Victoria Engberg Lowe, Vlada Caraman &
Natasha Becker Bertelsen

BSc in Cognitive Science
Methods 3 E2022
12/10-2022

Methods 3 Assignment 3

Part 1: Simulation

1.1 The goal of the simulation

The goal of the simulation is to produce a skeptical and an informed data set with 2000 observations in each containing information about 100 matched pairs of controls and schizophrenia. The simulation emulates that each participant goes through 10 trials while their speech is recorded. The informed dataset produced by the simulation includes 6 acoustic measures from the meta-analysis and 4 variables representing noise. The skeptical one includes only noise and therefore functions as a baseline data set. We have an informed and a skeptic dataset so that we can investigate if we can produce a model which can pick up on signal rather than random noise.

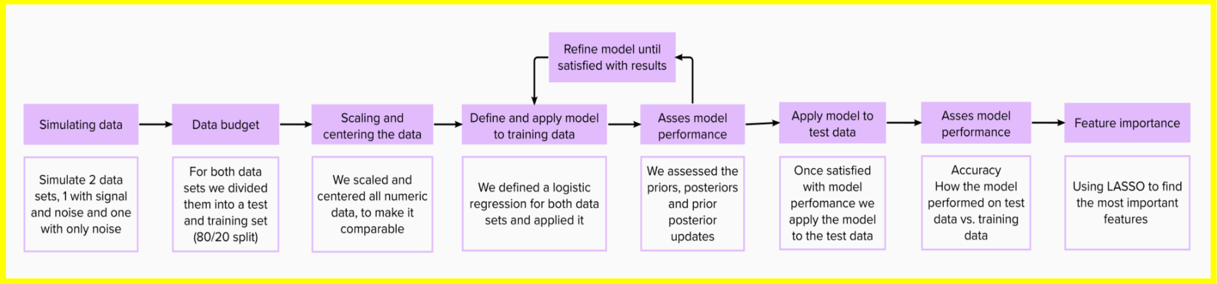
1.2 The simulation process

To begin with, we set a seed at 1000. Then, we define the parameters for the informed setup as the following:

Pitch	0.25
Pitch variability	-0.55
Speech rate	-0.75
Proportion of spoken time	-1.26
Pause number	0.05
Pause length	1.89
Noise 1	0
Noise 2	0
Noise 3	0
Noise 4	0

For the simulation of these two datasets, we defined the individual variability (Individual_SD = 1), the variability across trials (Trial_SD = 0.5) and the measurement error (Error= 0.2). After this, we identified the true effect sizes for each variable and created both the informed and skeptic datasets.

Part 2: ML pipeline on simulated data


Refine model until satisfied with results
Simulating data
Data budget
Scaling and centering the data
Define and apply model to training data
Asses model performance
Apply model to test data
Asses model performance
Feature importance
Simulate 2 data sets, 1 with signal and noise and one with only noise
For both data sets we divided them into a test and training set (80/20 split)
We scaled and centered all numeric data, to make it comparable
We defined a logistic regression for both data sets and applied it
We assessed the priors, posteriors and prior posterior updates
Once satisfied with model perfomance we apply the model to the test data
Accuracy How the model performed on test data vs. training data
Using LASSO to find the most important features

2.1 Data budgeting

The next step after simulating the datasets was data budgeting. We created a train and a test set for each dataset using an 80/20 split. We are doing this because we want to try our model on the train sets first and only after we are satisfied with the model, will we apply it to the test sets. Using tidymodels, we created a recipe that helped us with scaling and centering all numeric data. At this point, we had four sets: two for the informed data: *train_informed_s* and *test_informed_s* and two for the skeptic data: *train_skeptic_s* and *test_skeptic_s.*

