Cluster analysis

Community health workers in a school, Rabai county, Kenya, September 2021

WASHA, Takwimu, UKZN, 25 August 2023

Till Bärnighausen, Heidelberg Institute of Global Health, University Hospital and Medical Faculty, Heidelberg University

PCA is easy to program and easy to implement and artful to interpret

2 TACTICS OF PCA

- Dimension reduction: How many principal components should we use?
- Discovery: What do the principal components mean?

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

Principal component analysis uses two foundations of statistics, which come with assumptions

ASSUMPTIONS OF CANONICAL PCA

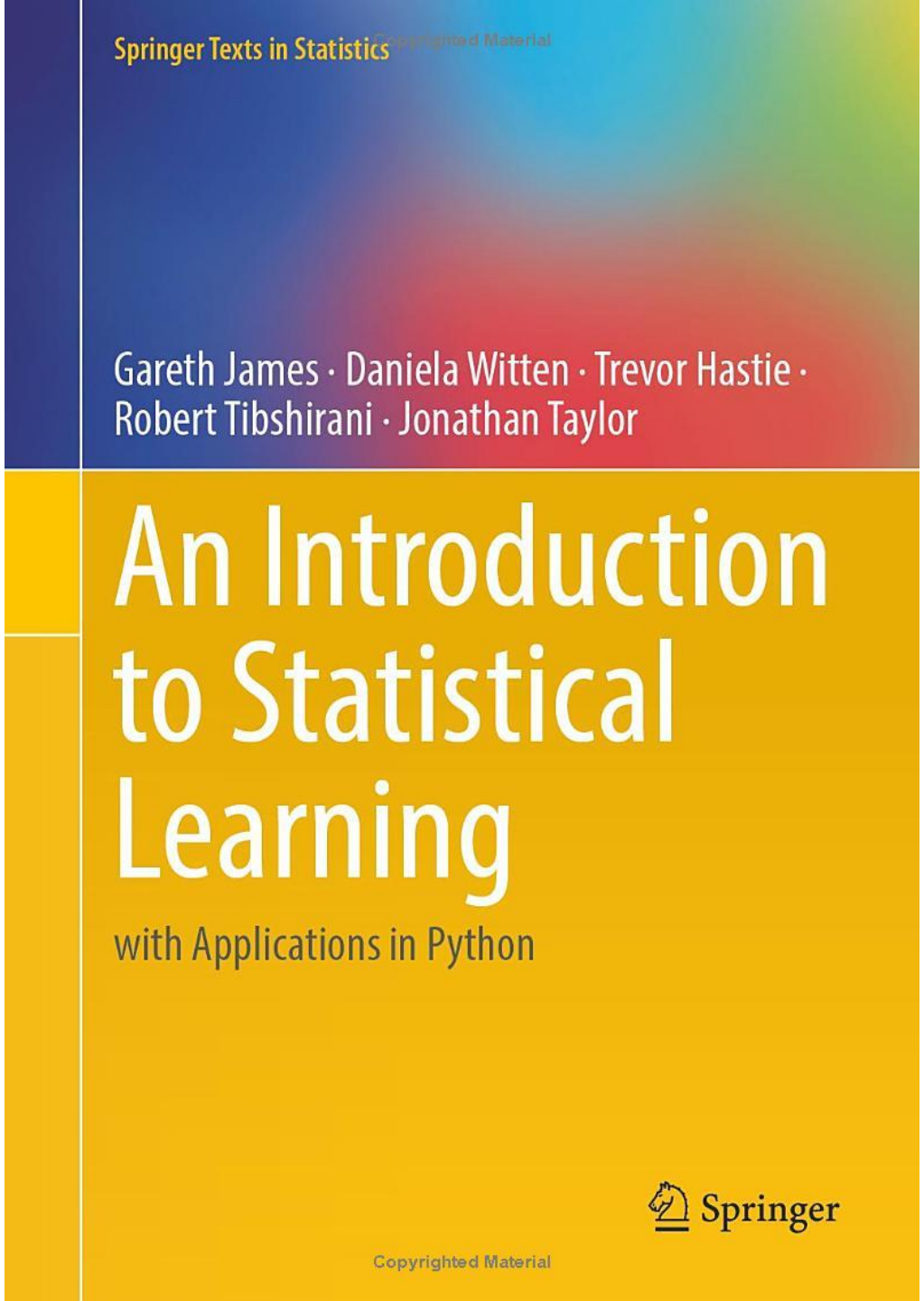
1. Linearity
2. Mean and variance sufficient statistics
3. Large variance mean important dynamics (high SNR)
4. Principal components are orthogonal

