Modeling individual differences and assigning individuals to latent groups

Daniel Porawski
7. August 2023

Approach

Compare alternative statistical methods for discovering latent classes in data with properties such as typical psycholinguistic data (in terms of repeated measures design) and specifically properties that reflect data collection with many subjects and few observations per subject, by performing simulation experiments

Motivation

What are latent classes?

Why are they useful?

What are latent classes?

Latent classes are homogeneous subgroups of subjects, that are not directly observable by the experiment.

The membership to a certain class of a subject describes the expected answering pattern of this subject.

Correlations between items are explained by the
presence of a priori unknown subpopulations (latent classes).

Why are latent classes useful?

Helpful for decision and policy making

Reveal hidden factors

Confirm existence of typological differences

Method

Using real data from “Partner effects and individual differences on perspective taking” (Loy & Demberg, 2022) and generate synthetic data based on that.

Creating multiple data sets that reflect difficulty levels and study designs.

Apply the considered methods on the scenarios to determine how well they can retrieve the underlying classes and preserve cluster structure.

Original Data Part I: Perspective taking

Critical trials
Participants of the study had to decide whether to drag and drop the object from their own or their partners perspective.

Different or same perspective tasks with each:
2 “left”, 2 “right”, 2 “front”, 2 “back”
critical trials for 180 subjects

Filler trials
Uniquely identifiable objects

Figure 1: A critical different perspective left-trial (Loy & Demberg, 2022)

Original Data Part I: Perspective taking

Critical trials
Participants of the study had to decide whether to drag and drop the object from their own or their partners perspective.

Different or same perspective tasks with each:
2 “left”, 2 “right”, 2 “front”, 2 “back”
critical trials for 180 subjects

Filler trials
Uniquely identifiable objects

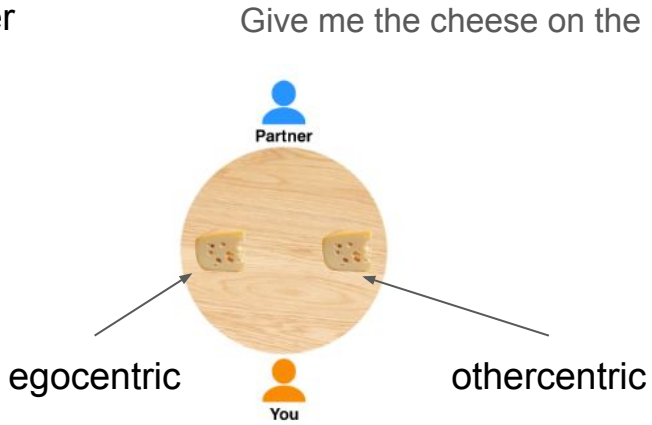


Figure 1: A critical different perspective left-trial (Loy & Demberg, 2022)

Original Data Part II : Individual differences battery

Also the scores of 3 tests for individual differences (instruments) were recorded

1. Object Perspective Test (OPT)
2. Stroop Task
3. Autistic Quotient

The original study had the goal to investigate their combined effect on perspective taking

The Data Part III : Overview

	workerID	perspective	targetObj	targetPos	respObjPos	respTime	gender	age	accuracy	acc	total	own.cod	other.cod	responderType	aq_score_subset	aq_score_total	opt_score_total
1	5RXRA10	same	shoe	L	L	1102	female	60	1	43	44	1	0	mixed	55	119	540
2	5RXRA10	different	clock	R	R	1104	female	60	NA	43	44	1	0	mixed	55	119	540
3	5RXRA10	same	lemon	R	R	1491	female	60	1	43	44	1	0	mixed	55	119	540
4	5RXRA10	different	vase	L	L	901	female	60	NA	43	44	1	0	mixed	55	119	540
5	5RXRA10	different	stapler	L	L	1002	female	60	NA	43	44	1	0	mixed	55	119	540
6	5RXRA10	same	glove	F	F	1571	female	60	1	43	44	1	0	mixed	55	119	540
7	5RXRA10	different	frog	B	F	1042	female	60	NA	43	44	0	1	mixed	55	119	540
8	5RXRA10	same	plant	B	B	997	female	60	1	43	44	1	0	mixed	55	119	540
9	5RXRA10	same	kettle	F	F	1678	female	60	1	43	44	1	0	mixed	55	119	540
10	5RXRA10	different	potato	R	L	2571	female	60	NA	43	44	0	1	mixed	55	119	540
11	5RXRA10	same	crab	L	L	2463	female	60	1	43	44	1	0	mixed	55	119	540
12	5RXRA10	different	cheese	F	B	766	female	60	NA	43	44	0	1	mixed	55	119	540
13	5RXRA10	same	coconut	B	B	2076	female	60	1	43	44	1	0	mixed	55	119	540
14	5RXRA10	different	key	B	F	1011	female	60	NA	43	44	0	1	mixed	55	119	540
15	5RXRA10	same	cap	R	R	844	female	60	1	43	44	1	0	mixed	55	119	540
16	5RXRA10	different	mushroom	F	B	640	female	60	NA	43	44	0	1	mixed	55	119	540

Statistical methods to investigate

1. Latent Class Analysis - A method usually used in these cases

2. Linear Mixed Effect Model - Analysis of random effect structure
   As an alternative, by-subject effects are utilized instead as seen as byproduct

3. Bayesian Mixed Effect Model - Analysis of random effect structure
   Due to few observations per subject, maybe performs better than LMEM

LCA

For a dichotomous variable the Latent-Class-Model for G classes can be written as:

$$p(X_{vi} = 1) = \sum_{g=1}^{G} \pi_g \, \pi_{ig}$$

with $p(X_{vi} = 1)$ being the probability that subject v has the value 1 for item i.

The two estimated parameters are

π_g is the probability to belong to the latent class g, with $\sum_{g=1}^{G} \pi_g = 1$, and

π_{ig} is the probability for a subject of class g to choose 1 for item i

The number of classes G is determined by model comparisons

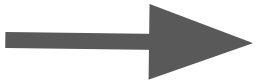
Christian Geiser (2010),
Datenanalyse mit Mplus