2.2 Defining models and priors

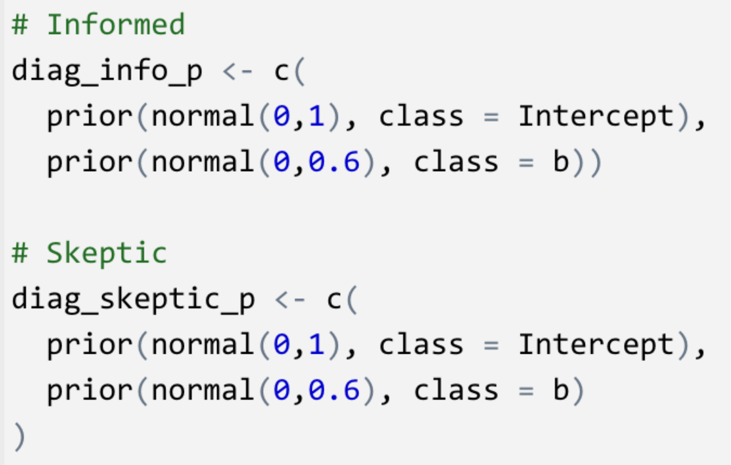
Following a Bayesian workflow, we defined the following two models for the informed and skeptical setup separately. Both models have the binary outcome "Group" (Schizophrenia or Control) and they are both predicted by the intercept as well as additional variables. Apart from the intercept, the informed model includes the 6 informed variables as well as the 4 random noise variables as predictors. Also, they include random slopes for trial and random intercepts for ID. This means that each ID will have an individual slope and intercept.

```
diag_info_f <- bf(Group ~ 1 + pitch + pitch_variability + speech_rate +
proportion_of_spoken_time + pause_number + pause_length + noise_1 +
noise_2  + noise_3 + noise_4 + (1+Trial|ID))
```

In the skeptic model, the group is predicted by 10 random noise variables, and there are also random slopes and intercepts in this model.

```
diag_skeptic_f <- bf(Group ~ 1 + noise_1 + noise_2 + noise_3 + noise_3 +
noise_4 + noise_5 + noise_6 + noise_7  + noise_8 + noise_9 + noise_10 +
(1+Trial|ID))
```

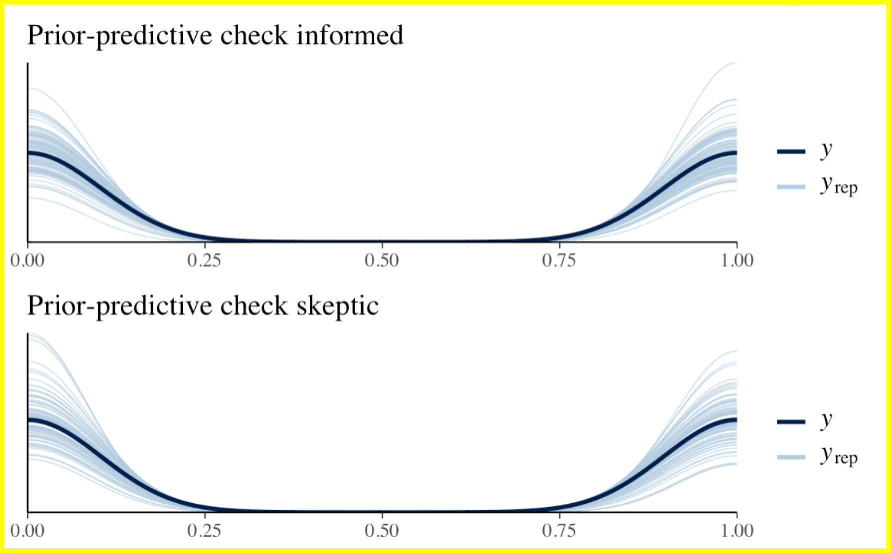
We specify the following priors for the two models. When choosing priors, we take into account that we are working with Bernoulli data, and therefore the priors are on log-odds scale.