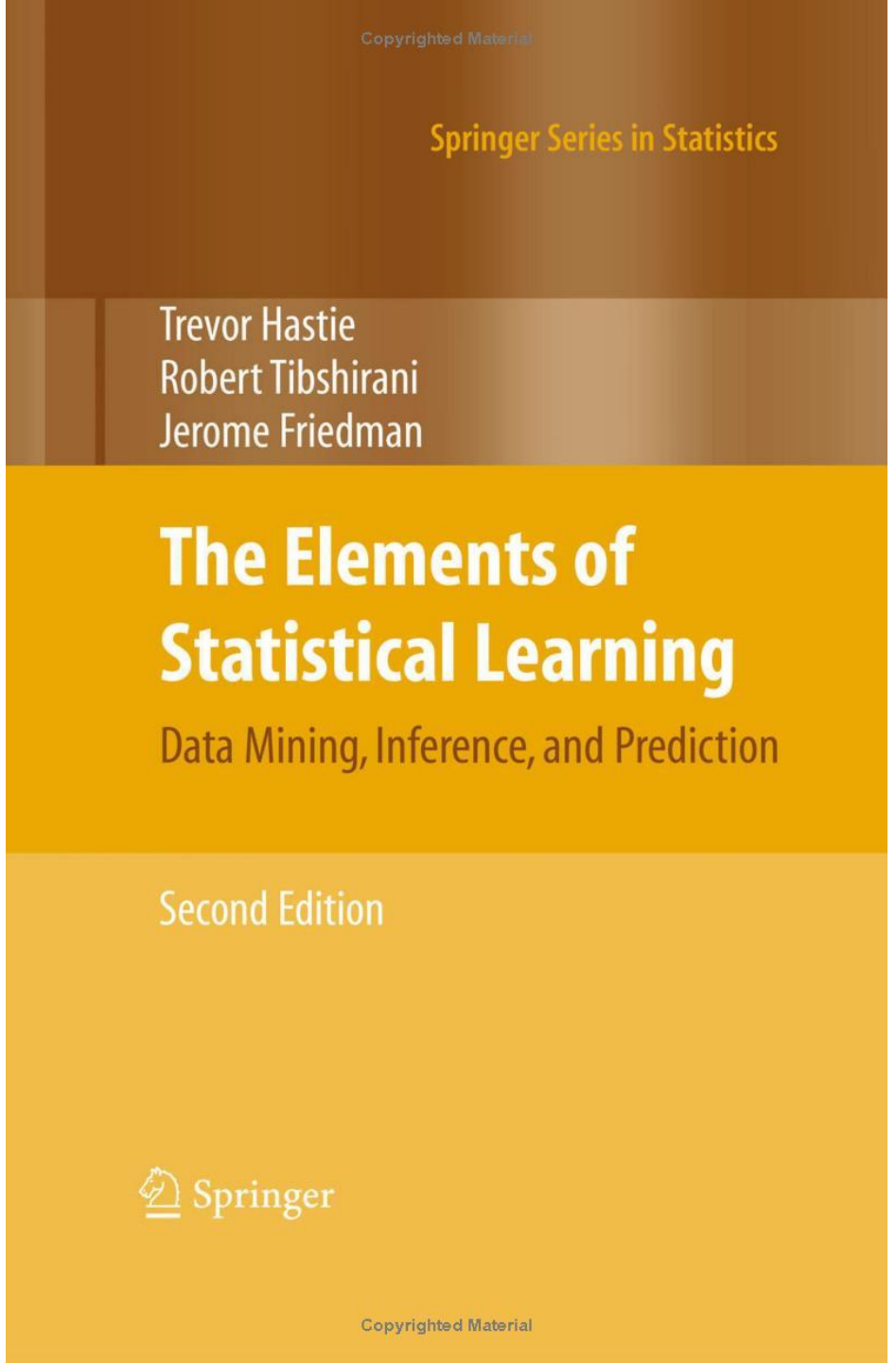
Quantitative analysis serve many important functions in health systems research

4 FUNCTIONS

1. Description
2. **Discovery – unsupervised machine learning**
    - Dimension reduction: low-dimensional representation of the observations that explain a large fraction of the total variance
    - **Cluster analysis:** homogeneous subgroups among the observations
3. Prediction
4. Causation

HEIDELBERG INSTITUTE OF GLOBAL HEALTH


Springer Texts in Statistics
Copyrighted Material
Gareth James · Daniela Witten · Trevor Hastie ·
Robert Tibshirani · Jonathan Taylor
An Introduction to Statistical Learning
with Applications in Python
Springer
Copyrighted Material


Copyrighted Material
Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer
Copyrighted Material

K means rests on simple mathematical foundations

MATHEMATICAL INTUITION (I)

- Minimize the total within-cluster variation, summed over all *K* clusters

$$min_{C_1, C_2, \ldots C_K} \sum_{k=1}^{K} W(C_k)$$

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

K means rests on simple mathematical foundations

MATHEMATICAL INTUITION (II)

- Minimize the sum of all the pairwise squared Euclidian distances between the observations in the k^{th} cluster, divided by the total number of observations in the k^{th} cluster – the summed over all K clusters

$$min_{C_1, C_2, \ldots C_K} \sum_{k=1}^{K} \frac{1}{C_k} W \sum_{i,i' \in C_k} \sum_{j=1}^{p} \left(x_{ij} - x_{i'j} \right)^2$$

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

What PCA is for dimension reduction, k means is for cluster analysis

PROCEDURAL INTUITION

- Two properties:
  - Each observation belongs to at least one of the k clusters
  - No observation belongs to more than one cluster
- Steps:
  1. Choose *K*
  2. Randomly assign a number from 1 to K to each observation
  3. Iterate until the assignments stop changing:
     a. For each of the *K* clusters, compute the cluster centroid (vector of the *p* feature means)
     b. Assign each observation to the cluster whose centroid is closest (using Euclidian distance)

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

How to choose the hyperparameter *K*?

INTUITION AND METRICS EXAMPLES (I)

- We have to quantify the goodness of a clustering
- All metrics capture some version of one or both of two high-level ‘goodness’ characteristics
  - Within: **cohesion** or **compactness** of clusters
  - Across: **separation** of clusters

$$\frac{(\alpha * separation)}{(\beta * compactness)}$$

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

How to choose the hyperparameter *K*?

INTUITION AND METRICS EXAMPLES (II)

- **Silhouette coefficient** (SC) – uses the mean intracluster distance (a) and the mean nearest-cluster distance (b)

$$\frac{(b-a)}{\max(a,b)}$$

- **Dunn index (DI)**

$$\frac{(\min separation = \min pairwise\ intercluster\ distance)}{(\max diameter = \max intracluster\ distance = \ compactness)}$$

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

What clusters of chronic disease dominate in India?

METHODS – FINGER EXERCISE (I)

1. Data source: National Family Health Survey – 639,661 women
2. Variables: biomarkers and anthropometric measures that indicate major chronic conditions – continuous variables
   - Height and weight
   - Waist and hip circumference
   - Hemoglobin
   - Random blood glucose
   - Systolic and diastolic blood pressure

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

What clusters of chronic disease dominate in India?

METHODS – FINGER EXERCISE (II)

3. Data preprocessing:
   - Observations removed because of implausible values (14 weight or height, 5 blood sugar, 11 systolic or diastolic blood pressure)
   - PCA: first five principal components using Kaiser's criterion – from 8 to 5 dimensions
4. Cluster analysis:
   - K means
   - *NbClust* R package for k hyperparameter optimization, using 26 cluster performance indices – majority vote

