pathways

Pathways Segmentation Methods Training

Session 3 – Latent Class Analysis

IDM Gates Foundation

Session 3 Outline

- Review of where we are in the Pathways process
- Cluster analysis
- Latent class analysis (LCA)
- Technical considerations of running the model
- Tools for selecting the best model
- LCA process

01

Session 1 Homework Review

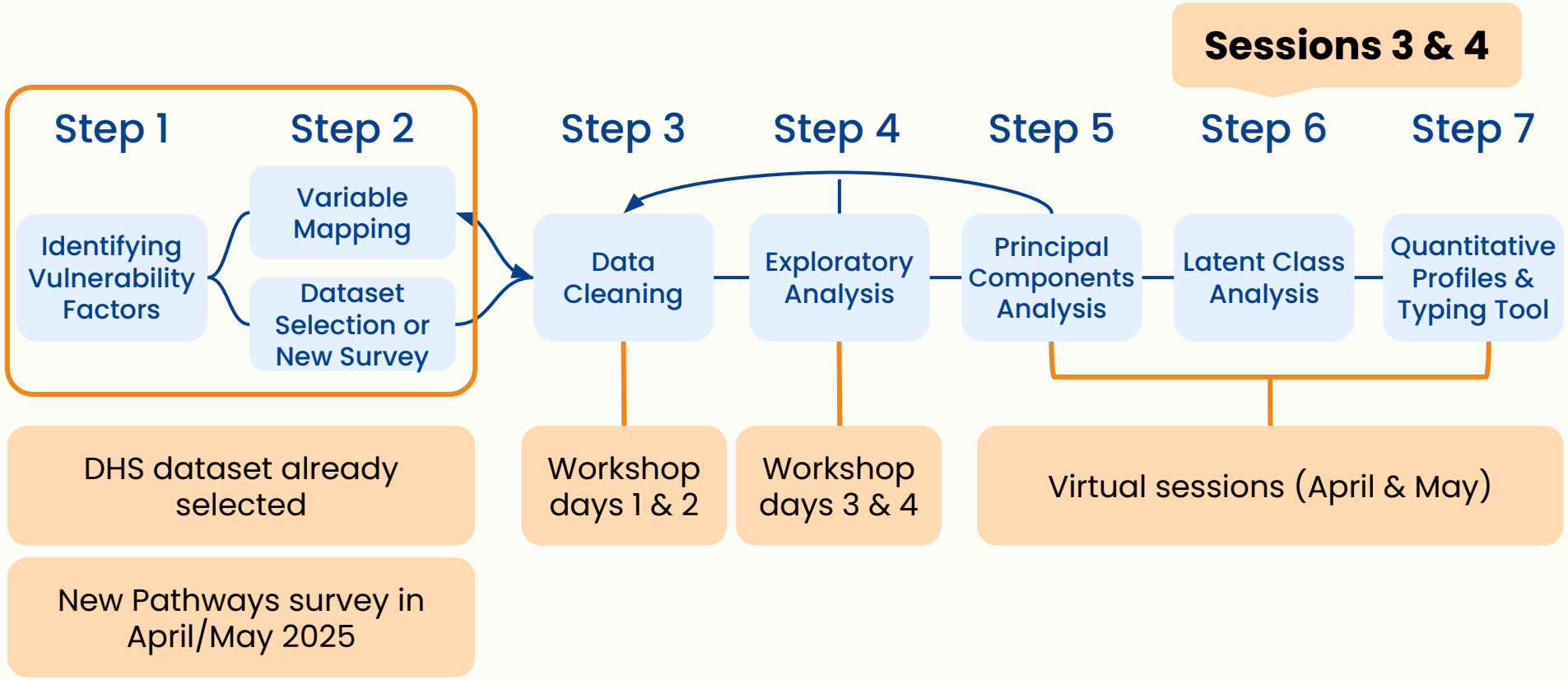
discussion

Do you have any questions about the homework?