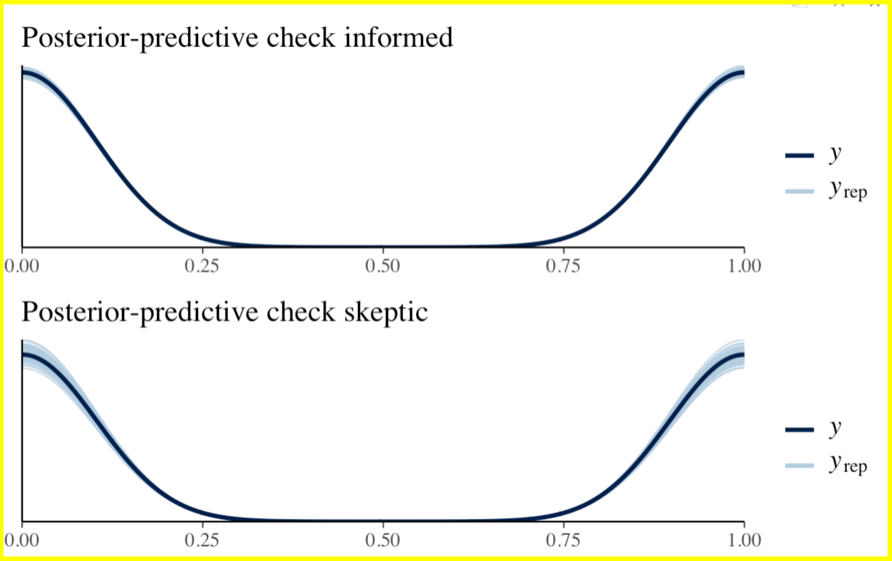
```
# Informed
diag_info_p <- c(
  prior(normal(0,1), class = Intercept),
  prior(normal(0,0.6), class = b))

# Skeptic
diag_skeptic_p <- c(
  prior(normal(0,1), class = Intercept),
  prior(normal(0,0.6), class = b)
)
```

2.3 Predictive checks

We do prior predictive checks and get the following plots.

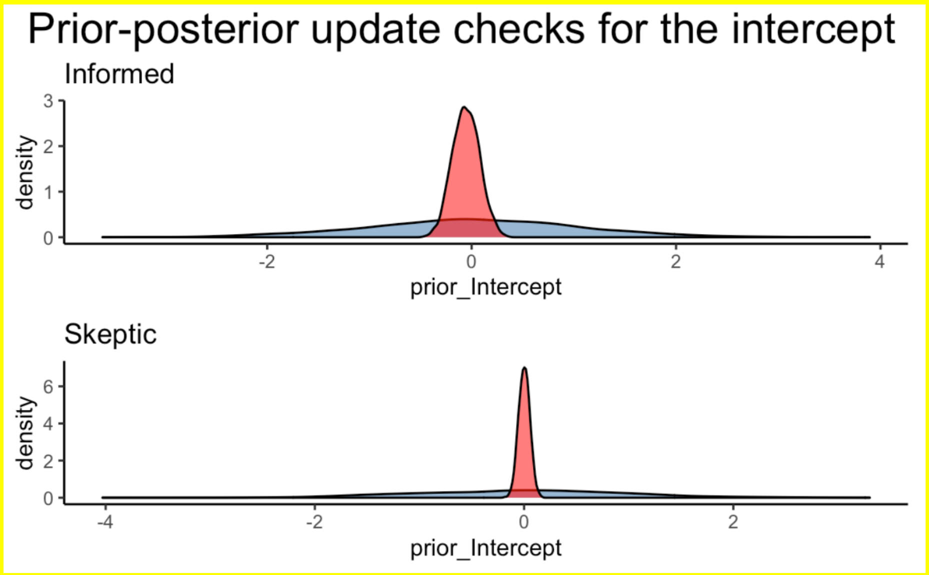


The two plots present the bimodal distributions. We can see that we have a symmetric distribution of both the priors and the data, which means that we have an even number of participants with schizophrenia and control participants.