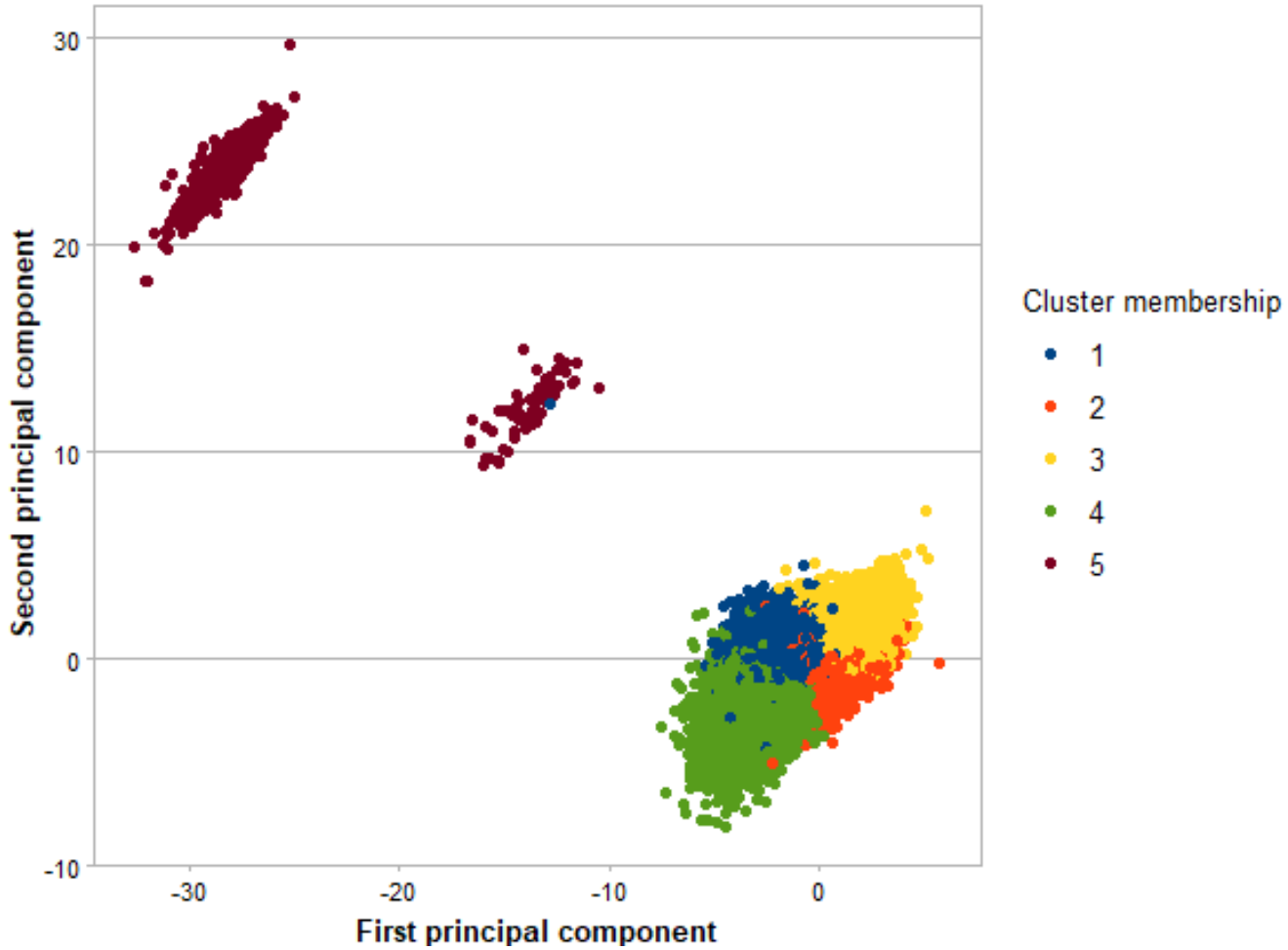
HEIDELBERG INSTITUTE OF GLOBAL HEALTH

5 clusters of chronic conditions emerge among women in India

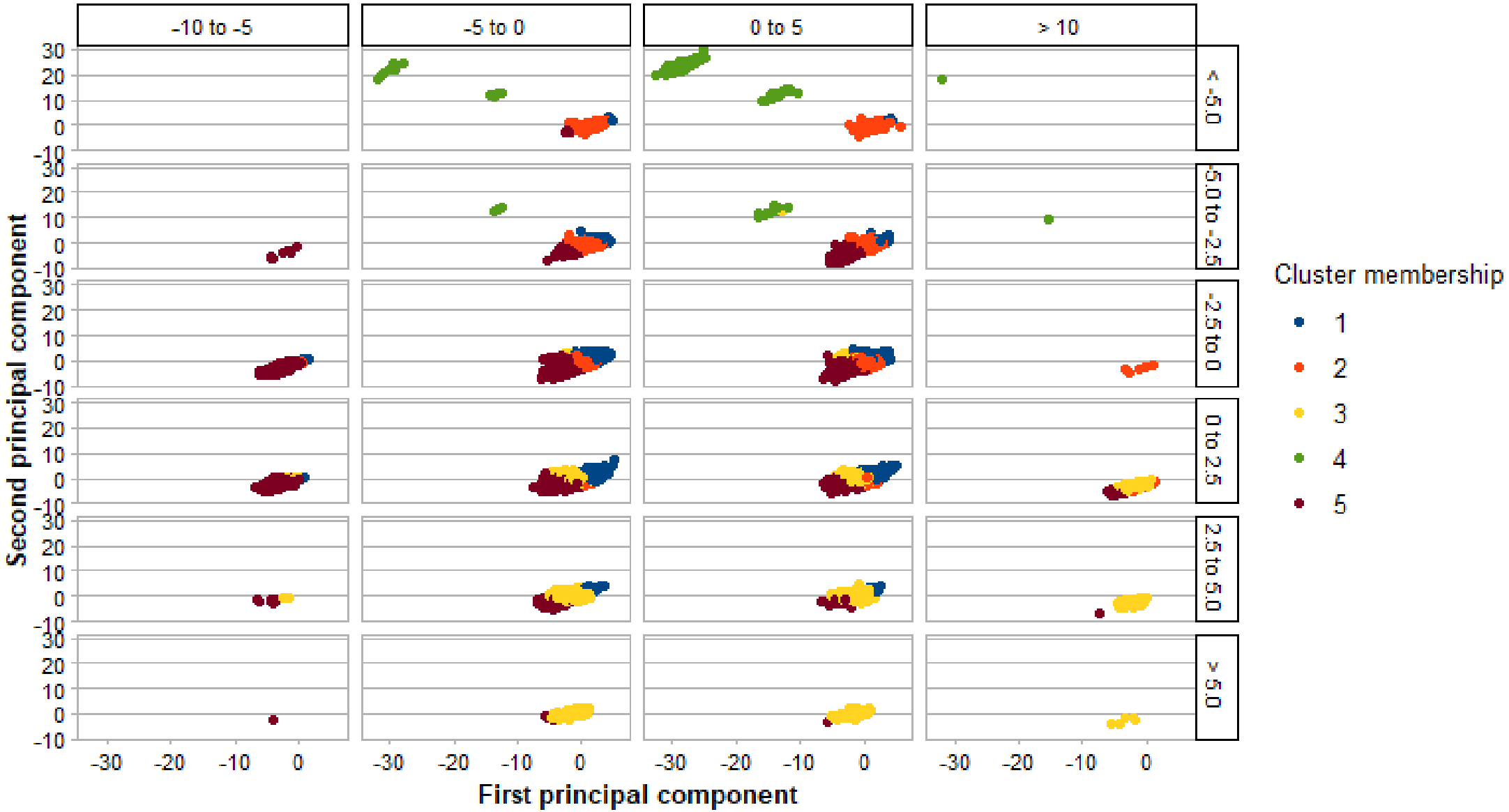
METHODS – FINGER EXERCISE (III)