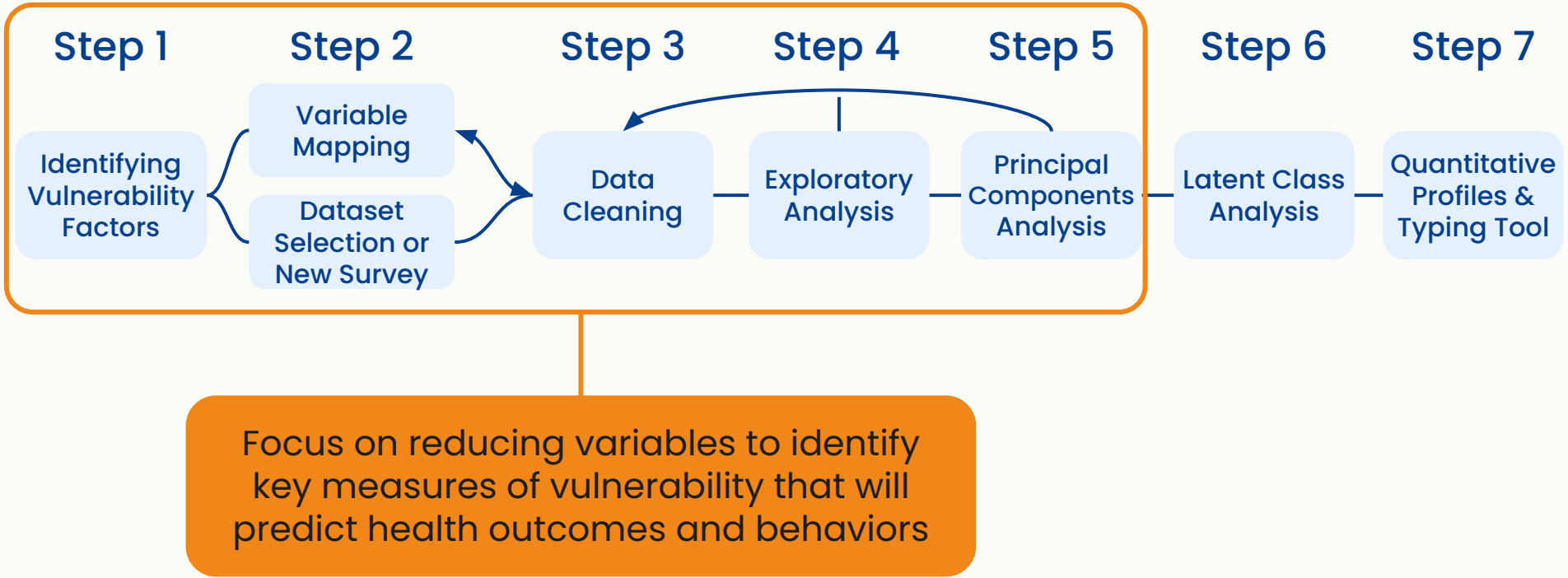
02

Where We Are in the Pathways Segmentation Method

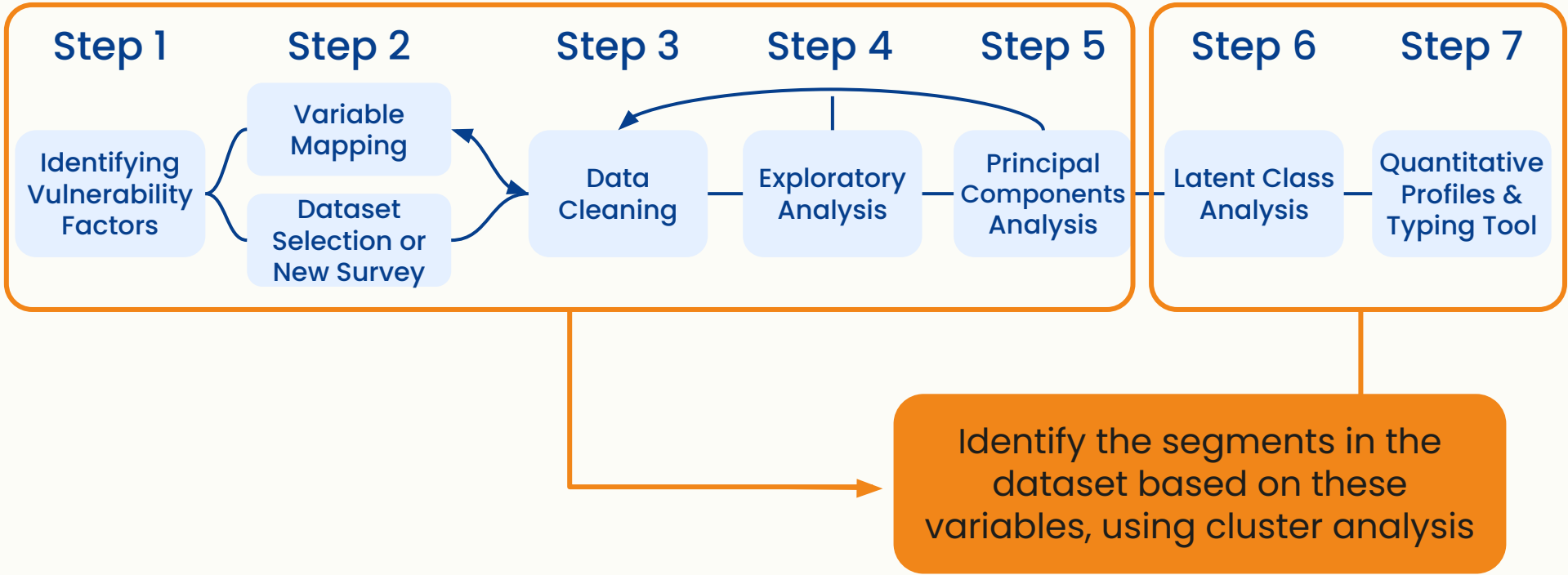
Timeline for segmentation training



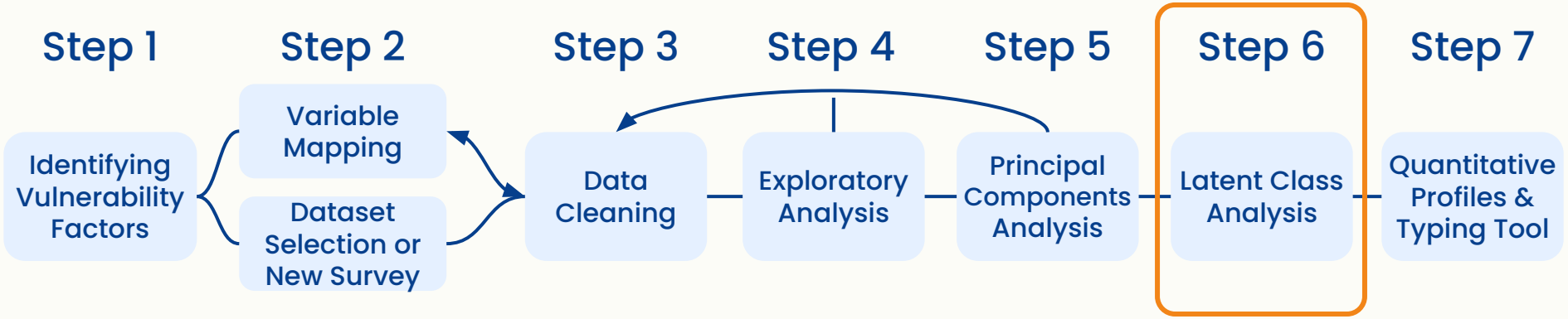
What are we trying to learn with step 6?



What are we trying to learn with step 6?



What are we trying to learn with step 6?



? How many segments exist in our dataset?

? What are the variables that drive separation between the segments and are they important to defining our segments?

03

Cluster Analysis

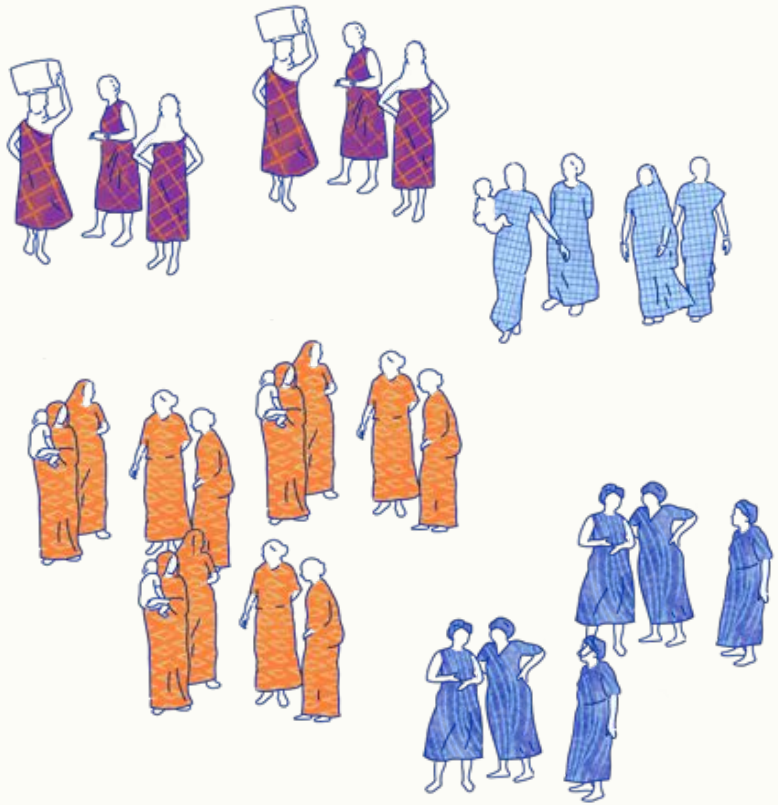
question

What is cluster analysis?

cluster analysis:

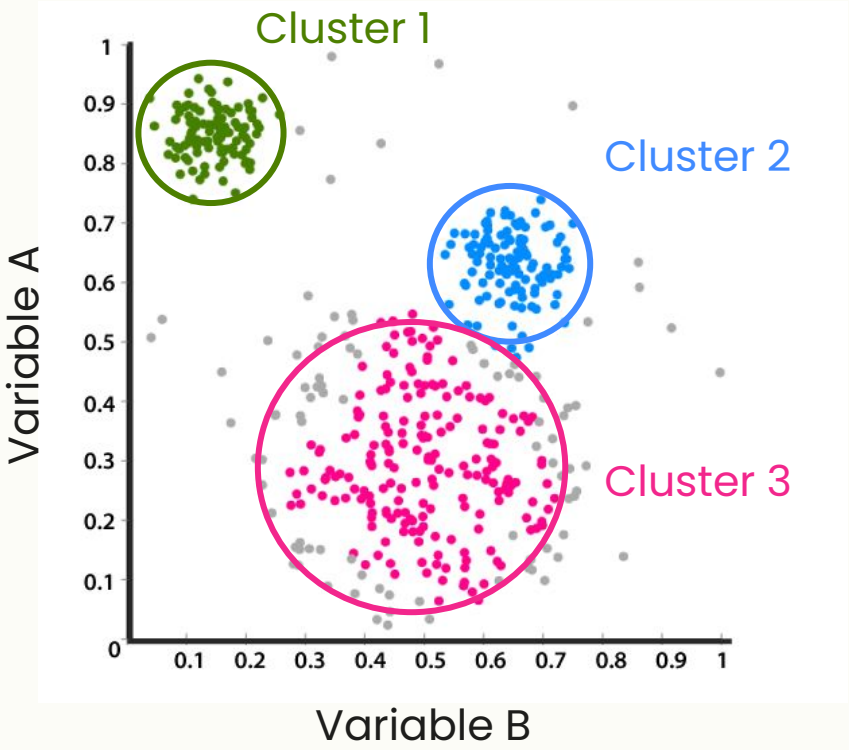
a set of statistical techniques which ask whether data can be grouped into categories on the basis of similarities or differences

The objective of cluster analysis is to find similar groups of subjects



Observations within each group are similar to one another with respect to variables or attributes of interest

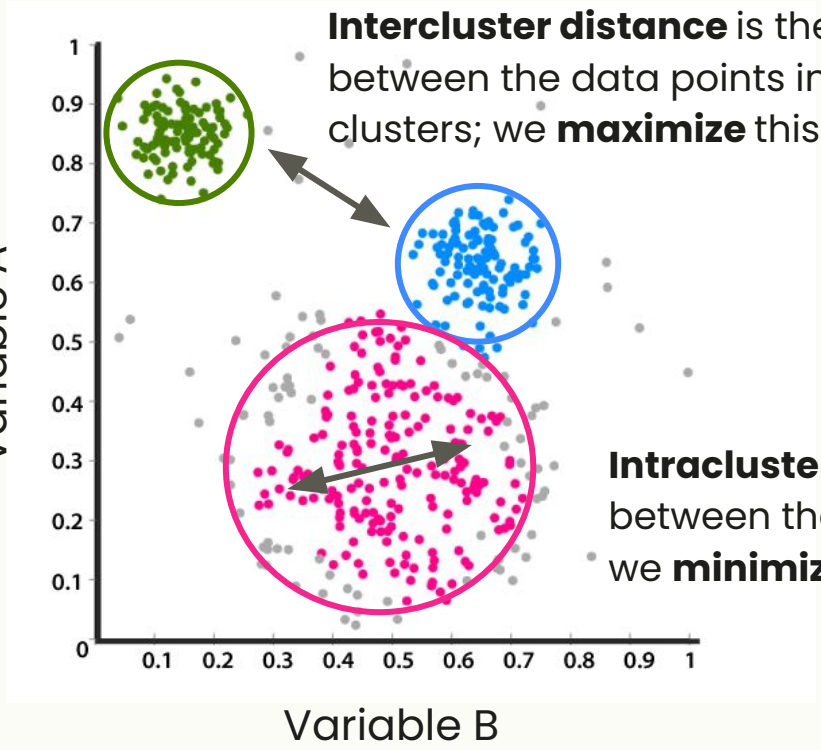
Observations between the groups are very different from one another



To develop clusters, we measure a set of variables on each subject in a dataset and try to either:

- minimize the differences among subjects within the same category

- maximize the differences between subjects belonging to different categories

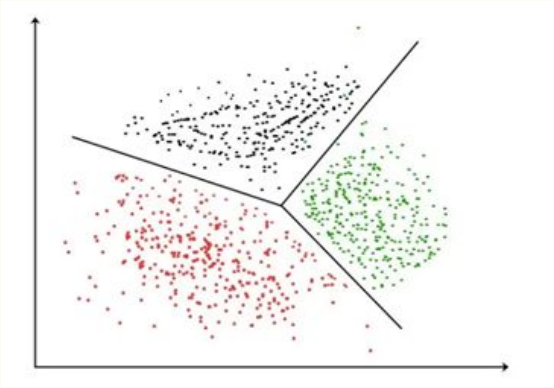


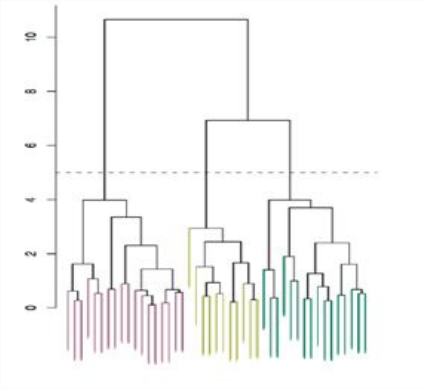
Intercluster distance is the difference between the data points in different clusters; we **maximize** this difference

Intracluster distance is the difference between the data points inside a cluster; we **minimize** this difference

Variables are compared between subjects and the clusters are developed.