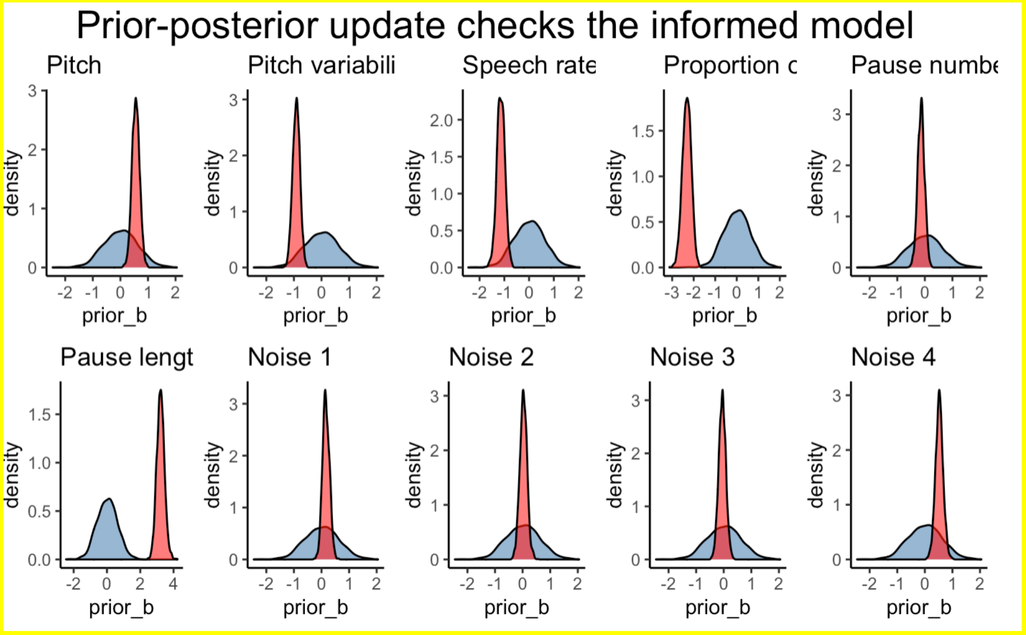


The posterior-predictive checks also show symmetric distributions of the two groups, and we can see that the posterior distributions fit the simulated data much better now. The posterior-predictive check for the skeptic model has a bit more noise than for the informed.

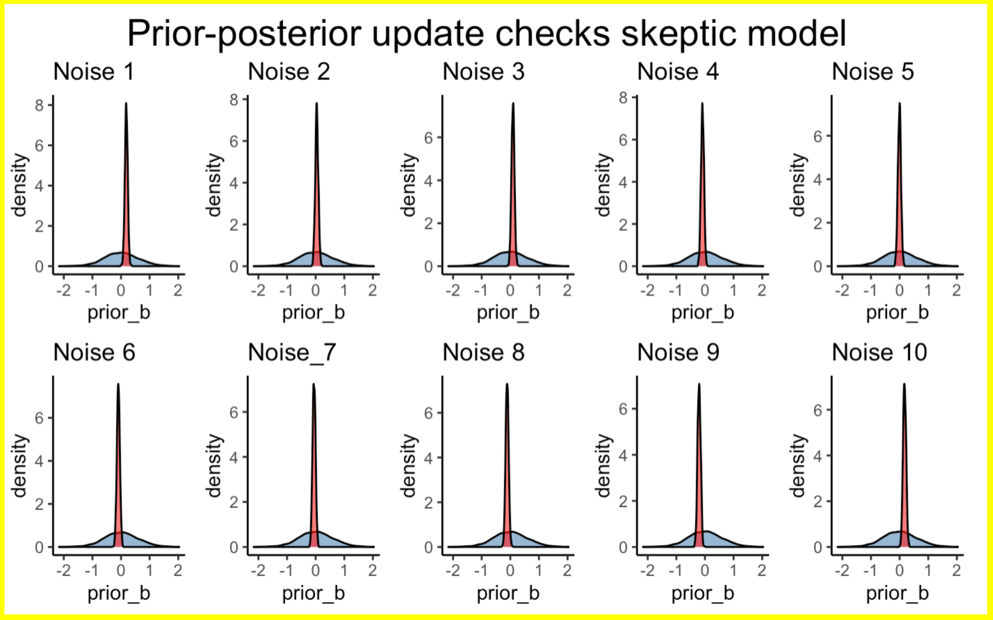
The prior-posterior update checks for the intercept shows that we have relatively uninformed priors for both setups. We observe that the posterior for the informed is wider than the one for the skeptic, indicating that it is taking more information into account, as well as being more uncertain. This is also supported by the posterior for the informed being less pointy than the one for the skeptic. The skeptic posterior is very slim and pointy, indicating that it is more confident, but suspiciously so, given that we know there is no actual signal in the data used to generate the prior and posterior.



