Portfolio 3 text

Part 1

Simulating data

Using the meta-analysis of Parola et al. (2020), we create a simulated dataset of 100 matched pairs of schizophrenic and healthy participants where each participant produced repeated measures across 10 trials. For each recording 10 acoustic features are produced. Specifically, six of these features are based on the meta-analysis and the remaining four are random noise.

Parameter	Name in code	Value(s)
Mean effect sizes of the informed dataset	EffectsMean_inform	0.25,-0.55,-0.75,-1.26,0.05,1.89,0,0,0,0
Standard deviation of the effect sizes of the informed dataset	EffectsSD_inform	0.5,0.29,0.39,0.63,0.59,0.62,1,1,1,1
Measurement error	Error	0.2
Standard deviation of the trials	TrialSD	0.5

Similarly, a baseline dataset, sim_d_noise, is simulated. However, this dataset only includes 10 noise variables.

Two figures are made to visualise the simulated effect sizes for each group (i.e. schizophrenic/SZ and healthy control/HC).

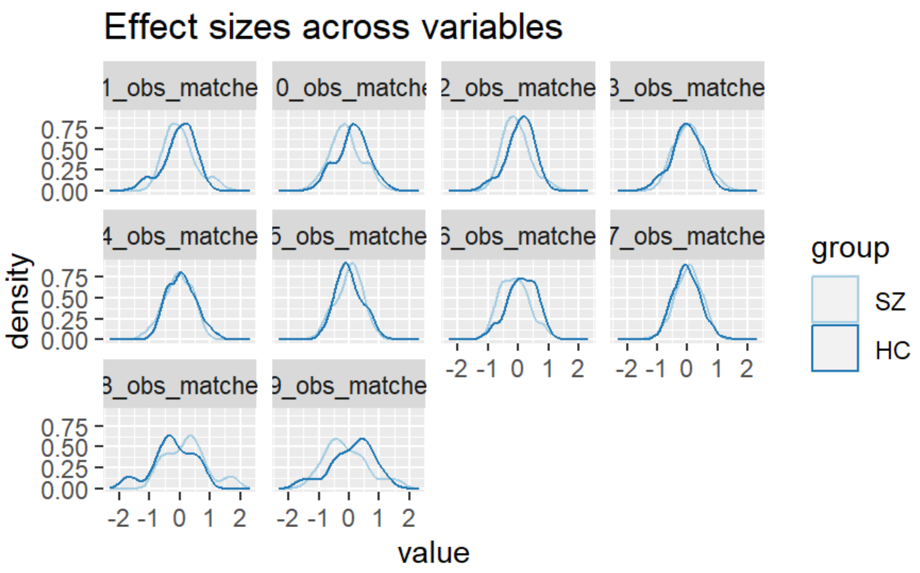

Figure 1. Effect sizes of the informed simulated data

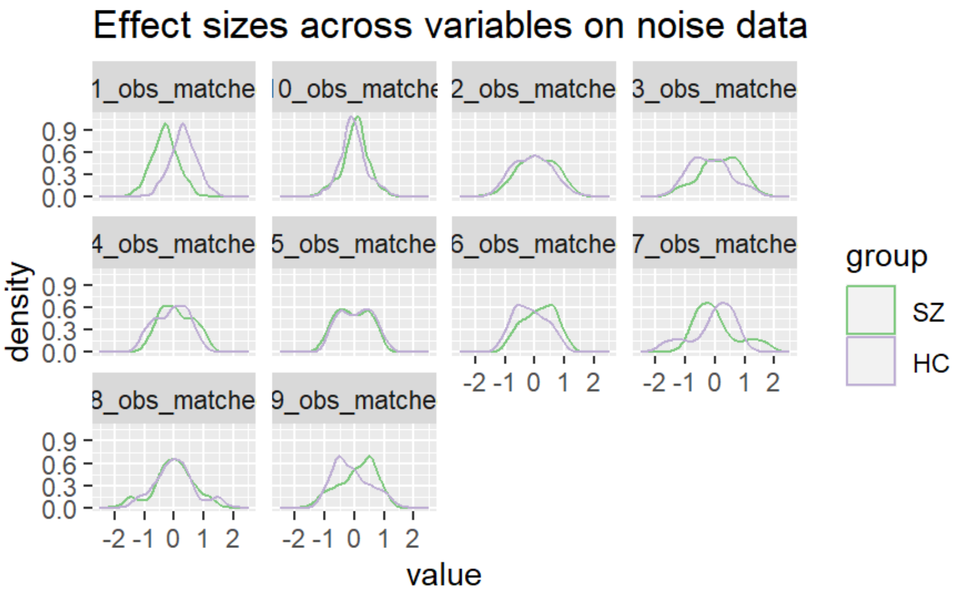

Figure 2. Effect sizes of the noise simulated data

The two figures show the effect size distributions for each acoustic feature across the two groups. From these, we can visually assess that the two datasets, i.e. informed and noise, differ.

Part 2

ML pipeline on simulated data

Describe your machine learning pipeline.

- Produce a diagram of it to guide the reader (e.g. see Rybner et al 2022 Vocal markers of autism: Assessing the generalizability of ML models).
- Describe the different parts: data budgeting, data pre-processing, model choice and training, assessment of performance. discuss whether performance is as expected and feature importance is as expected.

- Briefly justify and describe your use of simulated data, and results from the pipeline on them.

### Discuss why our noisy variables are "more" important and why we don't take them out

Part 3

Applying the ML pipeline to empirical data

- Third we apply the pipeline to the empirical data.
- Describe results from applying the ML pipeline to the empirical data and what can we learn from them.