Types of clustering methods





$$P(y \in \textbf{cluster1}) = f(x\textbf{1}, x\textbf{2}, \ldots)$$

Partition methods

Split up observations by measuring distance between observations

Hierarchical clustering methods

Assume smaller clusters merge into larger clusters, similarly, measure distance between observations

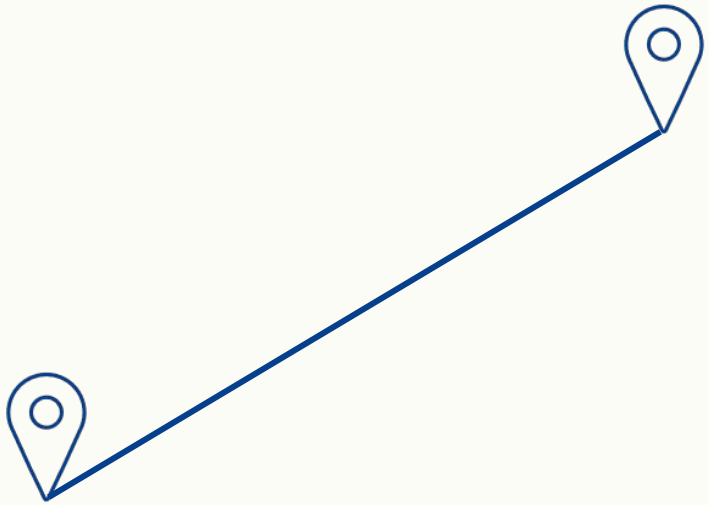
Model-based clustering methods

Assume an underlying model, which uses predicted probability of cluster membership and assigns clusters based on probability

Latent class analysis:

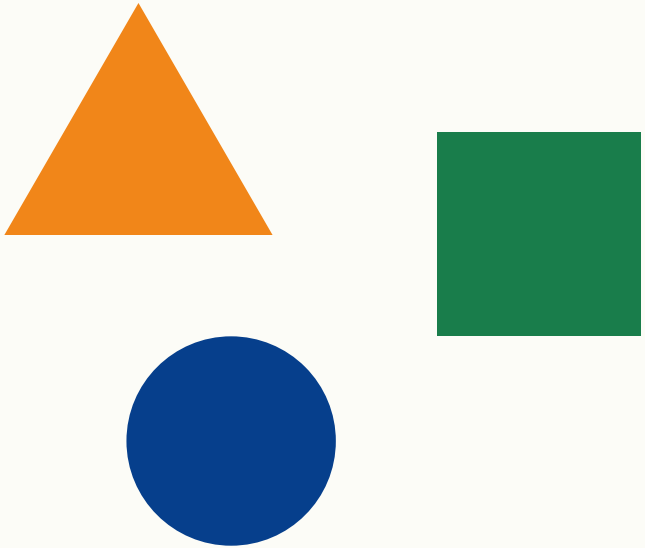
model based clustering method

Why use LCA versus other clustering methods?



Most clustering methods use distance metrics to measure this similarity between variables.

Why use LCA versus other clustering methods?



Distance metrics do not work well for categorical variables.

LCA is a model-based approach that can appropriately handle categorical variables.

04

Latent Class Analysis (LCA)

latent variable:

a variable that represents the clusters in our data, but we do not observe

likelihood:

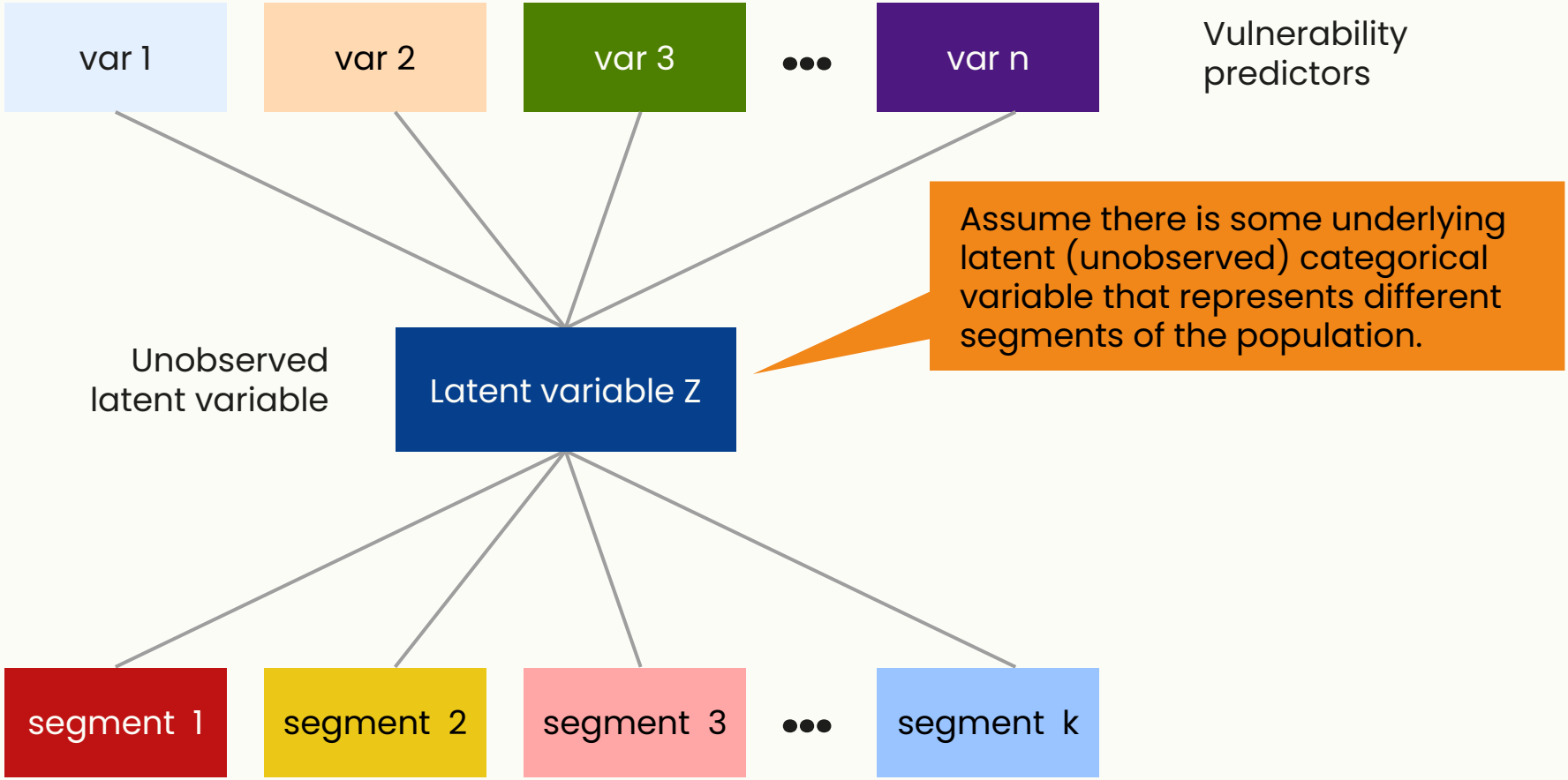
the model we are using to describe the distribution of the latent variable, and is a function of the variables we feed into our model

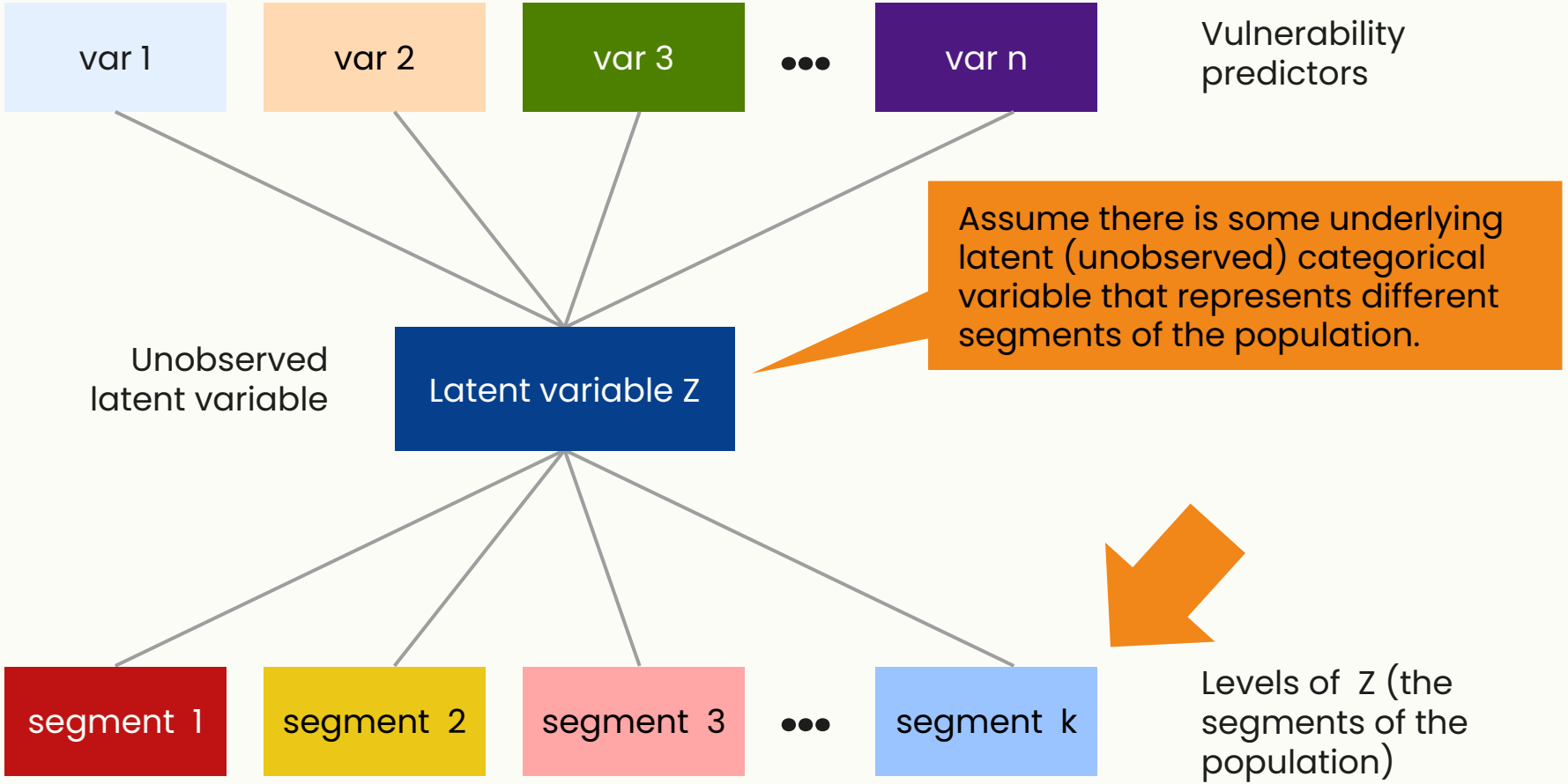
posterior probability:

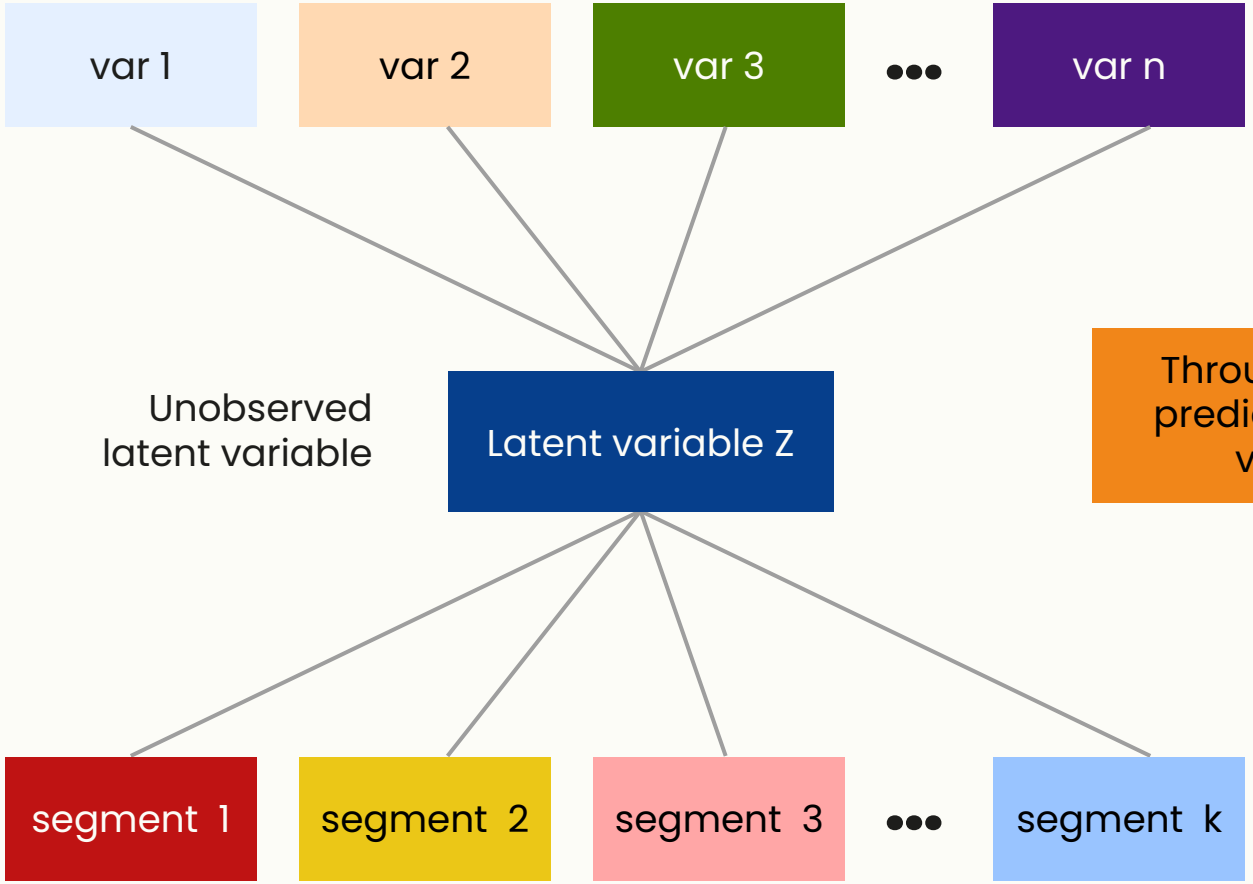
a predicted probability of belonging to a segment, given some set of observed variables (estimated by finding the **maximum likelihood estimate [MLE]** or **global maximum likelihood** of the model using a statistical algorithm)

Bayesian information criterion (BIC):

a statistic of model performance, which is a function of the likelihood. Lower values indicate a better model


var 1
var 2
var 3
var n
Vulnerability predictors
Assume there is some underlying latent (unobserved) categorical variable that represents different segments of the population.
Unobserved latent variable
Latent variable Z
segment 1
segment 2
segment 3
segment k


var 1
var 2
var 3
var n
Vulnerability predictors
Assume there is some underlying latent (unobserved) categorical variable that represents different segments of the population.
Unobserved latent variable
Latent variable Z
segment 1
segment 2
segment 3
segment k
Levels of Z (the segments of the population)



Vulnerability predictors

Through LCA use the vulnerability predictors to we predict the latent variable Z (i.e. segments)

Levels of Z (the segments of the population)

How does LCA work? (Single LCA model)

Assume a fixed number of segments (k) exist in the dataset.

1

Typically start with k =2 → separate models assume more segments.

How does LCA work? (Single LCA model)

Assume a fixed number of segments (k) exist in the dataset.

1

2

Specify the underlying model for predicting the latent class.

Model will be defined by the variables you are clustering on and R package used.

Pathways process uses the poLCA package, which has an appropriate underlying model for categorical variables (multinomial)

How does LCA work? (Single LCA model)

Assume a fixed number of segments (k) exist in the dataset.

1

2

Specify the underlying model for predicting the latent class.

For each observation in the dataset, estimate the posterior probability that the observation belong to each segment.

3

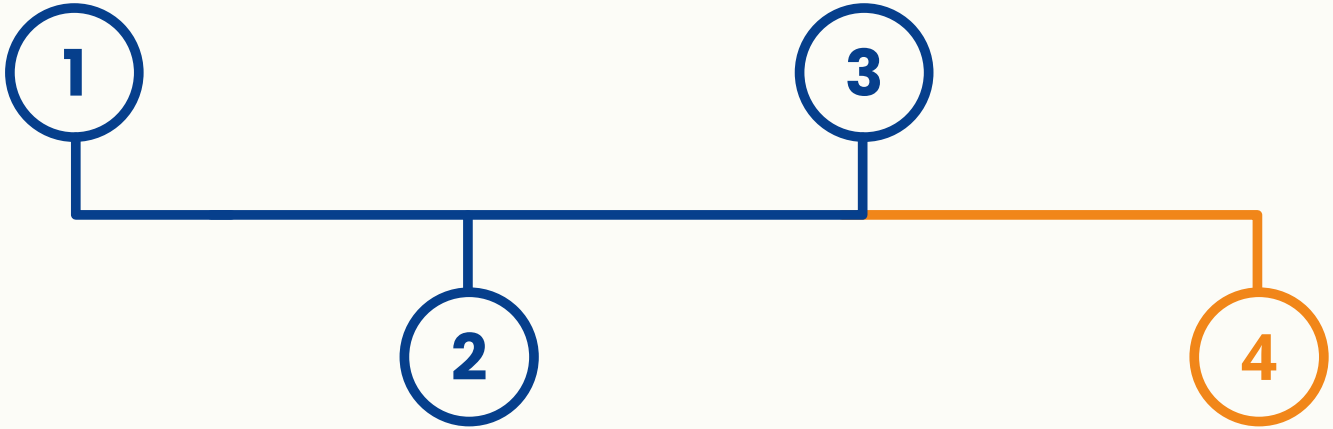
Estimate the posterior probability that the observation belong to each segment for each observation in the dataset.

These probabilities depend on the variables fed into the model.

The posterior probabilities are estimated by maximizing the likelihood using the EM algorithm.

How does LCA work? (Single LCA model)

1. Assume a fixed number of segments (k) exist in the dataset.
2. Specify the underlying model for predicting the latent class.
3. For each observation in the dataset, estimate the posterior probability that the observation belong to each segment.
4. Using the posterior probabilities, assign each observation to the segment for which their posterior probability is largest.