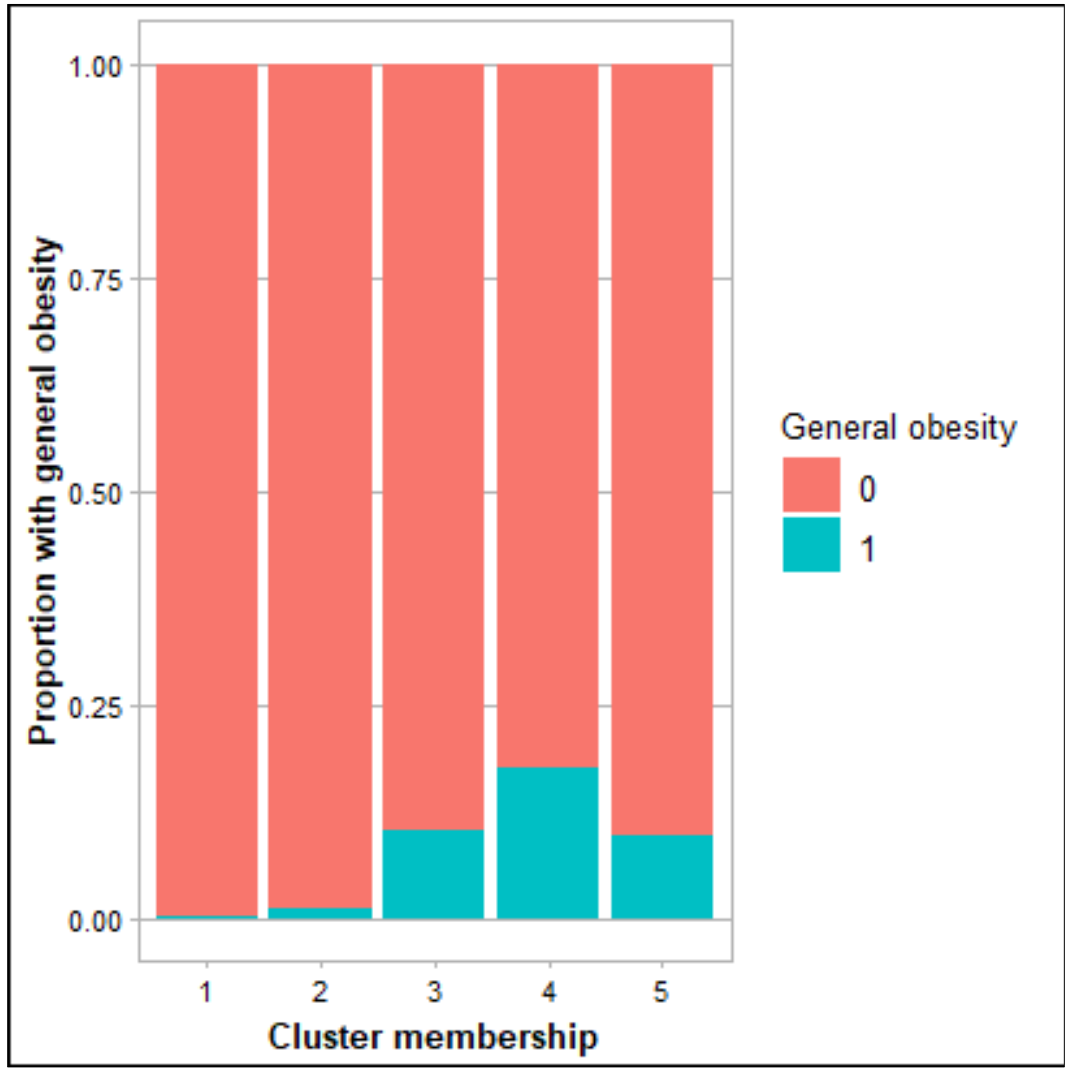
Index name	# clusters
KL	5
CH	5
Hartigan	4
CCC	5
Scott	5
Marriot	5
TrCovW	5
TraceW	5
Friedman	5
Rubin	5
Cindex	4
DB	5
Silhouette	2
Duda	2
PseudoT2	2
Beale	2
Ratkowsky	5
Ball	3
PtBiserial	5
Frey	1
McClain	2
Dunn	9
Hubert	0
Sdindex	5
Dindex	0
SDbw	10

