Associazione Italiana di Psicologia | SEZIONE SPERIMENTALE

Problems with clustering cases from quantitative psychological data on individual differences

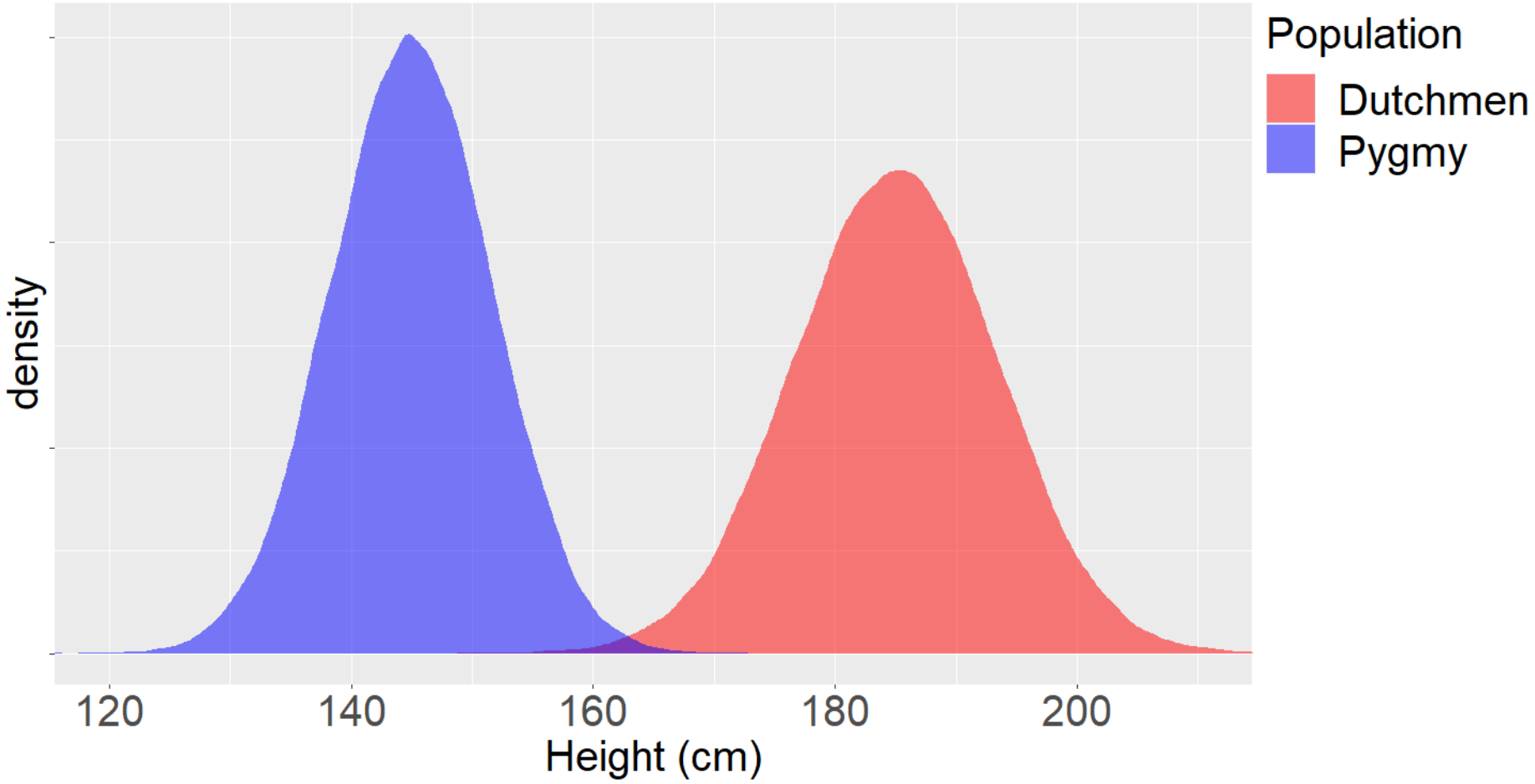
ENRICO TOFFALINI DPG, University of Padova
PAOLO GIRARDI DSS + DPSS, University of Padova
DAVID GIOFRÈ DISFOR, University of Genoa
GIANMARCO ALTOÈ DPSS, University of Padova