LCA

Using R package poLCA : Polytomous Variable Latent Class Analysis

Necessary to transform input to wide format.
2 ≙ egocentric choice, 1 ≙ othercentric choice (numbering due to poLCA limitations)

workerID	perspective	targetObj	targetPos	respObjPos
5RXRA10	different	clock	R	R
5RXRA10	different	vase	L	L
5RXRA10	different	stapler	L	L
5RXRA10	different	frog	B	F
5RXRA10	different	potato	R	L
5RXRA10	different	cheese	F	B
5RXRA10	different	key	B	F
5RXRA10	different	mushroom	F	B



workerID	B1	B2	F1	F2	L1	L2	R1	R2
5RXRA10	1	1	1	1	2	2	2	1

With no covariates BIC suggests 3 class model

LCA - Result on base data

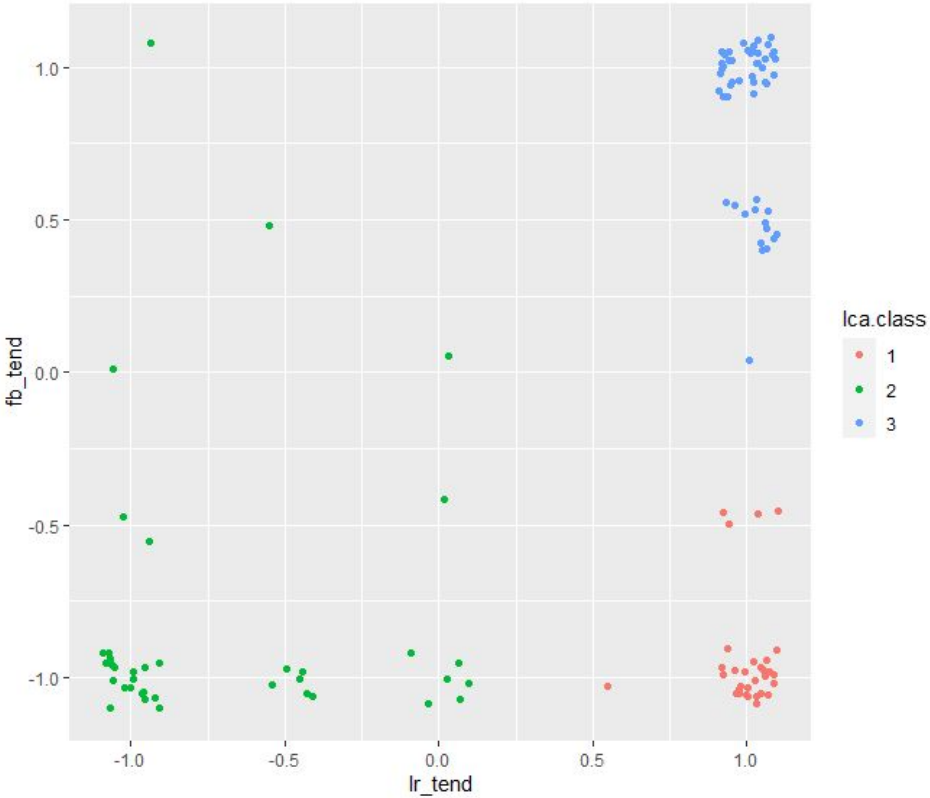


Figure 2: LCA clustering on original data

When assigning names to classes, researcher should be aware to not engage in “naming fallacy”

Number of classes determined by information criteria like BIC, AIC, Log Likelihood

The lowest BIC can still result in classes that are difficult to interpret.
It tends to put isolated points into their own class.

Analytic Hierarchy Process (Akogul & Erisoglu, 2017) can be used to find explainable classes

LMEM

Random effects capture differences between subjects and between items

```
m0 <- lmer(own.cod ~ targetPos +
                    (targetPos|workerID),
                     data=lmem_dat)
```

Each subject has their own random

slope for different perspective trial types.

We can also observer the similarity between
F and B trial types, and L and R trial types.
Hence, the idea to group them into FB and LR.

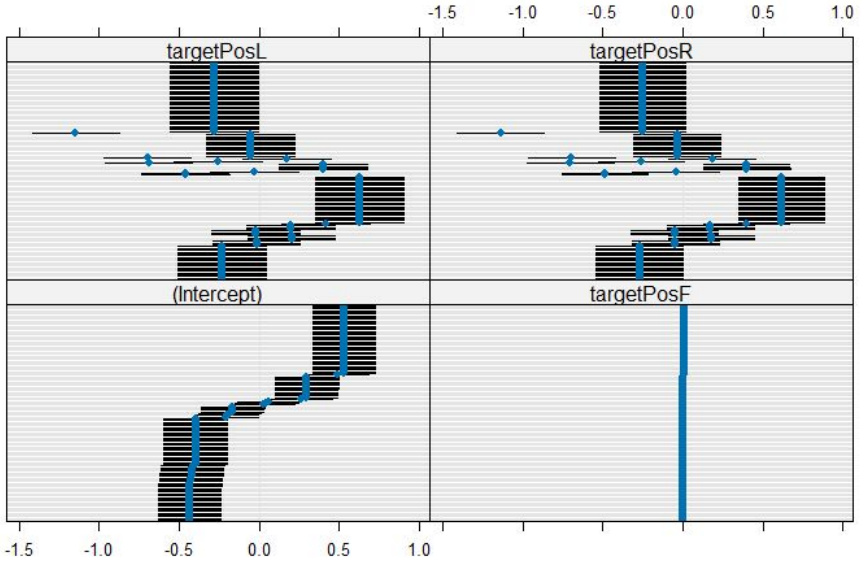


Figure 3: Caterpillar plots for random intercepts

By analyzing the random effect structure,

it is possible to group subjects into classes by their random slopes.

LMEM

From coef(m0)$workerID :

	(Intercept)	targetPosF	targetPosL	targetPosR
1R9GA1U	0.26488438	-0.006036502	-0.26043484	-0.24740217
1ZD7A3L	0.95959968	-0.001326146	-0.94292187	-0.89707071

prediction of othercentric choice for:

First subject 1R9GA1U

targetPos B: **0.265** F: 0.265-0.006≈**0.26** L: 0.265-0.26≈**0.01** R: 0.265-0.247≈**0.02**

Is egocentric but with a slight tendency to choose othercentric for FB

Second subject 1ZD7A3L

targetPos B: **0.96** F: 0.96-0.001≈**0.96** L: 0.96-0.943≈**0.02** R: 0.96-0.9≈**0.04**

Is FB othercentric but LR egocentric

LMEM - Results on original data

Comparing Random Slopes for B, F, L and R

Combine to $FB = \frac{F+B}{2}$ and $LR = \frac{L+R}{2}$

for each subject.