Example

The number of segments we assume to exist in our dataset (k=2)

Obs	Head of household	Working	Earns more than partner	Prob. Of being in segment 1	Prob. Of being in segment 2	Assigned Segment
1	Male	No	No			
2	Male	No	No			
3	Male	Yes	No			
4	Female	No	No			
5	Female	Yes	No			
6	Female	Yes	Yes			

Example

The vulnerabilities that predict segment membership

Obs	Head of household	Working	Earns more than partner	Prob. Of being in segment 1	Prob. Of being in segment 2	Assigned Segment
1	Male	No	No			
2	Male	No	No			
3	Male	Yes	No			
4	Female	No	No			
5	Female	Yes	No			
6	Female	Yes	Yes			

Example

Obs	Head of household	Working	Earns more than partner	Prob. Of being in segment 1	Prob. Of being in segment 2	Assigned Segment
1	Male	No	No	0.9	0.1	
2	Male	No	No	0.9	0.1	
3	Male	Yes	No	0.65	0.35	
4	Female	No	No	0.45	0.55	
5	Female	Yes	No	0.25	0.75	
6	Female	Yes	Yes	0.05	0.95	

Observations with similar variables will have similar probabilities estimated by the model.

Example

Obs	Head of household	Working	Earns more than partner	Prob. Of being in segment 1	Prob. Of being in segment 2	Assigned Segment
1	Male	No	No	0.9	0.1	1
2	Male	No	No	0.9	0.1	1
3	Male	Yes	No	0.65	0.35	1
4	Female	No	No	0.45	0.55	2
5	Female	Yes	No	0.25	0.75	2
6	Female	Yes	Yes	0.05	0.95	2

The probability that is largest determines the segment membership.

Questions of interest

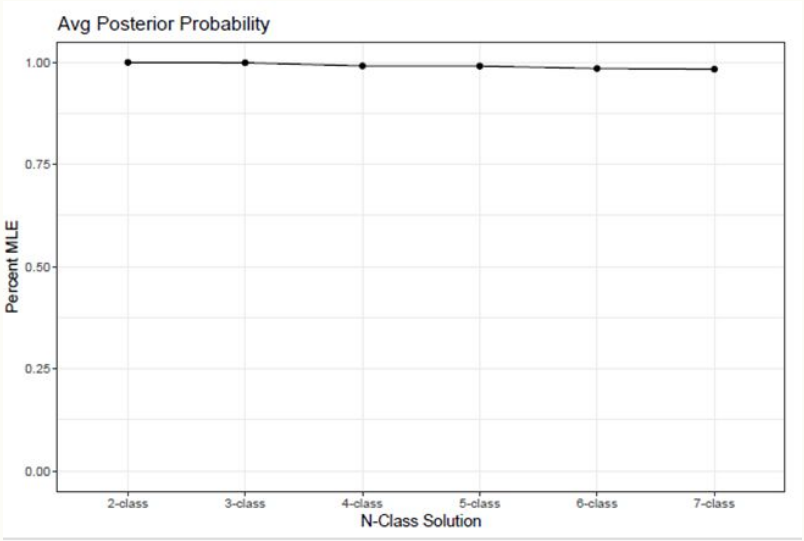
? How many segments exist in our dataset?

? What are the variables that drive separation between the segments and are they important to defining our segments?

- Must run several LCA models assuming different number of clusters (k) and compare the results across models (typically 10 models, with k = 2, 3, 4, …, 10).
- 'Final segmentation solution' is the best model in terms of statistical fit and meaningful (epidemiologic, social, etc.) differences between segments.

05

Technical Considerations of Running the Model



The posterior probabilities are estimated by maximizing the likelihood of the model, using a statistical algorithm.

Assumptions

1

A global maximum exists:
There is single best way to split up the dataset into clusters.

2

You've found the global maximum:
The algorithm successfully ran and found the best solution. It did not get 'stuck' at a subpar solution (e.g., a local maximum).

Reasons these assumptions might not be met

1

A global maximum exists:
There is single best way to split up the dataset into clusters.

Not met because splitting the data based on random noise, and not actual patterns in the data (especially when considering large number of clusters)

2

You've found the global maximum:
The algorithm successfully ran and found the best solution. It did not get 'stuck' at a subpar solution (e.g., a local maximum).

Not met because if the first isn't true **OR** if you just got unlucky in running the algorithm

Reasons these assumptions might not be met

Re-run the model a few times to see if you are consistently finding the same maximum likelihood (built in function in R) and check diagnostic plots.

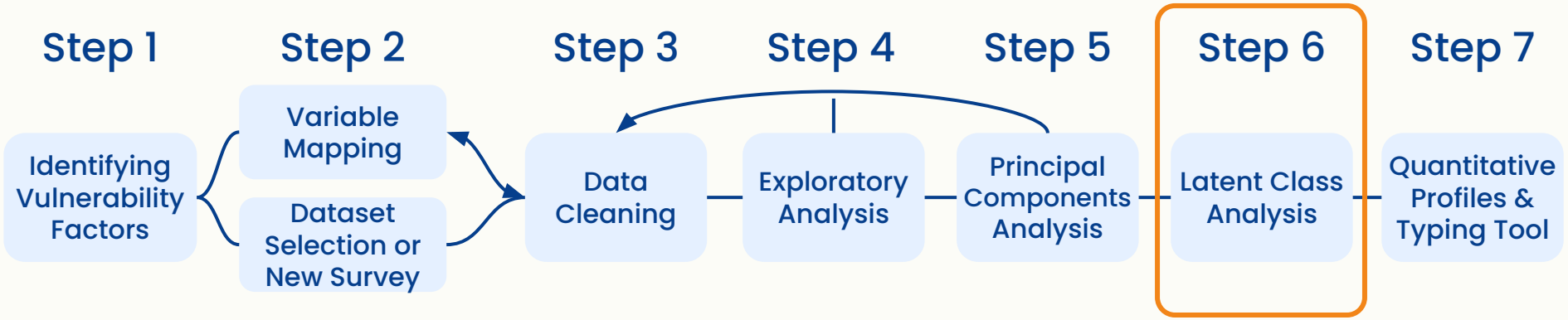
A global maximum exists:
There is single best way to split up the dataset into clusters.

Not met because splitting the data based on random noise, and not actual patterns in the data (especially when considering large number of clusters)

You've found the global maximum:
The algorithm successfully ran and found the best solution. It did not get 'stuck' at a subpar solution (e.g., a local maximum).

Not met because if the first isn't true **OR** if you just got unlucky in running the algorithm

Questions of interest



? How many segments exist in our dataset?

? What are the variables that drive separation between the segments and are they important to defining our segments?

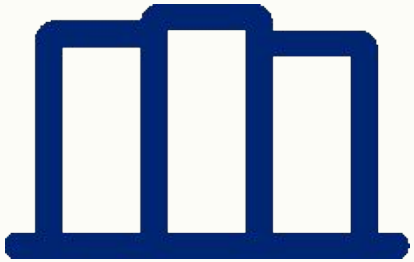
06

Tools for Selecting the Best Model

Tools for selecting the best model

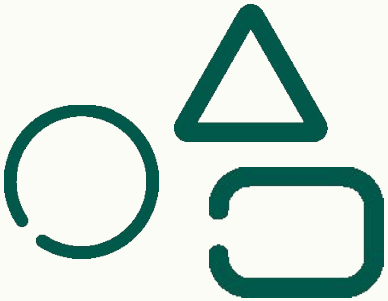
Diagnostics & fit

Used to eliminate models/solutions that are unlikely to capture reliable patterns



Stability

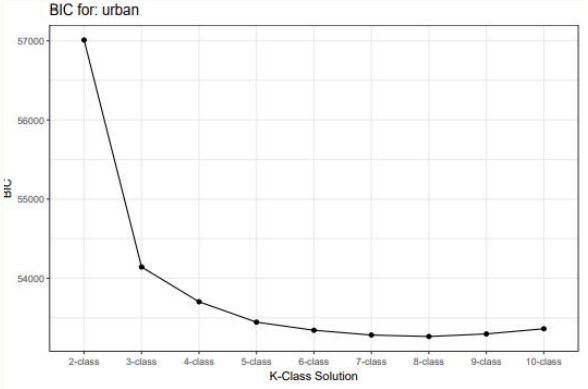
Used to understand which segments of the population are consistently found across models or if they are just ‘noisy’ groupings of observations

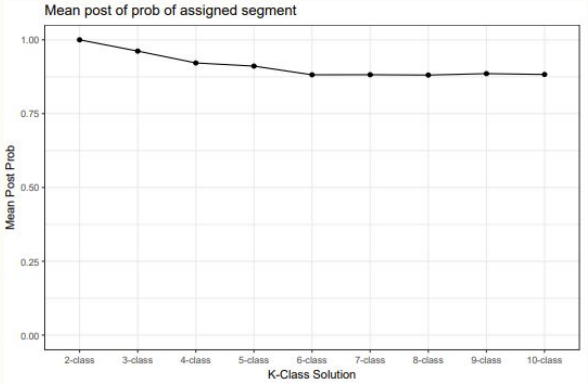


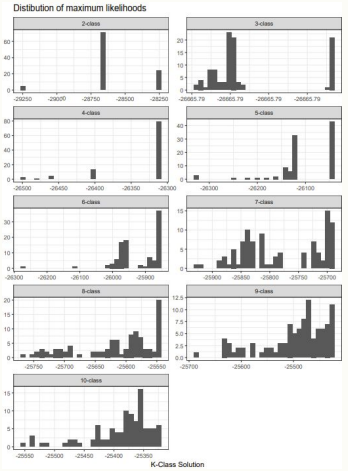
Meaningful variation

Used to understand if adding a segment captures meaningful differences within our dataset

Diagnostics & fit







BIC plot

Plot of the BIC statistics across models. Lower BIC better, and more complex models penalized.