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

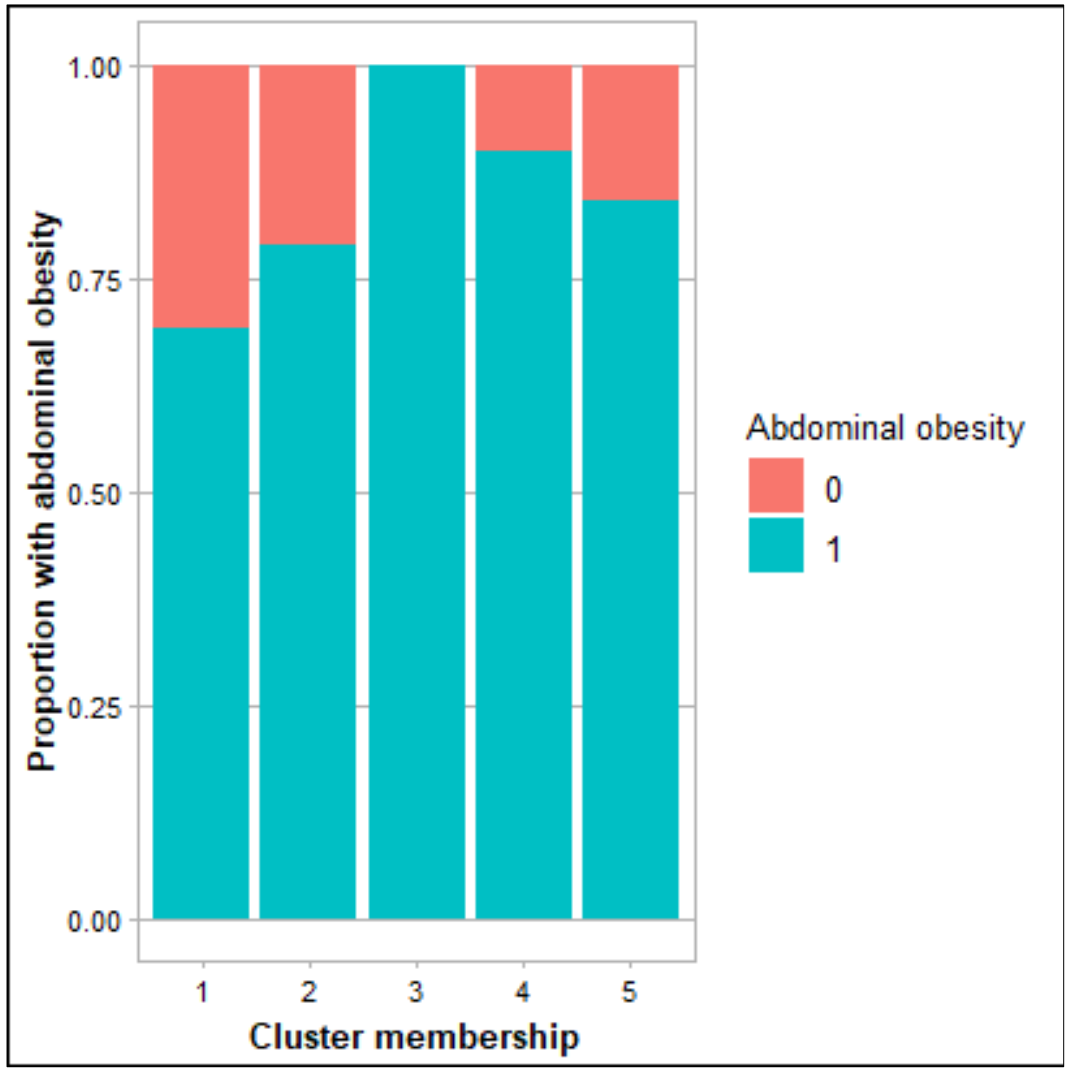


HEIDELBERG INSTITUTE OF GLOBAL HEALTH

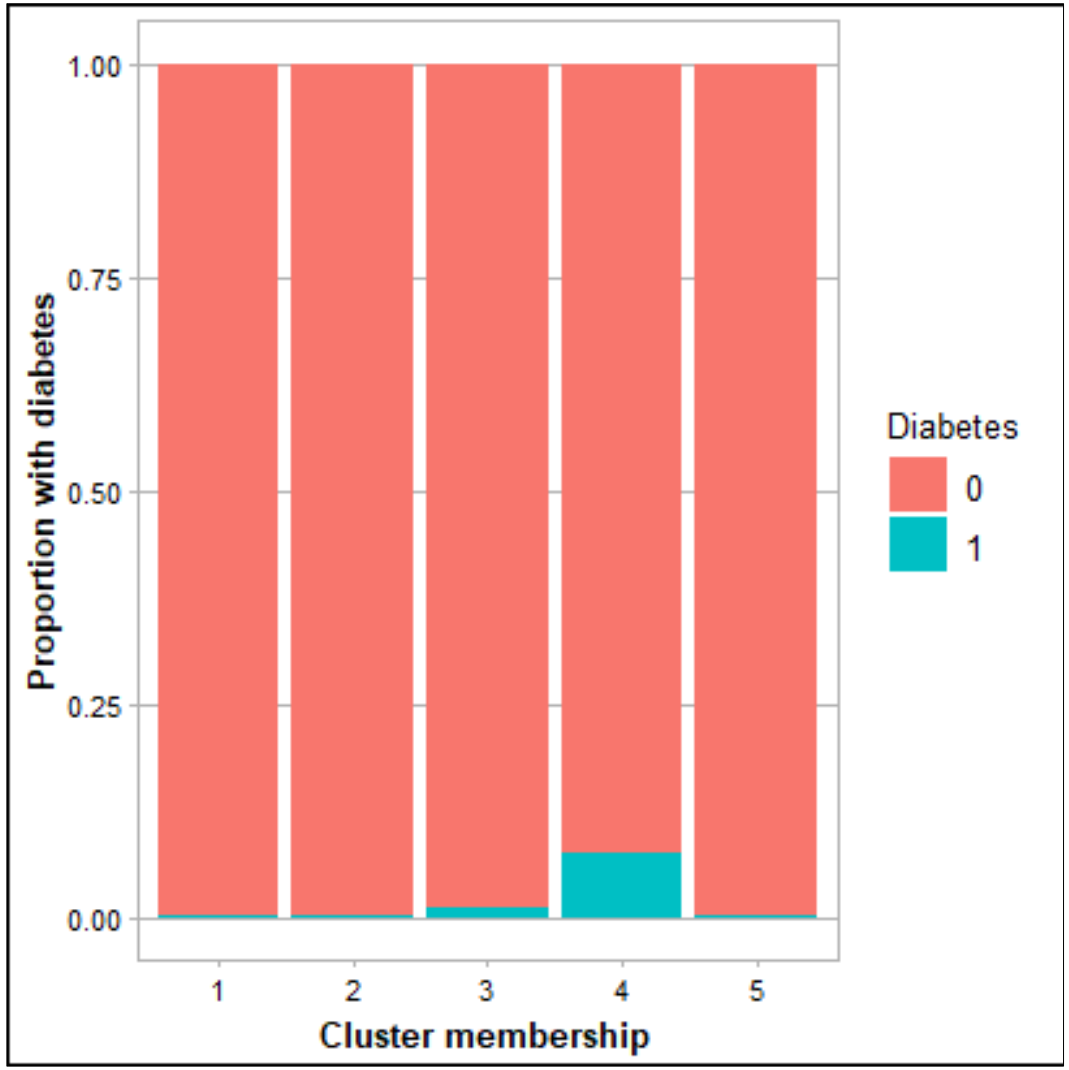

-10 to -5
-5 to 0
0 to 5
> 10
< -5.0
-5.0 to -2.5
-2.5 to 0
0 to 2.5
2.5 to 5.0
> 5.0
Second principal component
First principal component
Cluster membership
1
2
3
4
5