Apply latent profile analysis (LPA)

with Analytic Hierarchy Process (AHP)

to determine good thresholds for classes

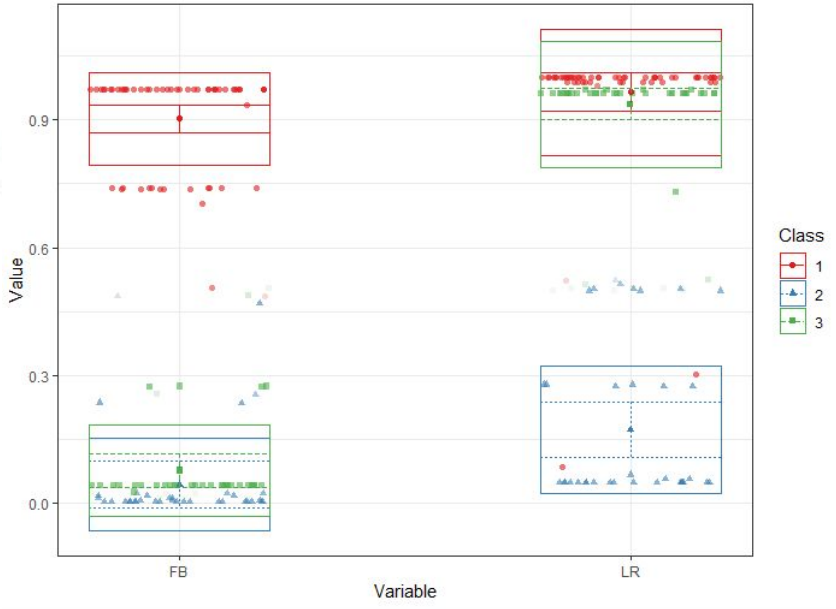


Figure 4: LPA clustering for 2 variables

LMEM - Results on original data

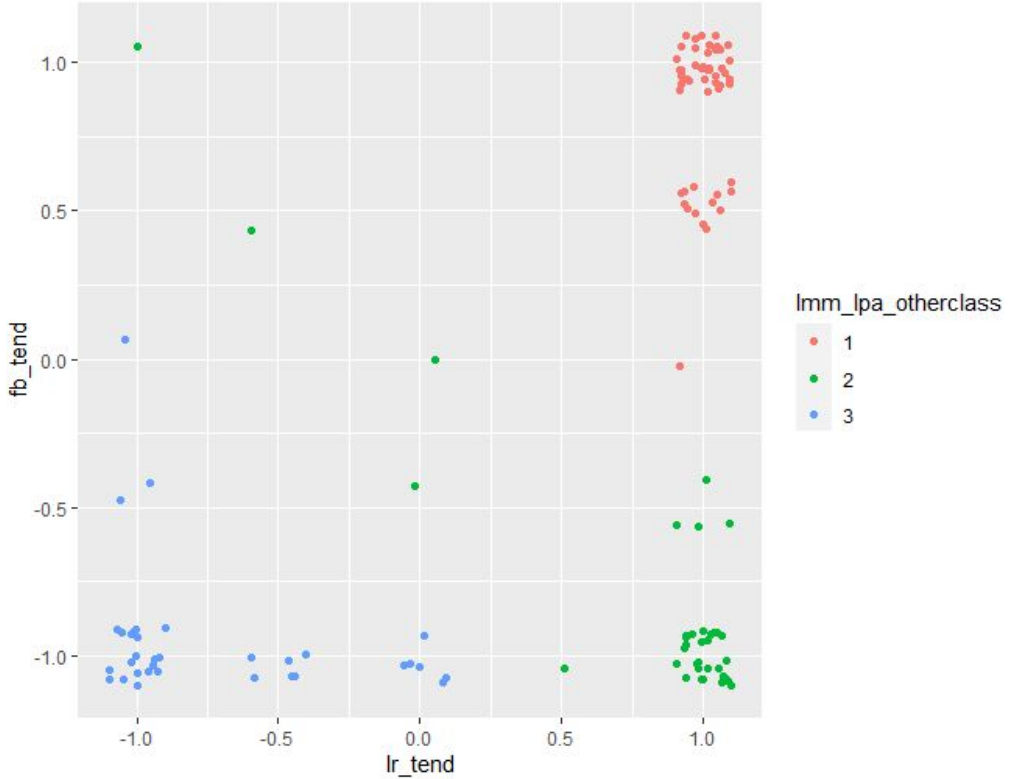


Figure 5: Clustering from LMEM

Slightly different results from LCA

Easier to interpret as

1 Ego-centrics

2 Mixed responders

3 Other-centrics

BMEM

Specify a bayesian model with BRMS package

Apply the same random structure analysis from LMEM

```
bm1 <- brm(data = lmem_dat, own.cod ~ targetPos + (targetPos | workerID),
           prior = c(prior(normal(0.5,3), class = Intercept),
                     prior(normal(0,2), class = b),
                     prior(normal(0, 1), class = sd)),
           seed = 123,
           cores = 8, iter = 4000, warmup = 2000)
```

BMEM

The model reveals a quite bad fit

with `pp_check`.

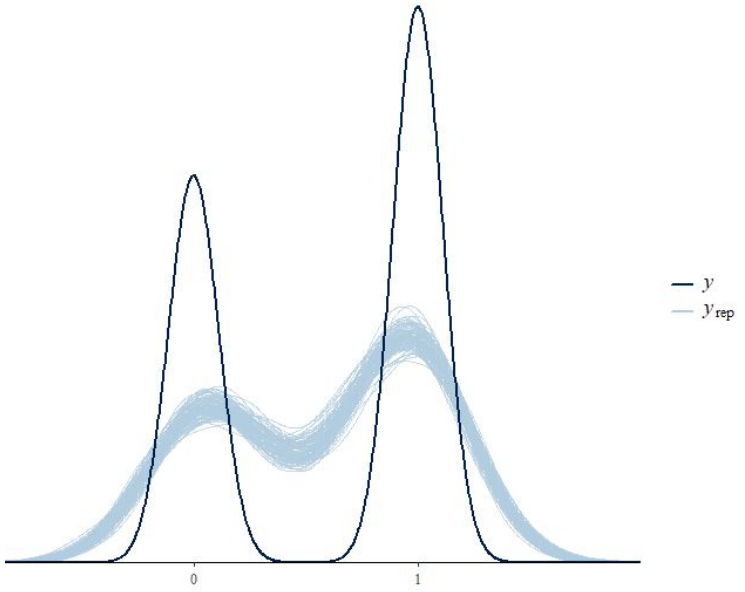


Figure 6: Posterior predictive check for the BMEM with gaussian family

BMEM

The model reveals a quite bad fit.