Overall the prior-posterior update checks for the informed show that most posteriors are within the range of the priors, indicating that the priors mostly encapsulates the data. There are some posteriors that are pushing on their priors (especially for the variables *pause length* and *proportion of spoken time*). We were not able to completely rectify this, we went back and adjusted the priors in an attempt to better capture the data, but in the end it was not sufficient for all variables. Generally the posteriors are quite pointy (i.e., confident), and we see that many of them have learnt a lot from the priors.



The prior-posterior update checks for the skeptic model show similar patterns for all predictors, namely that all posteriors are within the range of the priors and are very confident. None of the posteriors are skewed.

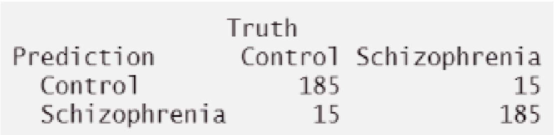


When comparing the update checks for the informed and skeptic model, we see that the informed plots show more variation than the skeptic plot. This indicates that the informed model is picking up on some signal, this is supported by the fact that all noise variables (including those in the informed model) are centered around 0, where the 6 values from the meta analysis are not.

2.4 Assessing performance on the test set

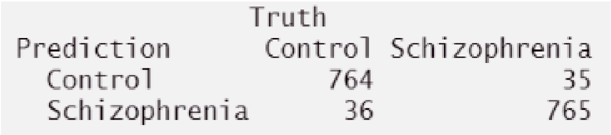
We used confusion matrices to assess performance on our test sets, we made a matrix for both the train and test datasets for comparison. This gives us the following matrices for the informed model:

from the test dataset



```
              Truth
Prediction     Control Schizophrenia
  Control          185            15
  Schizophrenia     15           185
```

from the training dataset

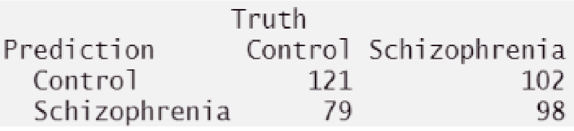


```
               Truth
Prediction       Control Schizophrenia
  Control            764            35
  Schizophrenia       36           765
```

This matrix demonstrates that the model is quite good at correctly identifying participants, with only 30 out of 400 participants in the test dataset being incorrectly classified as either schizophrenia or control. This result further indicates that our informed model was able to pick up on reliable signals and use those to correctly classify the participants.

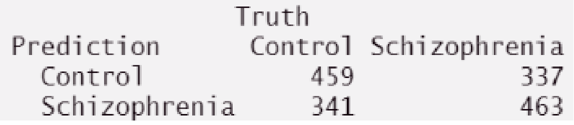
We get the following matrices for the skeptic model:

from the test dataset



```
              Truth
Prediction     Control Schizophrenia
  Control          121           102
  Schizophrenia     79            98
```

from the training dataset



```
              Truth
Prediction     Control Schizophrenia
  Control          459           337
  Schizophrenia    341           463
```

For the skeptic model, we see that the model is not very good at identifying participants, it does it at around chance level. This indicates that the skeptic model has not been able to detect reliable signals to classify based on.