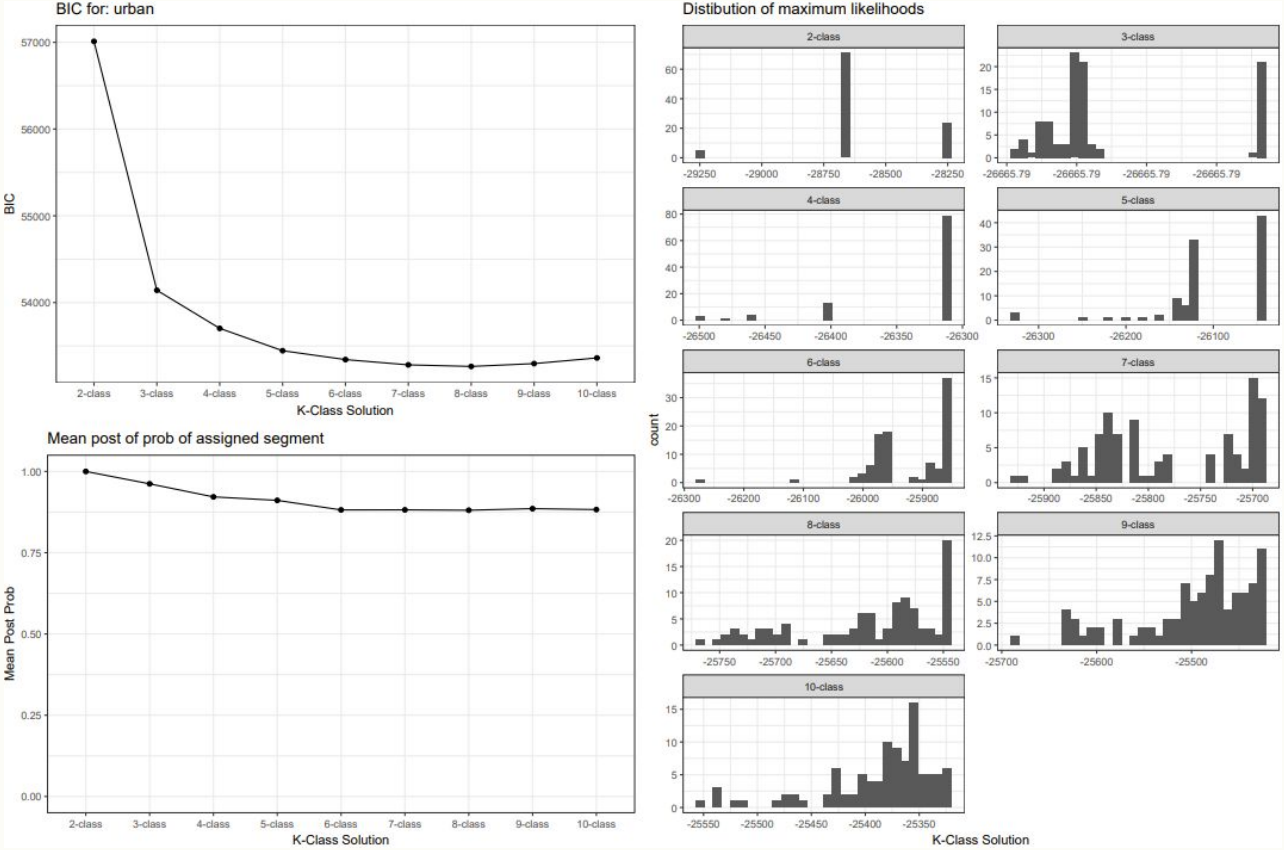
Avg posterior probability

Across all observations, for the posterior probability of the segment they were assigned to. If not large, could signal that solution is not much better than a random split.

Distribution of ML

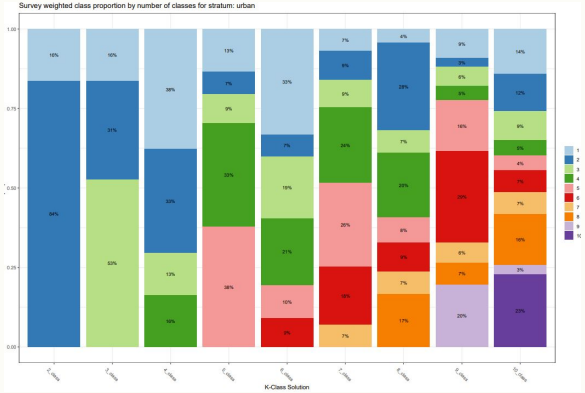
For a single model repeatedly run, what is the distribution of maximum likelihoods? Is the global maximum clear?

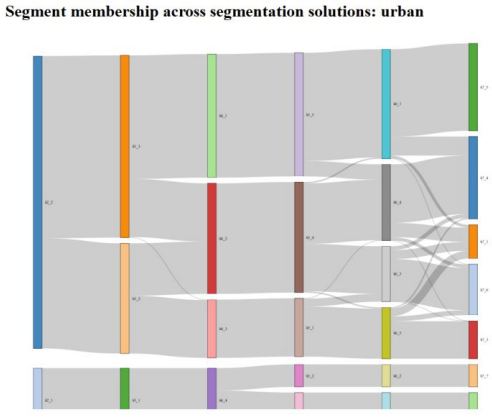
Diagnostics & fit



- Can we eliminate models that are comparatively worse than other, in terms of fit statistics?
- Do any models seem unstable?
- Do any models indicate random splits?

Stability





Segment size plot

For each model, the proportion of observations within each segment.

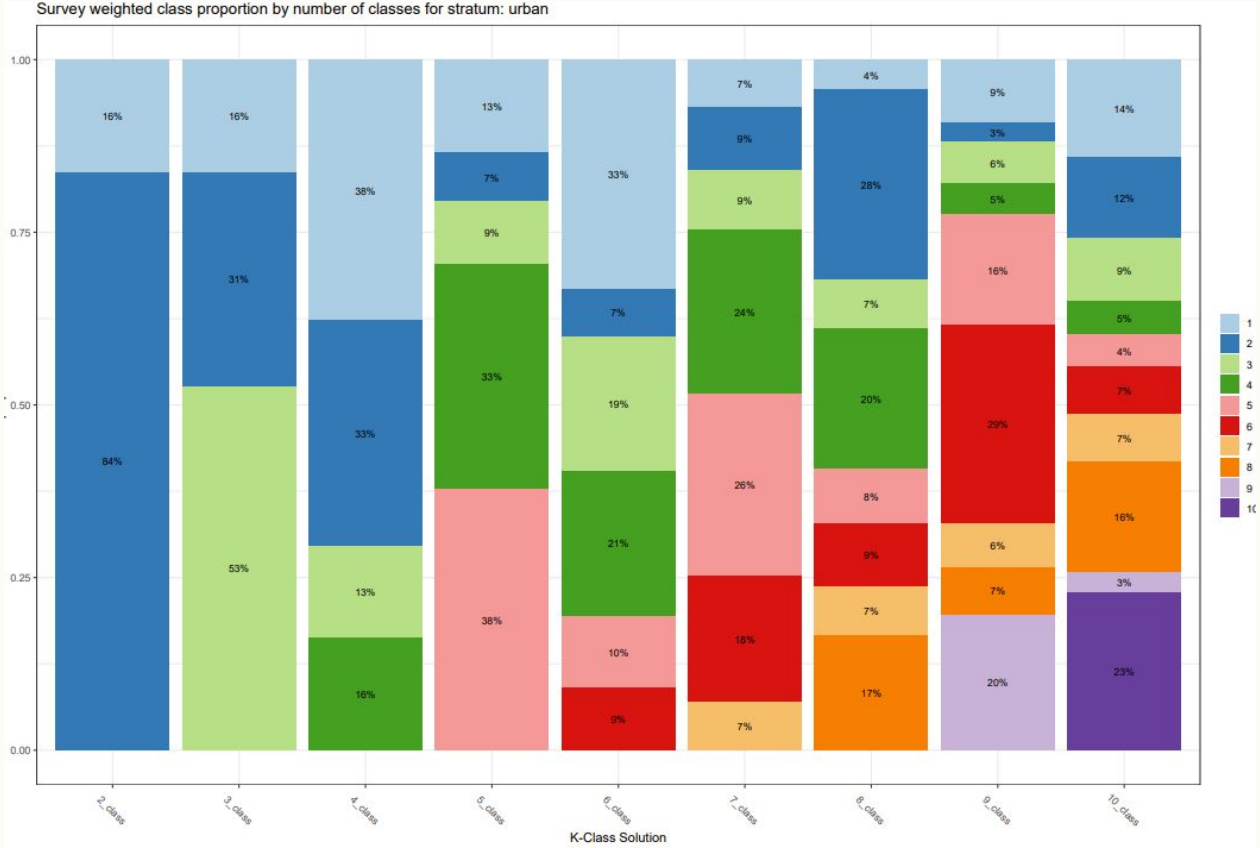
Flags any unusually small segments, that might be unstable or non-representative.

Sankey diagram

Tracking across models, segment membership of each observation.