The extracted random slopes suggest 3 classes. Like LCA it generates a difficult to interpret pattern:

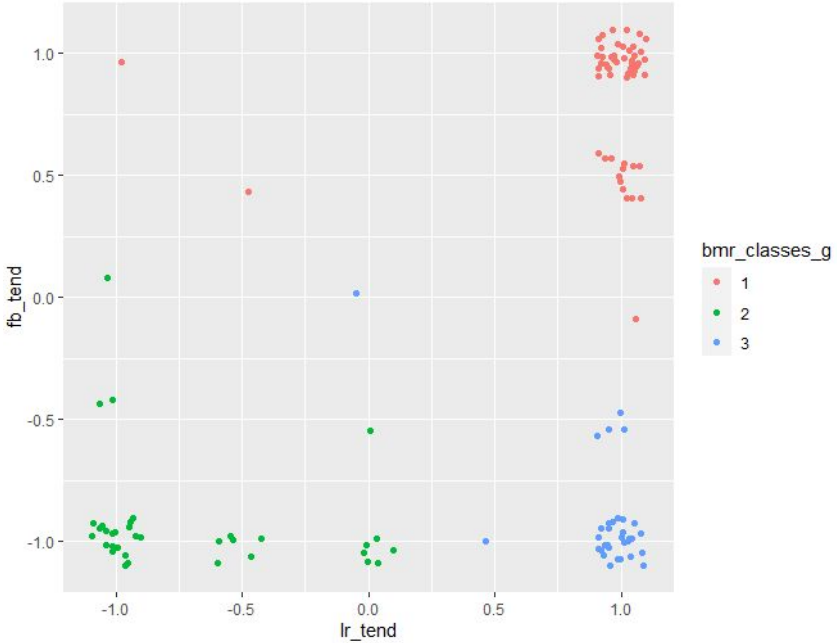


Figure 7: Clustering for gaussian BMEM

BMEM - Bernoulli approach

Specifying an other BMEM

with `family = bernoulli()`

produces a nicer fit

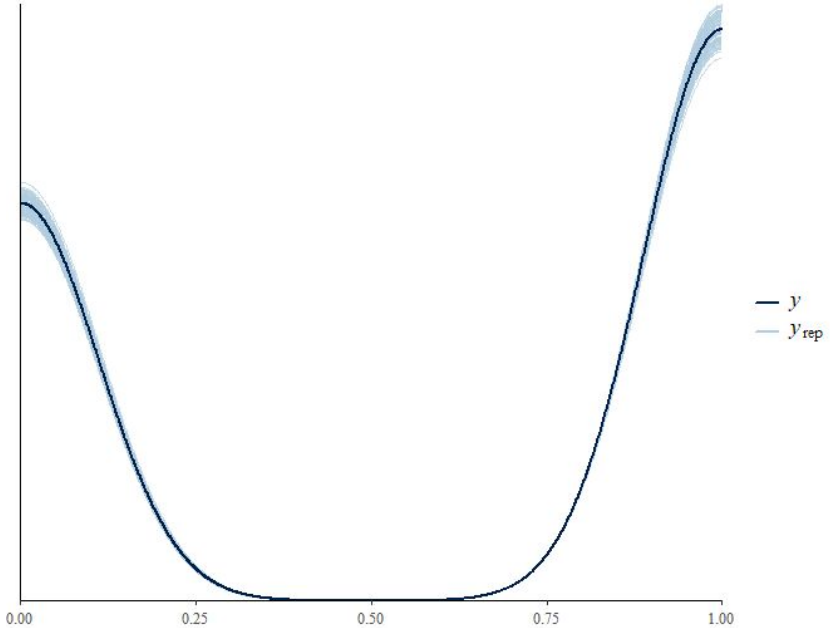


Figure 8: Posterior predictive check for BMEM with Bernoulli family

and the extracted slopes …

BMEM - Bernoulli results

… result in many more classes.

This seems more promising.

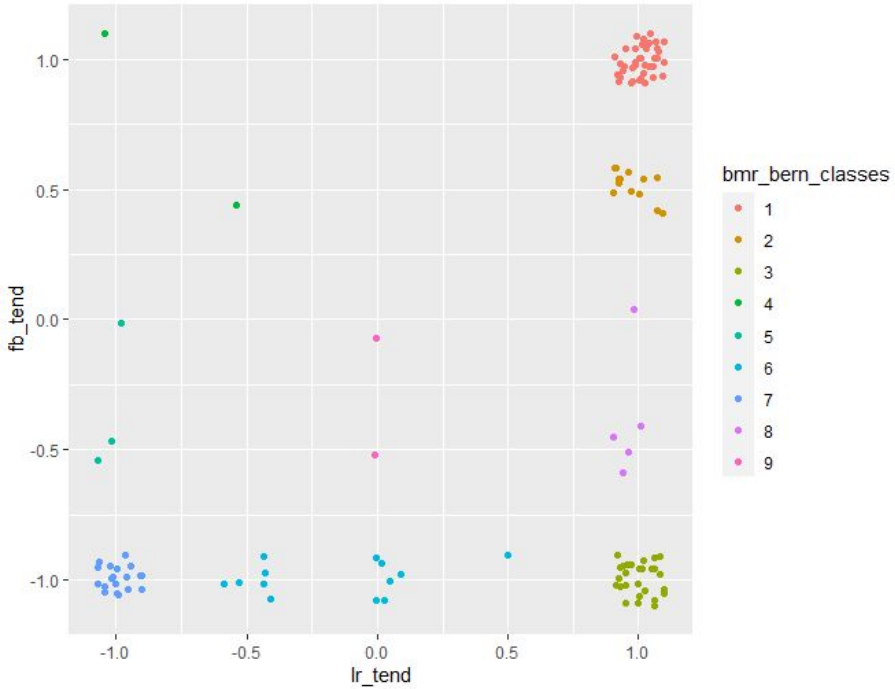


Figure 9: Clustering for bernoulli BMEM

Simulating data

Each subject belongs to one of 25 base classes, that will be used as underlying ground truth.

A class consist of a probability for egocentric perspective taking on front-back and one for left-right trials.

Class sizes can be adjusted to remove classes or create major and minor classes.

Each subject has also an individual tendency ~ N(0, 0.05) to choose egocentric that is added to their class probabilities.

Figure 10: Centroid positions and number(name) of basic class centroids

Simulating data

Modeled by

$$P_s(FB = 1) = g_s(FB) + s_{FB}$$

$$s_{FB} \sim \mathcal{N}(0, 0.05)$$

g being the class number of subject *s*

$P_s(FB = 1)$ Probability for subject *s* to choose egocentric (=1) in a front or back trial

