 | | -0.181 | -0.163 | -0.391 | | | 0.178 | |
| HRF_Mean | -0.151 | 0.213 | -0.297 | -0.521 | | -0.141 | -0.127 | 0.162 | -0.250 |
| HRF_Median | -0.150 | 0.218 | -0.299 | -0.495 | | -0.158 | -0.129 | 0.168 | -0.263 |
| HRF_SD | -0.128 | 0.150 | -0.226 | -0.555 | | | | 0.206 | -0.104 |
| HRF_IQR | | 0.172 | -0.260 | -0.640 | | -0.106 | | 0.245 | |
| HRF_MAD | | 0.175 | -0.262 | -0.632 | | -0.112 | | 0.251 | |
| MDQ_Mean | -0.465 | | 0.167 | 0.236 | -0.115 | -0.110 | 0.200 | -0.129 | 0.212 |
| MDQ_Median | -0.473 | | 0.172 | 0.239 | | | 0.198 | -0.136 | 0.195 |
| MDQ_SD | | | -0.191 | -0.365 | 0.327 | 0.203 | | 0.192 | |
| MDQ_IQR | 0.119 | | -0.166 | -0.318 | 0.404 | 0.235 | | 0.201 | |
| MDQ_MAD | 0.122 | | -0.157 | -0.327 | 0.407 | 0.245 | | 0.183 | |
| peakSlope_Mean | -0.760 | 0.120 | | | | | 0.206 | | 0.340 |
| peakSlope_Median | -0.763 | 0.140 | | | | -0.113 | 0.204 | | 0.306 |
| peakSlope_SD | -0.455 | | | | | | | | 0.355 |
| peakSlope_IQR | -0.682 | | | | | | 0.103 | | 0.322 |
| peakSlope_MAD | -0.726 | | | | | | 0.118 | | 0.318 |
| Rd_Mean | | | | | | -0.675 | | | 0.150 |
| Rd_Median | | | | | | -0.636 | | | 0.167 |
| Rd_SD | | 0.118 | | -0.153 | | -0.507 | | 0.266 | -0.155 |
| Rd_IQR | | 0.113 | | -0.141 | | -0.494 | | 0.263 | -0.162 |
| Rd_MAD | | | | -0.128 | | -0.524 | | 0.234 | -0.129 |
| Rd_conf_Mean | | | | | 0.803 | 0.279 | | | |
| Rd_conf_Median | | | | | 0.780 | 0.286 | | | |
| Rd_conf_SD | -0.107 | 0.165 | | -0.102 | 0.381 | -0.217 | | 0.261 | -0.353 |
| Rd_conf_IQR | -0.110 | 0.155 | | -0.120 | 0.363 | -0.217 | | 0.262 | -0.354 |
| Rd_conf_MAD | -0.113 | 0.156 | | -0.127 | 0.353 | -0.219 | | 0.246 | -0.354 |
| VAD_Mean | 0.121 | 0.194 | -0.207 | 0.731 | | -0.122 | 0.137 | | |
| MCEP0_Mean | -0.635 | | | 0.184 | | -0.100 | | -0.244 | |
| MCEP0_Median | -0.642 | 0.104 | | 0.165 | | -0.123 | | -0.191 | |
| MCEP0_SD | -0.474 | | | | | | 0.229 | 0.222 | |
| MCEP0_IQR | -0.356 | | | | | | 0.177 | | |
| MCEP0_MAD | -0.352 | | | | | 0.109 | 0.167 | | |
| MCEP1_Mean | 0.837 | | | 0.103 | | | | | -0.148 |
| MCEP1_Median | 0.821 | | | 0.106 | | | | 0.112 | -0.142 |
| MCEP1_SD | -0.418 | | | | | | | | |
| MCEP1_IQR | -0.342 | | | | | | | | 0.112 |