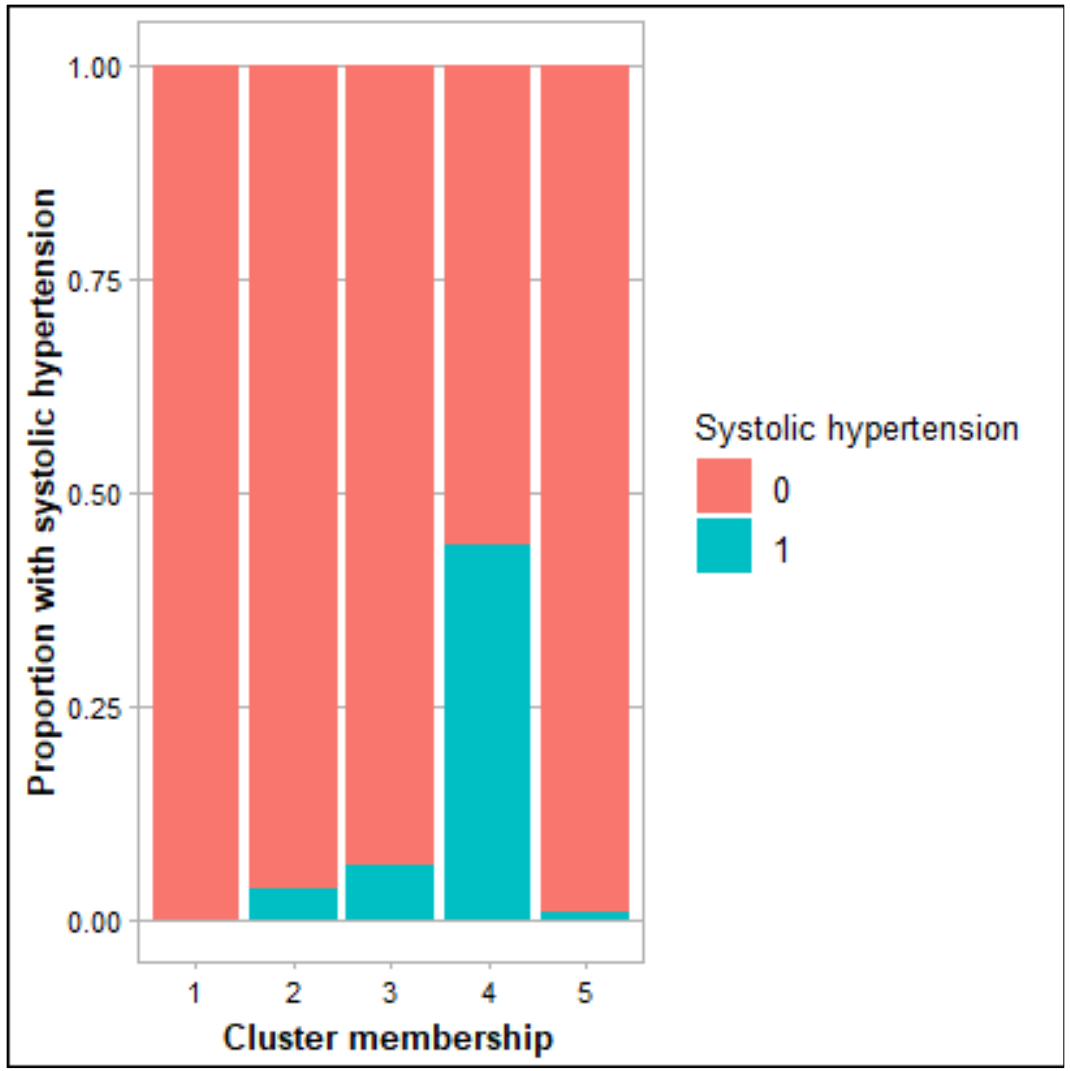
HEIDELBERG INSTITUTE OF GLOBAL HEALTH



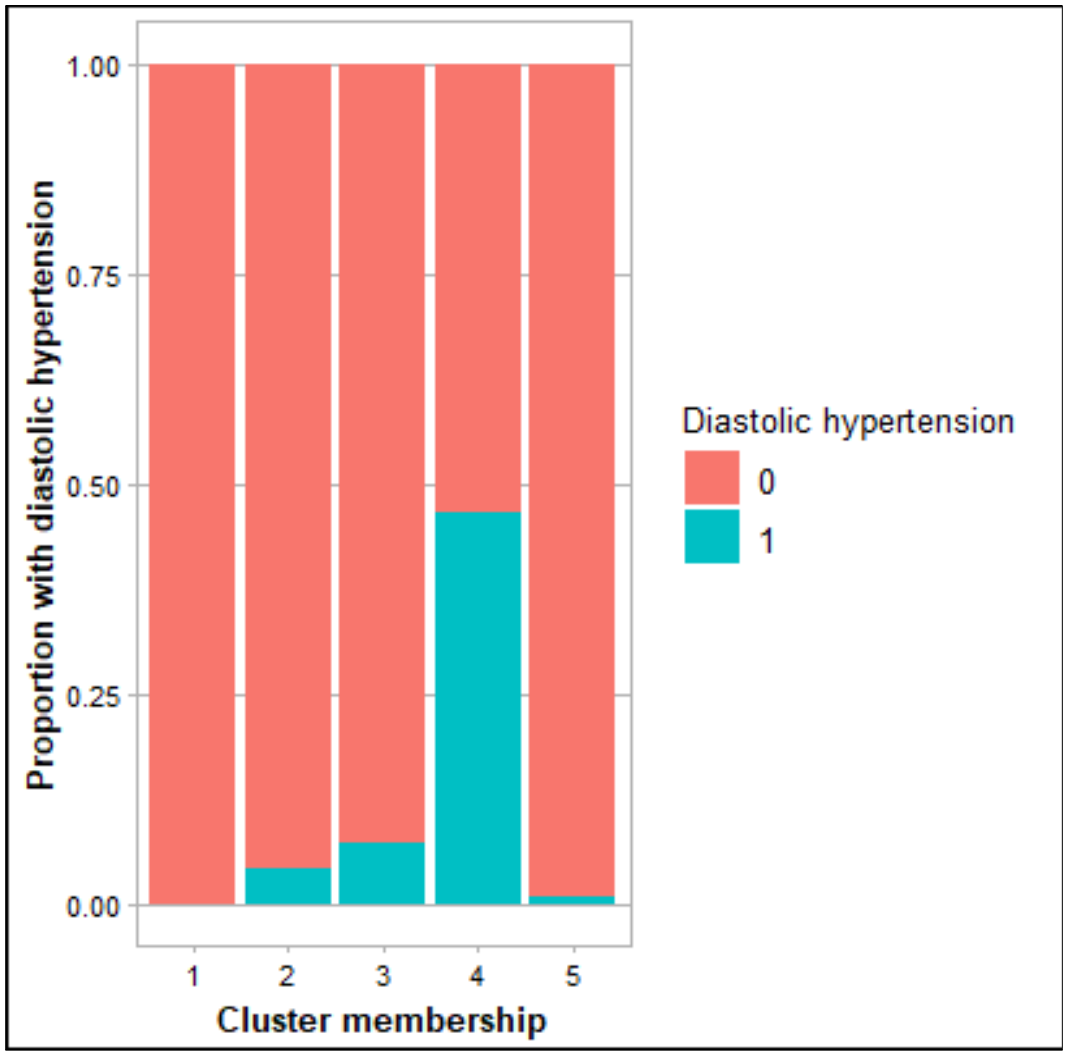
HEIDELBERG INSTITUTE OF GLOBAL HEALTH



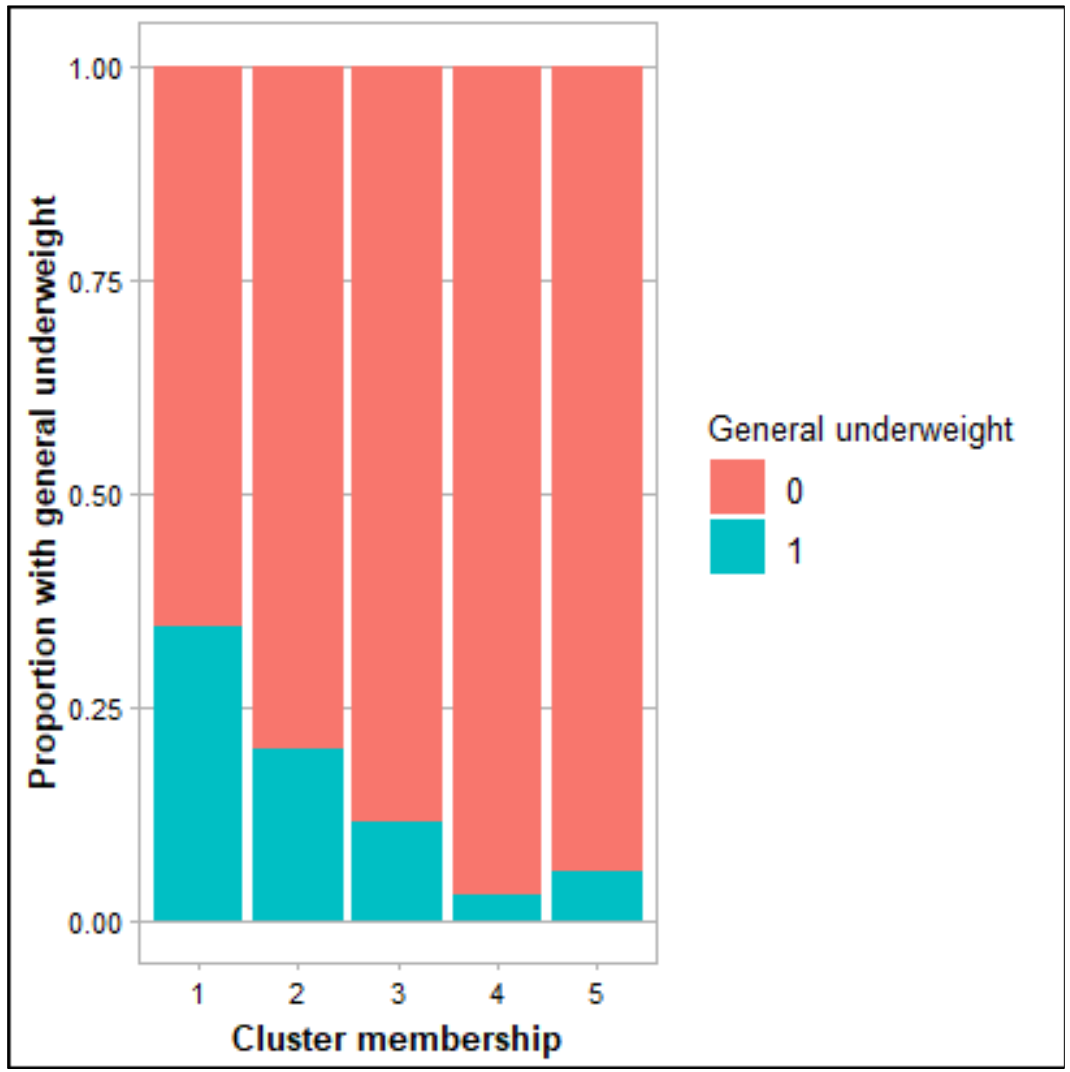
HEIDELBERG INSTITUTE OF GLOBAL HEALTH



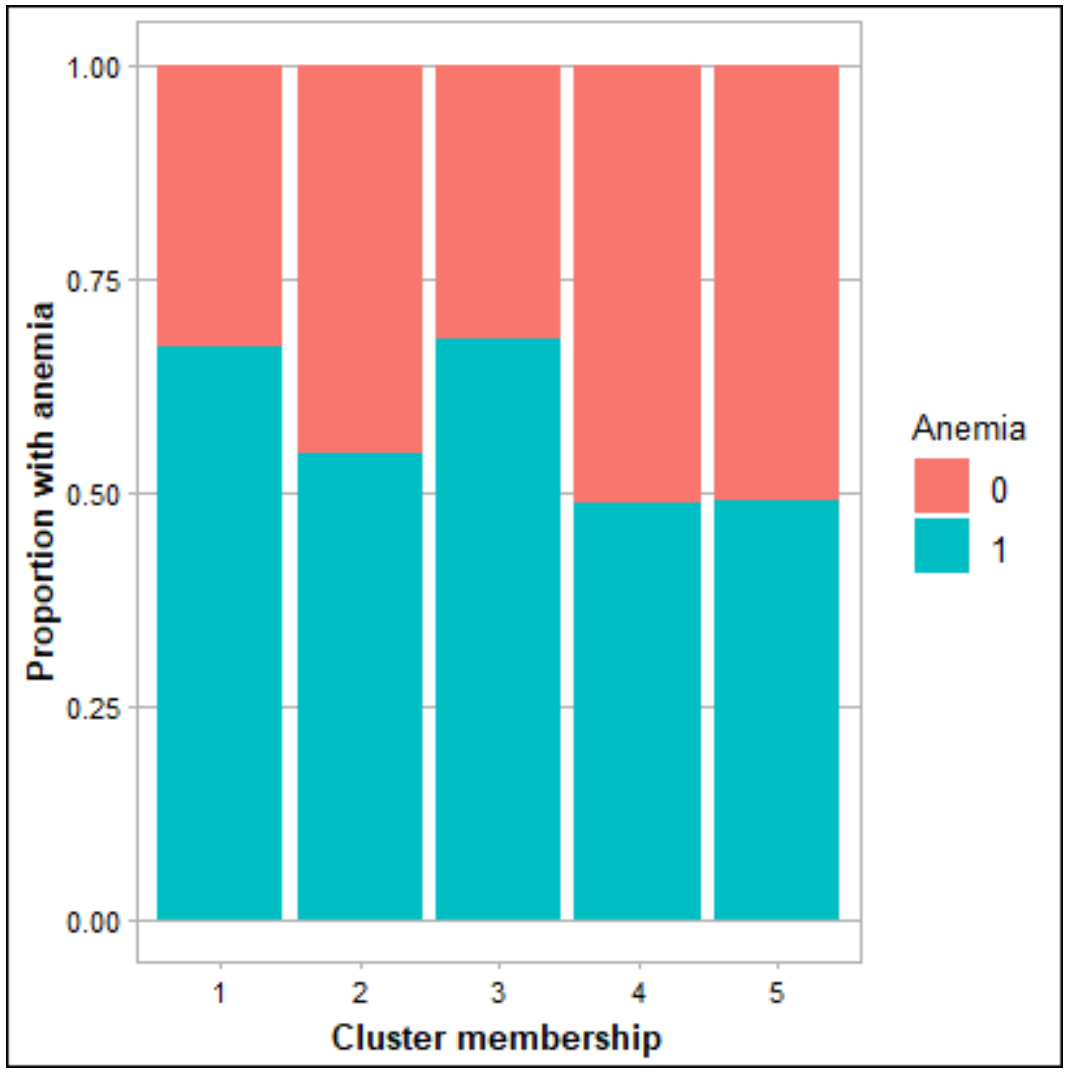
HEIDELBERG INSTITUTE OF GLOBAL HEALTH



HEIDELBERG INSTITUTE OF GLOBAL HEALTH



HEIDELBERG INSTITUTE OF GLOBAL HEALTH



HEIDELBERG INSTITUTE OF GLOBAL HEALTH

What factors determine cluster membership?

METHODS – FINGER EXERCISE (IV)

5. Regression:
   - Multinomial logistic regression, using the *nnet* R package
   - Dependent variable: cluster membership
   - Independent variables: socio-demographic
     - Age
     - Education
     - Rural vs urban
     - Wealth
     - Religion (Hindu, Buddhist, Muslim, Sikh, Christian, Jain, Jewish, Zoroastrian, other, none)
   - Causal intention