$g_s(FB)$ Probability of subject from class *g* ∈ *{1,...,25}* to choose egocentric in a front or back trial

s_{FB} Individual front-back difference for *s*

Probabilities <0 or >1 will be set to 0 or 1 respectively

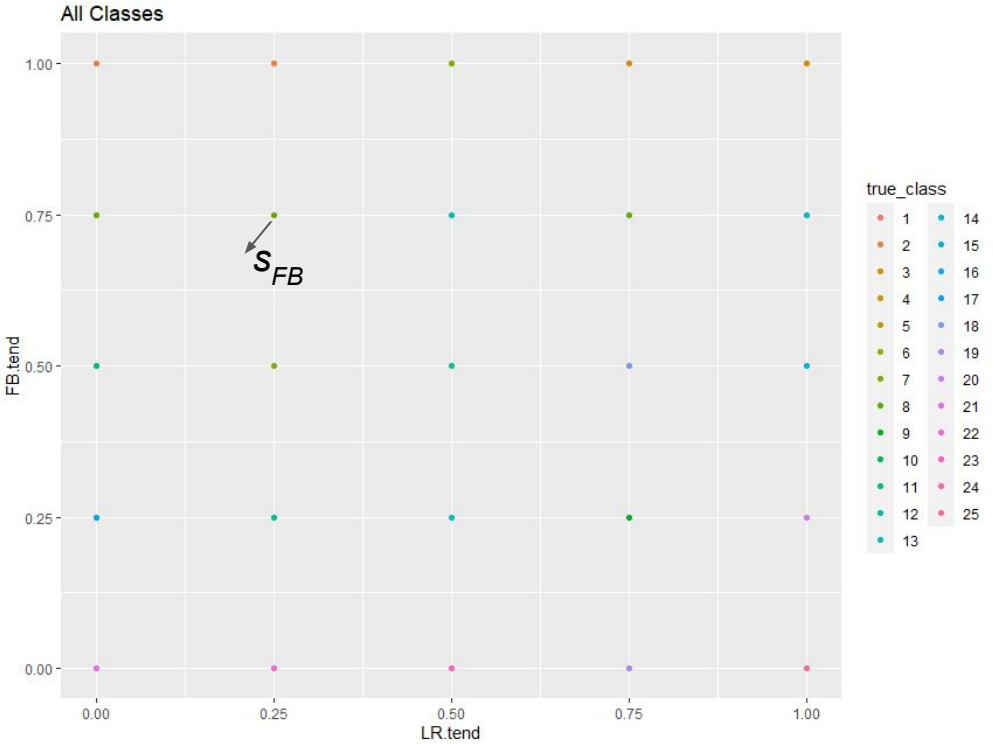


Figure 11: Each participant has an own individual variation from their class centroid

Simulating data : Influence of observations

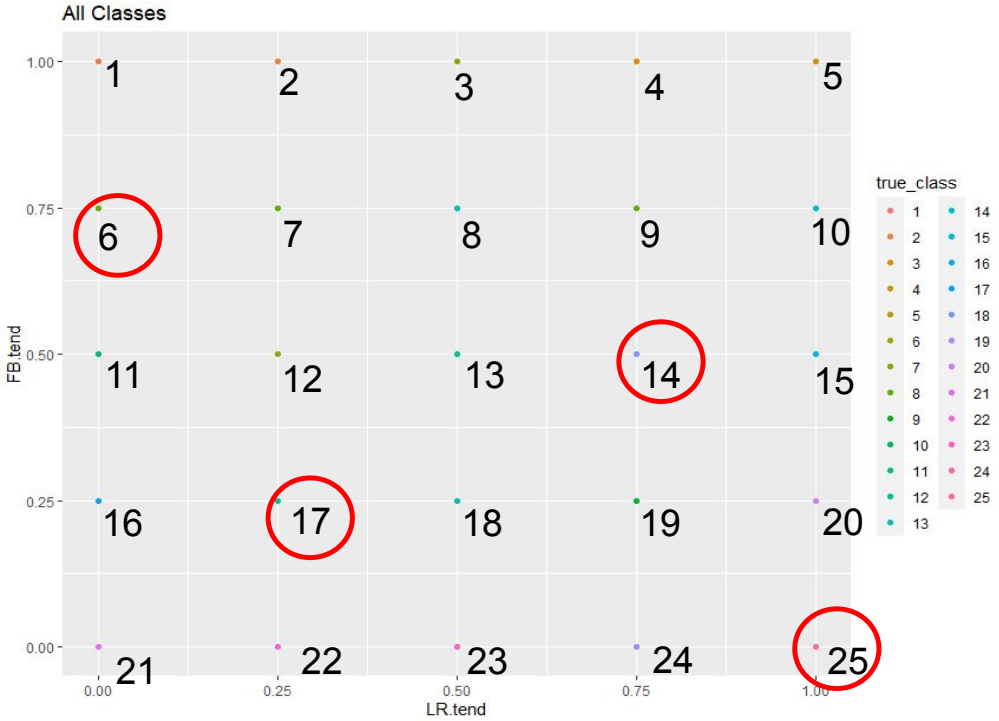


Figure 12: Class centroids determine the probability to choose egocentric for FB or LR trial types

Lets generate a dataset with only the classes 6, 14, 17, and 25.
We use 2 observations per trial type and 100 subjects.

A generated subject will have the same probability to belong to one of the classes. Represented by the class distribution matrix *C*

$$C = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Simulating data : Influence of observations per trial type

The generated points overlap,

since for 2 obs. per trial type only

5 tendency values for each

dimension are possible.

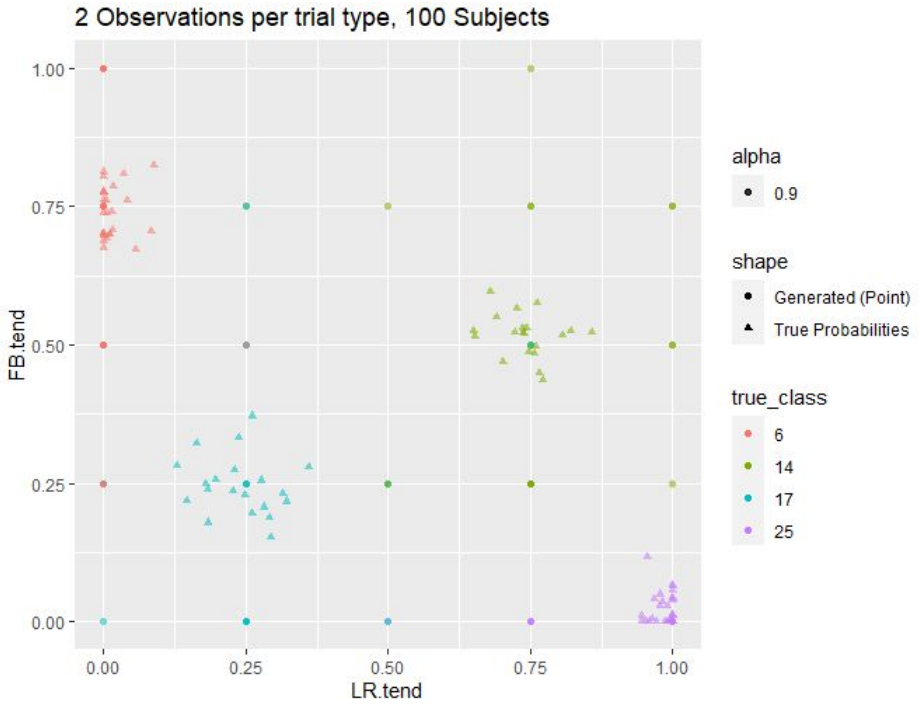


Figure 13a: Individual variation can be spotted for true probabilities. Generated points are discretized.