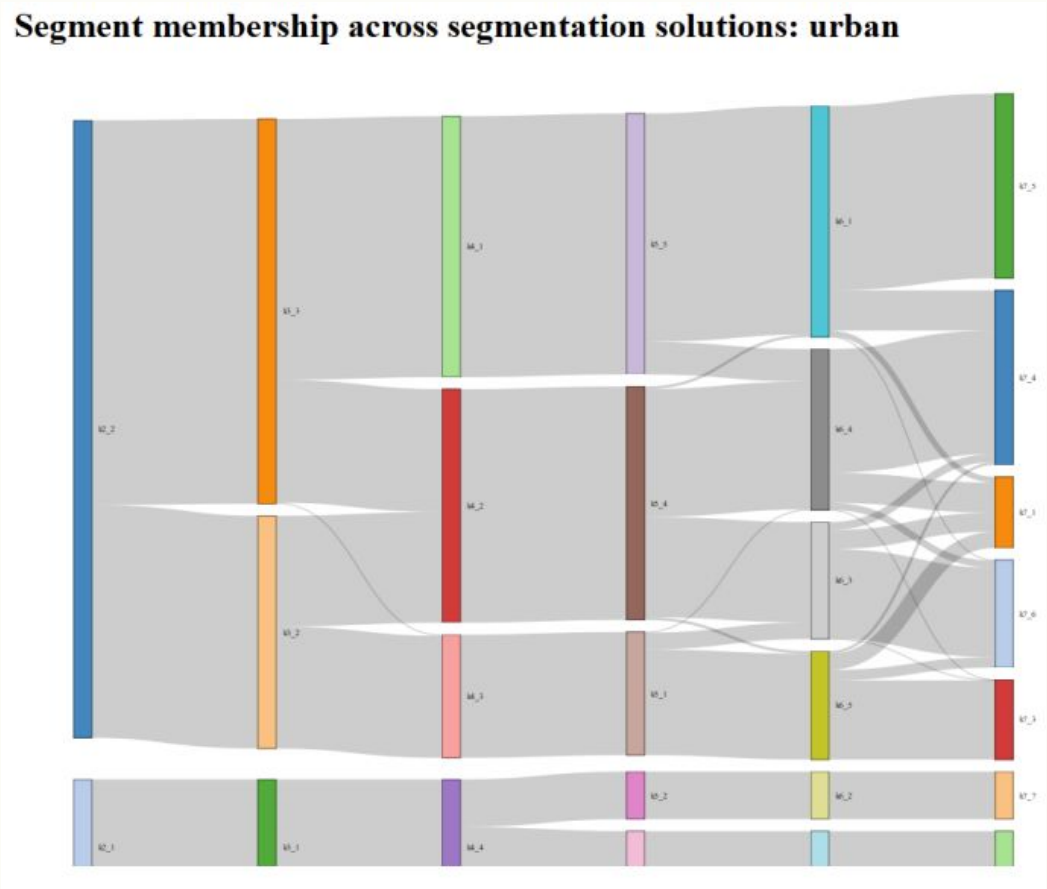
Flags any models where segments break apart easily, and observations within these segments may not have much in common.

Stability visualizations



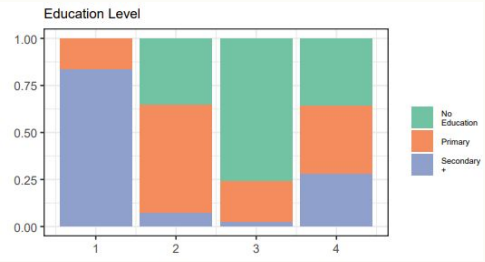
- Do any models show very small or large clusters?
  - Too small could be noise
  - Too large could mask important difference
- General guideline: segments should be around 10-30% of the population

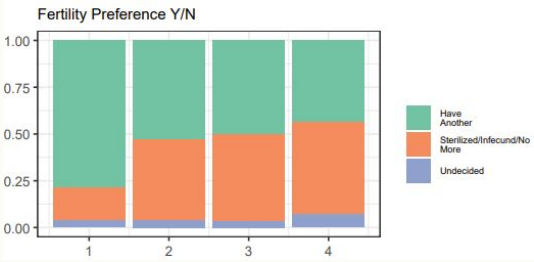
Stability visualizations

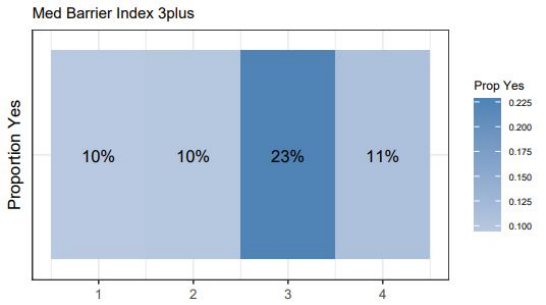


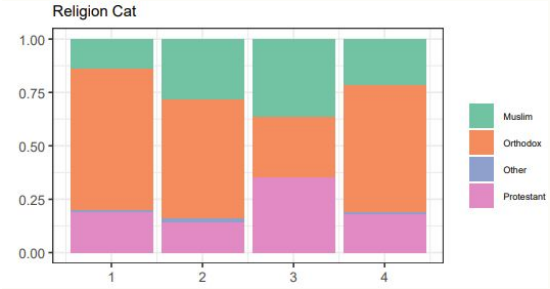
- Is any segment membership consistent across classes?
- Are new segments made up of splits of a previous segment?
- Are new segments made up of small proportions of other segments?
  Do these observations have something meaningful in common?

Meaningful variation visualizations within the same model



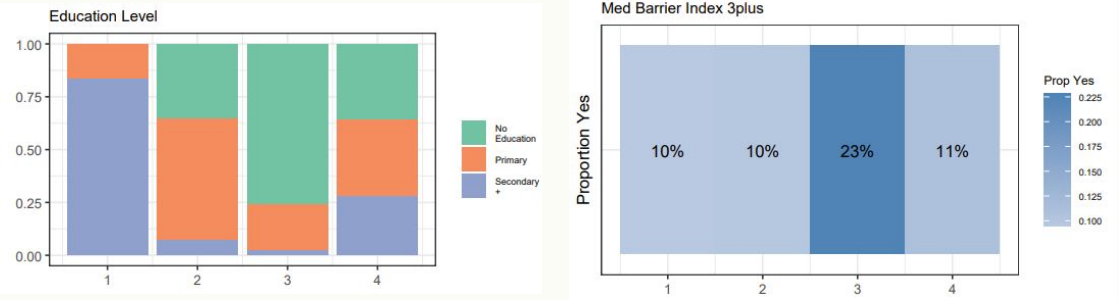


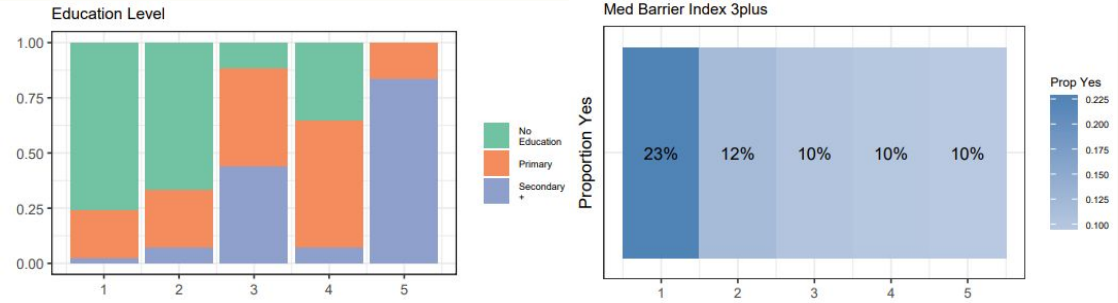




- What variables show large differences between segments?
  - Are the distributions different between segments?
- Which segments appear relatively more vulnerable than others?