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

	Cluster membership (reference: cluster 1)			
	Cluster 2	*Cluster 3*	*Cluster 4*	*Cluster 5*
Age (years)	1.055***	1.038***	1.165***	1.084***
	(0.0004)	(0.005)	(0.001)	(0.0004)
Education (years)	0.986***	1.063***	1.007***	1.031***
	(0.001)	(0.010)	(0.001)	(0.001)
Wealth index (standard deviation)	0.910***	1.412***	1.443***	1.628***
	(0.004)	(0.051)	(0.006)	(0.005)
Rural *Residence (reference: urban)*	1.000***	0.951***	0.990***	1.116***
	(0.009)	(0.091)	(0.012)	(0.009)
Religion *(reference: Hindu)*				
Muslim	1.124***	1.373***	1.558***	1.522***
	(0.011)	(0.116)	(0.015)	(0.011)
Christian	1.289***	0.986***	1.570***	1.290***
	(0.013)	(0.165)	(0.018)	(0.014)
Sikh	0.919***	1.170***	2.189***	1.768***
	(0.030)	(0.252)	(0.030)	(0.025)
Buddhist	1.260***	0.915**	1.810***	1.228***
	(0.033)	(0.410)	(0.040)	(0.035)
Jain	1.290***	0.018***	1.199***	1.013***
	(0.112)	(0.00005)	(0.128)	(0.105)
Jewish	0.003***	0.526***	0.029***	1.334***
	(0.00000)	(0.00001)	(0.00000)	(0.00001)
Zoroastrian	4.175***	0.580***	3.354***	3.313***
	(0.257)	(0.0002)	(0.278)	(0.282)
No religion	2.492***	0.112***	5.235***	2.173***
	(0.168)	(0.0002)	(0.193)	(0.186)
Other religions	1.860***	0.190***	2.425***	1.505***
	(0.031)	(0.001)	(0.041)	(0.036)
Constant	0.244***	0.001	0.003	0.056***
	(0.018)	(0.196)	(0.027)	(0.019)

Different families of cluster analysis are useful for particular applications

TYPES

- **Centroid-based** (k means, medians, medoids …)
- **Distribution-based** (Gaussian, Bernoulli … mixture models)
- **Density-based** (HDBSCAN, DBSCAN, …)
- **Hierarchical** (trees)
- …

HEIDELBERG INSTITUTE OF GLOBAL HEALTH

How do health behavioral risks cluster?

BERNOULLI MIXTURE MODEL IN NAMIBIA