Simulating data : Influence of observations per trial type

Jittering helps to see the generated point better.

Let’s increase the number of obs. per trial type to 20 …

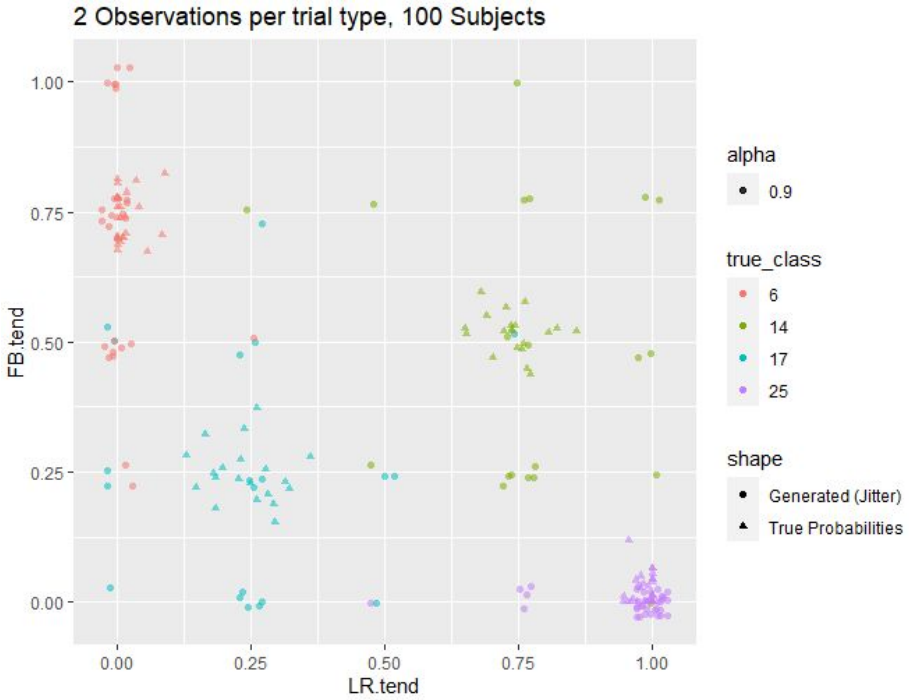


Figure 13b: With jitter we avoid overlapping to a certain degree

Simulating data : Influence of observations per trial type

The spread out cluster got much denser. No need to jitter anymore.

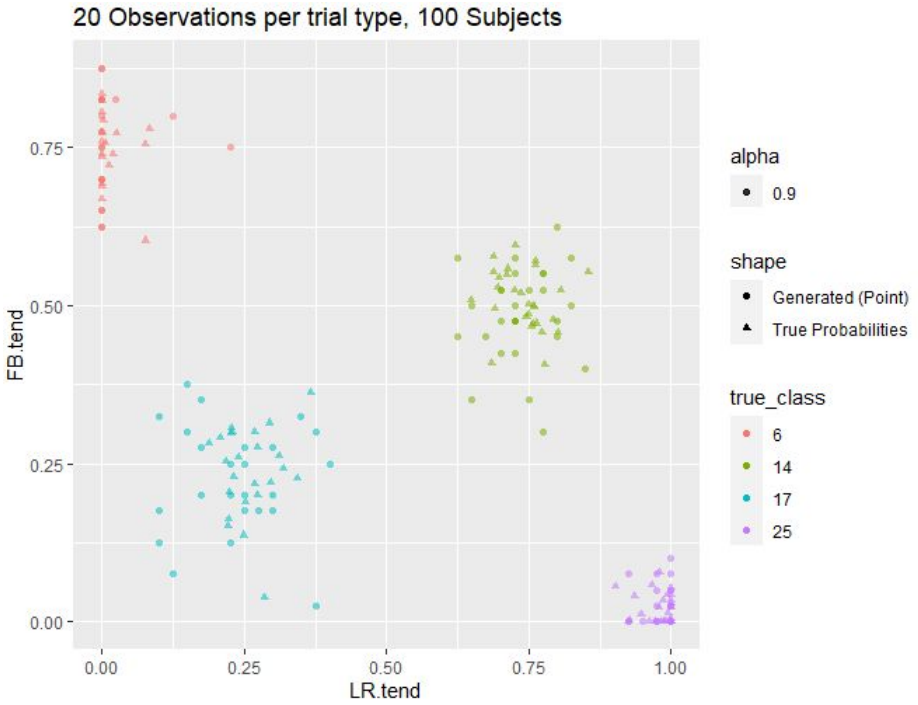


Figure 14: With more observations per trial type the clusters get more defined

Simulating data : Influence of number of subjects

Back to 2 obs. per trial type.
Now with 400 subjects.

Maybe the models will be able to estimate the true number of classes more accurately.

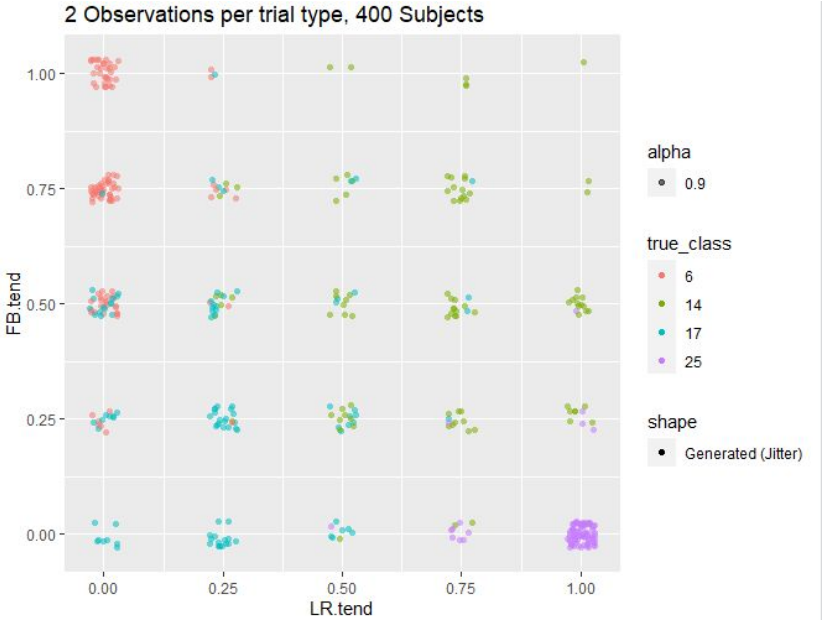


Figure 15: More subjects don't prevent classes overlapping, but maybe give more certainty of where centroids are.

Simulating data : Generated data sets

By changing the class distribution Matrix, data sets can be created that reflect more or less difficult scenarios.

Planned are 9 different scenarios, each 3 of easy, medium and hard difficulty.