Meaningful variation visualizations between models

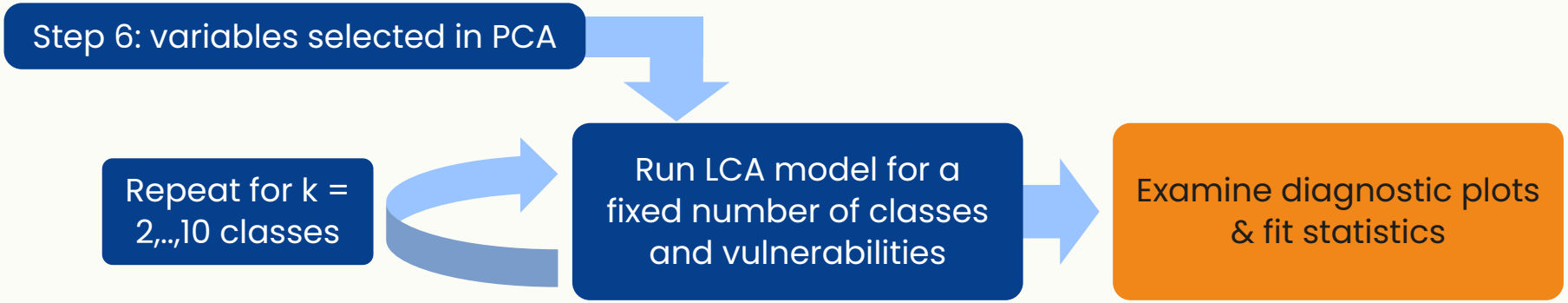




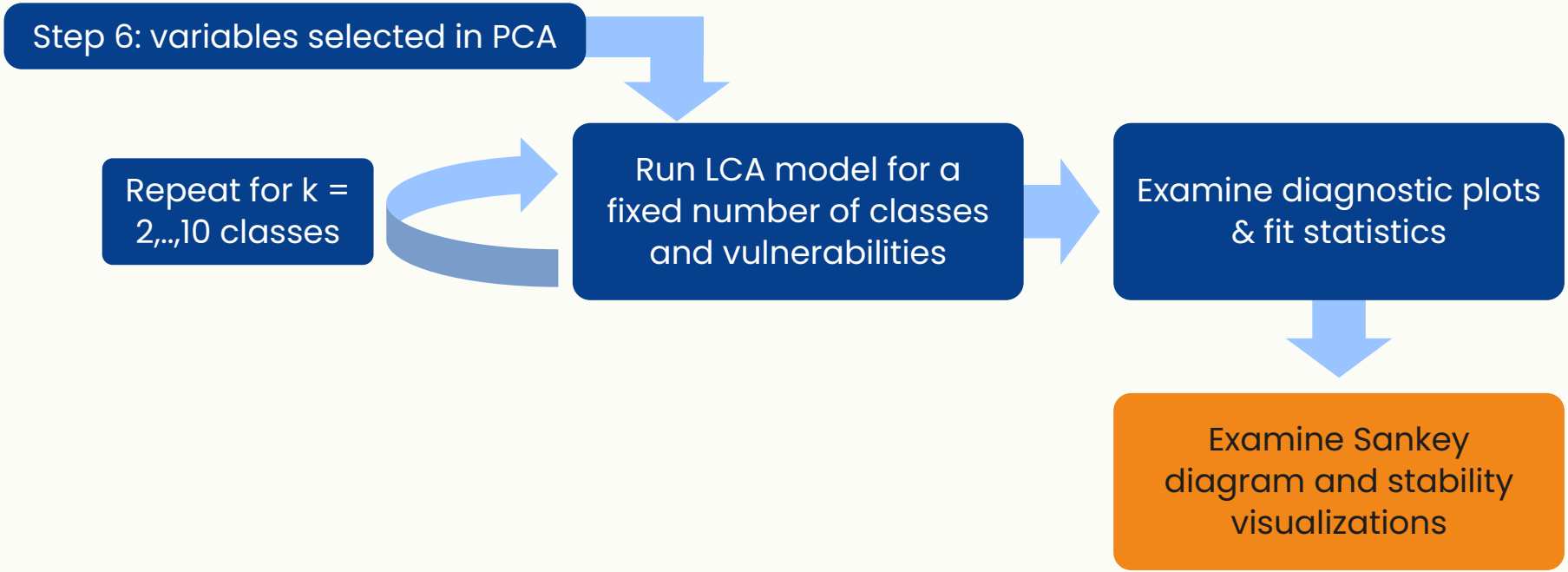
- Does increasing the number of assumed segments show a difference in a meaningful vulnerability?
- Note that segments are not directly comparable between models

07

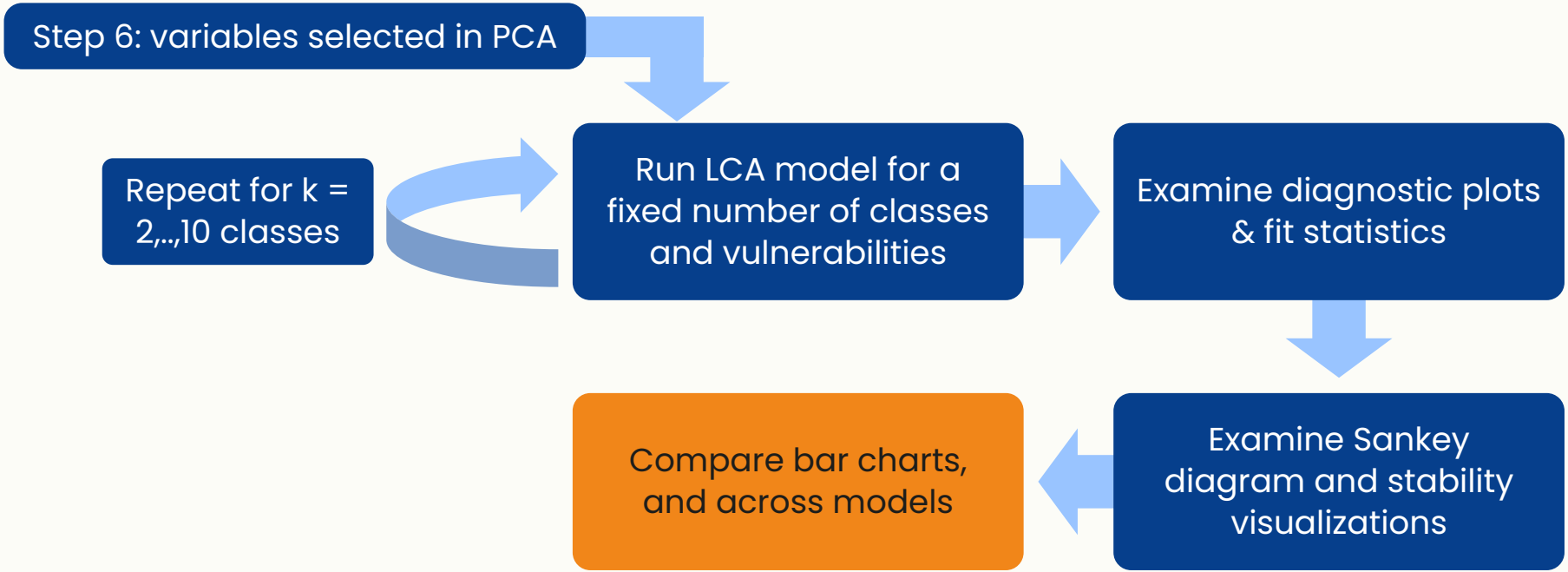
The LCA Process



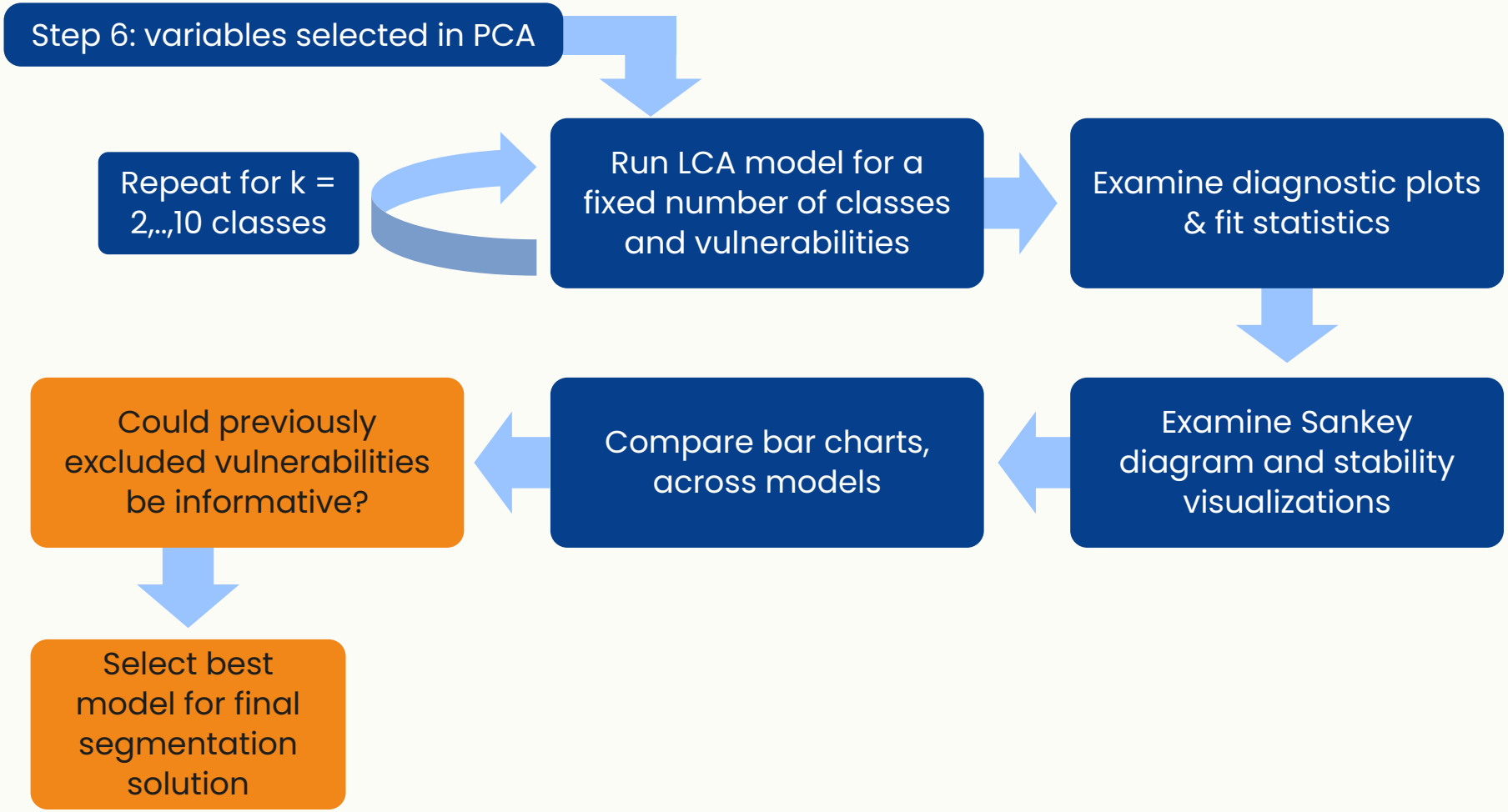
- Do any models seem to fit poorly?
- Can we eliminate any models based on fit statistics?



Does increasing the number of segments in the model decrease stability of the model?



- What variables differentiate segments?
- Are these variables meaningful?


Step 6: variables selected in PCA
Repeat for k = 2,..,10 classes
Run LCA model for a fixed number of classes and vulnerabilities
Examine diagnostic plots & fit statistics
Examine Sankey diagram and stability visualizations
Compare bar charts, across models
Could previously excluded vulnerabilities be informative?
Select best model for final segmentation solution

08

Group Activity

discussion

Which solution best balances segment sample size, segment algorithm statistics, and captures vulnerability?

homework

Let’s take a look at the R code and results pack for LCA.

https://uwashington.qualtrics.com/jfe/form/SV_eM1hAA6ruxWyzHw

Session survey