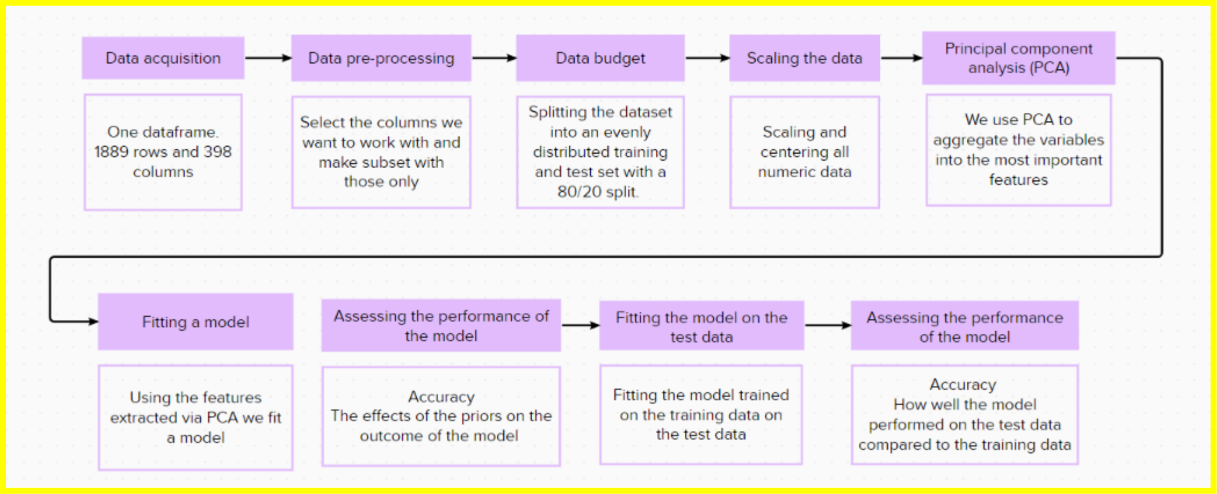
2.5 Feature importance

To determine which features are the most important and thereby the ones to keep in the model, we use the lasso procedure. Firstly, we run the logistic regression and take each of the beta coefficients and make them absolute. We sum the coefficients and find the weight of the features by dividing each feature with the sum. A threshold of 0.10 is chosen and therefore all the values that are below the threshold are removed. We run the lasso procedure again until we have no values left below the threshold. In the end, we have the following four features that are the most important: *pitch variability* (weight 0.12), *speech rate* (weight 0.15), *proportion of spoken time* (weight 0.31) and *pause length* (weight 0.42).

This is as we expected, we see that the 4 noise variables have been discarded, as well as 2 of the 6 acoustic variables. We are left with 4 of the 6 acoustic variables, and they are all congruent with the current literature on the effect of schizophrenia on voice patterns, e.g., that schizophrenia patients have longer speech pauses than healthy controls.

Part 3: Applying the ML pipeline to empirical data

3.1 The machine learning pipeline



Data acquisition

One dataframe. 1889 rows and 398 columns

Data pre-processing

Select the columns we want to work with and make subset with those only

Data budget

Splitting the dataset into an evenly distributed training and test set with a 80/20 split.

Scaling the data

Scaling and centering all numeric data

Principal component analysis (PCA)

We use PCA to aggregate the variables into the most important features

Fitting a model

Using the features extracted via PCA we fit a model

Assessing the performance of the model

Accuracy
The effects of the priors on the outcome of the model

Fitting the model on the test data

Fitting the model trained on the training data on the test data

Assessing the performance of the model

Accuracy
How well the model performed on the test data compared to the training data