To show the influence of observations per trial type all experiments are performed with 2, 4 and 6 observations per trial type.

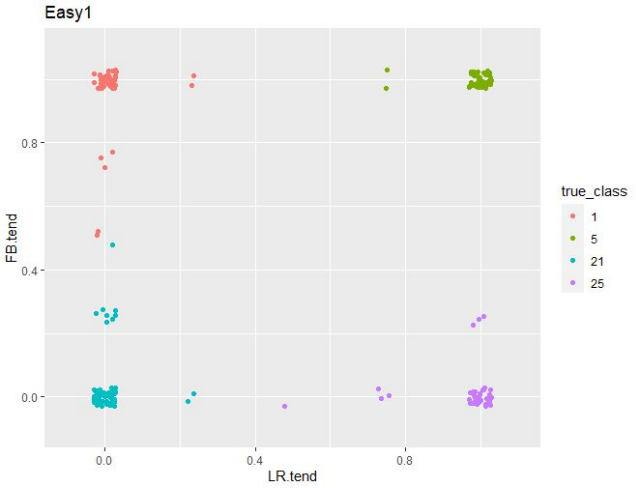
The number of subjects will be 100, 200 (original experiment), and 400

Total of 9 X 3 X 3 = 81 data sets

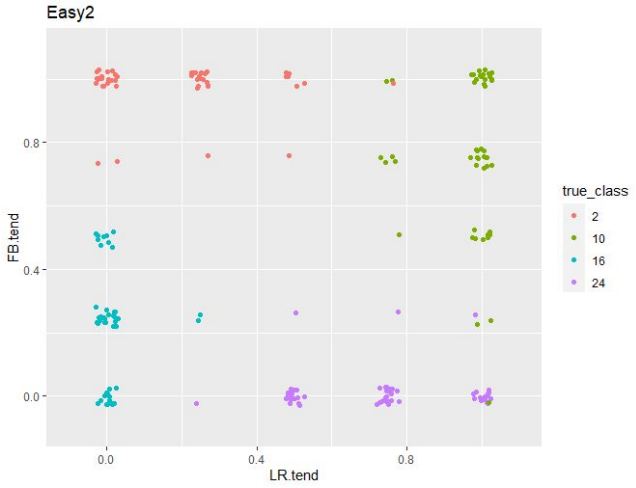
Next up: Examples based on 2 observations per trial type and 200 subjects.

Analyze Results : Scenario difficulties (Easy)

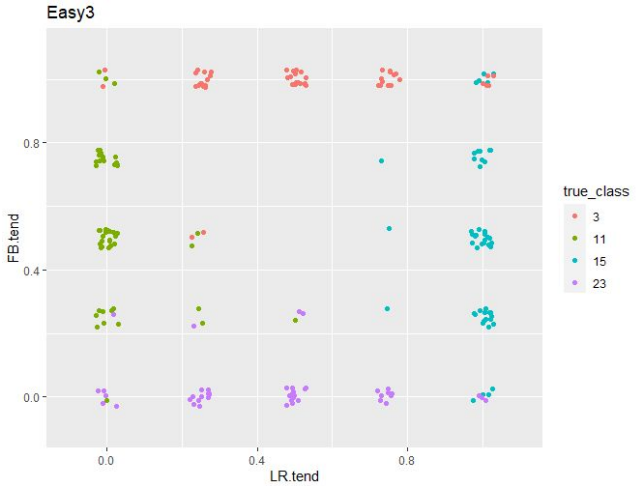
Easy scenarios have clearer tendencies and centroids far apart, resulting in minor overlap.




```
c( 1,  0,  0,  0,  1,
   0,  0,  0,  0,  0,
   0,  0,  0,  0,  0,
   0,  0,  0,  0,  0,
   1,  0,  0,  0,  1)
```




```
c( 0,  1,  0,  0,  0,
   0,  0,  0,  0,  1,
   0,  0,  0,  0,  0,
   1,  0,  0,  0,  0,
   0,  0,  0,  1,  0)
```




```
c( 0,  0,  1,  0,  0,
   0,  0,  0,  0,  0,
   1,  0,  0,  0,  1,
   0,  0,  0,  0,  0,
   0,  0,  1,  0,  0)
```


Figure 16: The 3 easy scenarios and their corresponding class distribution.

Analyze Results : Scenario difficulties (Medium)

Medium scenarios have less clear tendencies, imbalances and more classes. Resulting in more overlap and spread out classes. Medium1 is meant to mimic the original data.

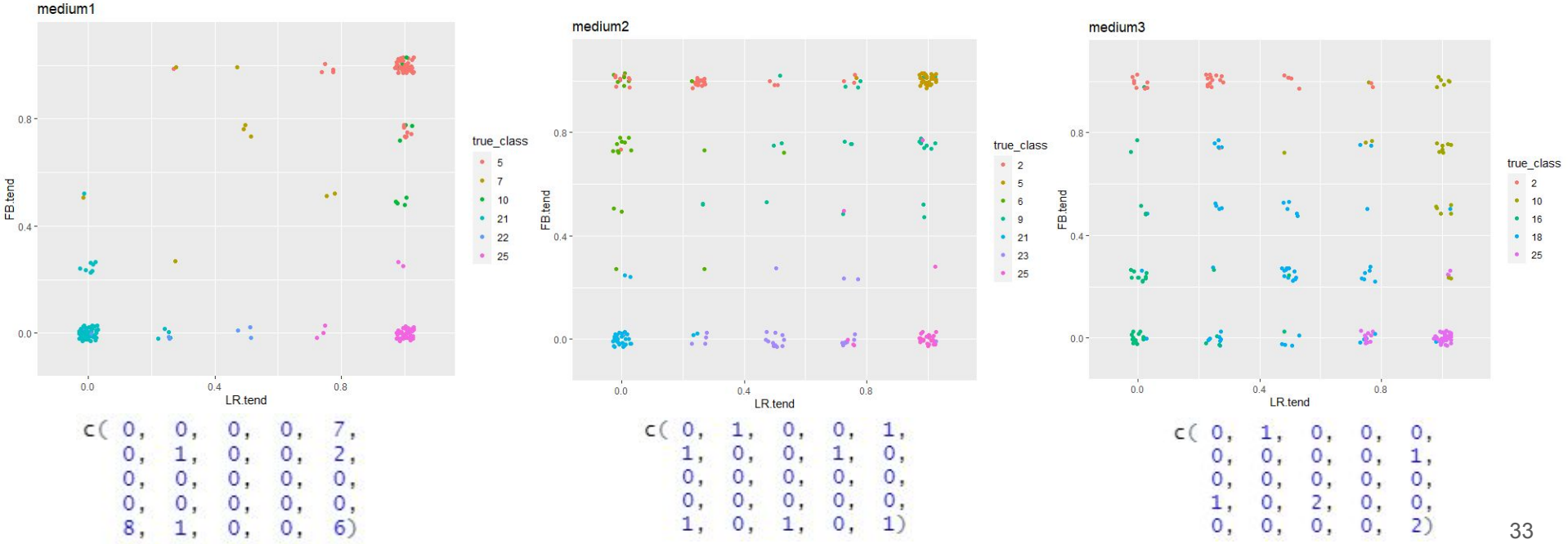


Figure 17: The 3 medium scenarios and their corresponding class distribution.

Analyze Results : Scenario difficulties (Hard)

Hard scenarios are extremely difficult for any clustering algorithm. They have less clear tendencies, strong imbalances and close centroids. Resulting in more overlap, spread out dominant classes that obscure others. It's also possible that noise from some classes will aggregate to new clusters (see hard2).

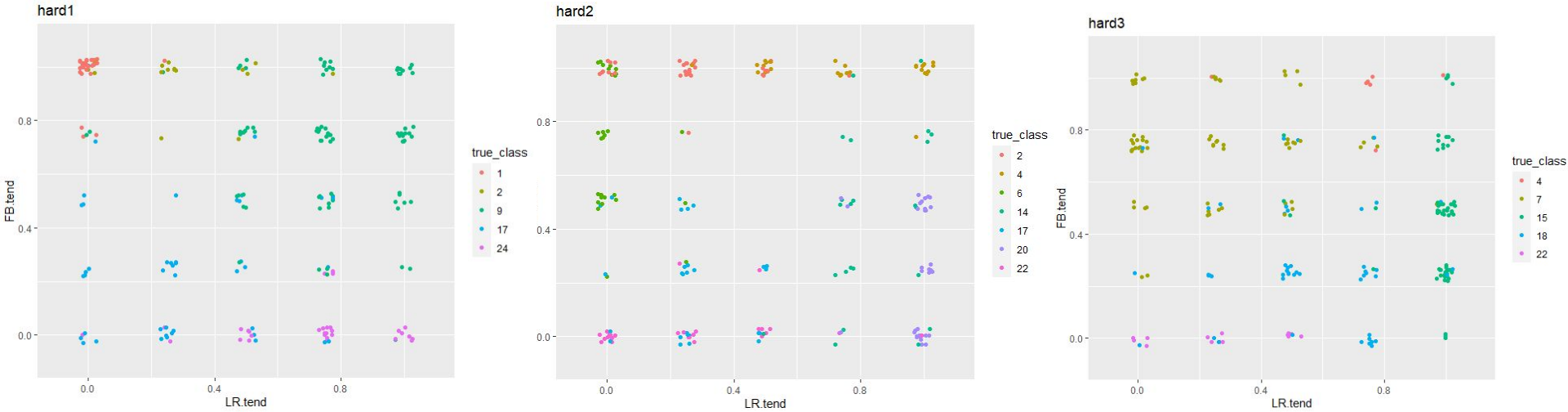


Figure 17: The 3 hard scenarios and their corresponding class distribution.

Analyze Results - Performance Measures

RI and NMI are both in range [0,1] and are popular performance measure in classification and clustering.

They compare the predicted class memberships (X) to the ground truth (Y).

$$RI = \frac{\text{Number of pair-wise correct predictions}}{\text{Total number of possible pairs}}$$

$$NMI = \frac{H(X) + H(Y) - H(X,Y)}{H(X,Y)}$$

RI is shown to be more sensible to false negatives (putting 2 subjects from the same ground truth in in different classes) than to false positives (putting 2 subjects from different ground truth in the same class).

Result tables for 100, 200, and 400 participants

		Easy_1				…				Hard_3			
		pred. clusters	true clusters	RI	NMI	pred. clusters	true clusters	RI	NMI	pred. clusters	true clusters	RI	NMI
2 obs.	LCA												
	LMEM												
	BMEM												
4 obs.	LCA												
	LMEM												
	BMEM												
8 obs.	LCA												
	…												

Table 1: Proposed result summary table

Contribution

- Better understanding the influence of data on retrieval of latent classes
- Finding different strengths and weaknesses of the statistical methods discussed
- How to identify the data structure to use the best statistical method for the case
- Show the limits of the methods by revealing non-ideal cases

Planned schedule

August Week 2 Generating synthetic data, R scripts	September Week 1 Related Work	October Week 1 Background + Introduction	November Week 1 Buffer week
August Week 3 Running experiments, R scripts, Results	September Week 2 Related Work, Introduction	October Week 2 Abstract	
August Week 4 Method	September Week 3 Conclusion	October Week 3 Finalizing writing process	
	September Week 4 Half-Time buffer week	October Week 4 Printing and hand-in	

Related Work

1. Barr, D.J., et al. (2012),Random effects structure for confirmatory hypothesis testing: Keep it maximal
   - Background information on LMEM, Simulation study

2. Akogul, S. & Erisoglu, M. (2017),An Approach for Determining the Number of Clusters in a Model-Based Cluster Analysis
   - introduces the AHP to determine the number of clusters

3. Nylund, K. L. et al. (2007), Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study
   - Monte Carlo Simulation example

Related Work

4. Kliegel, R. et al. (2011). Experimental effects and individual differences in linear mixed models: estimating the relationship between spatial, object, and attraction effects in visual attention
   - Using LMEM to analyse variance-covariance matrix of random effects in order to find subjects deviation from grand mean RT and from fixed effects parameters
   - Source of inspiration for method in this thesis to classify random slopes

5. Geiser, C. (2010).Datenanalyse mit Mplus
   - Describes goals and application of LCA with the Mplus software
   - Example of LCA with log-likelihood to determine number of classes
     - Multiple runs with different initial values to avoid local maxima

6. Francot, R. et al. (2020). Profiles of bilingualism in early childhood: A person-centred Latent Profile Transition Approach
   -LPA example

Related Work

7. Sun, X. et al. (2012). Credibility of claims of subgroup effects in randomised controlled trials: systematic review
   - Pitfalls and shortcomings of subgroup claims

8. Rouder, J. N. & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks
   - Individual differences explained

9. Loy, J. & Demberg, V. (2022). Partner effects and individual differences on perspective taking.
   - Original anchor paper, source for data

Related Work

10. Imaizumi, T. et al. (2020). *Advanced Studies in Classification and Data Science*.
    - Information on Rand Index and Normalized Mutual Information

11. Linzer, D. A. & Lewis, J. B. (2011). poLCA: An R Package for Polytomous Variable Latent Class Analysis
    - poLCA manual

12. Preud'homme, G. et al.(2021). Head-to-head comparison of clustering methods for heterogeneous data: a simulation-driven benchmark
    - Extensive overview and benchmark of clustering algorithms on simulated data

UNIVERSITÄT DES SAARLANDES

Thank you for your time!