SESSION: Methods-3

Friday, 30th of September 2022

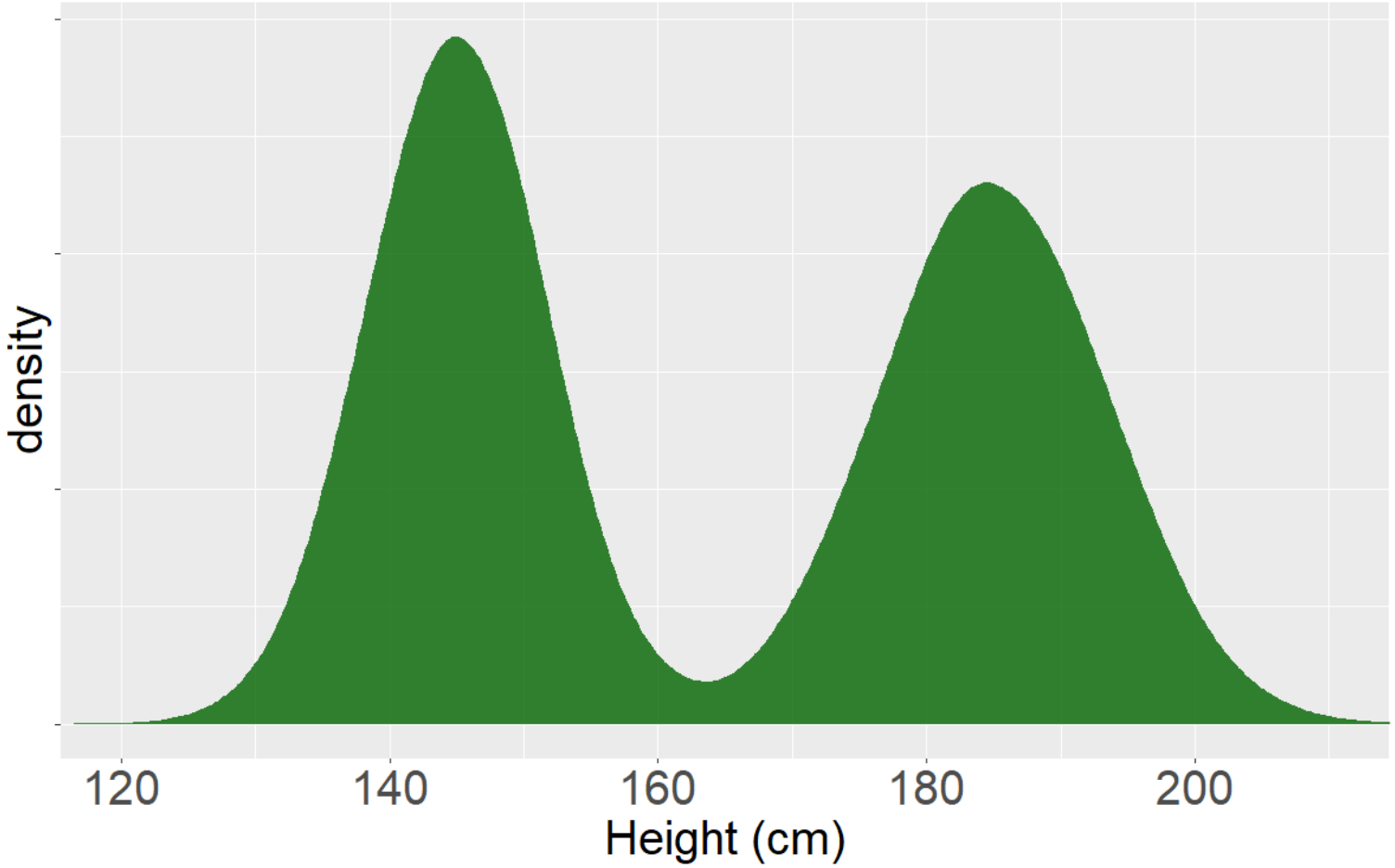
EXAMPLE 1

A convenient *unidimensional* example: dutchmen and pygmies are arguably the 2 most extremely different populations in terms of average height



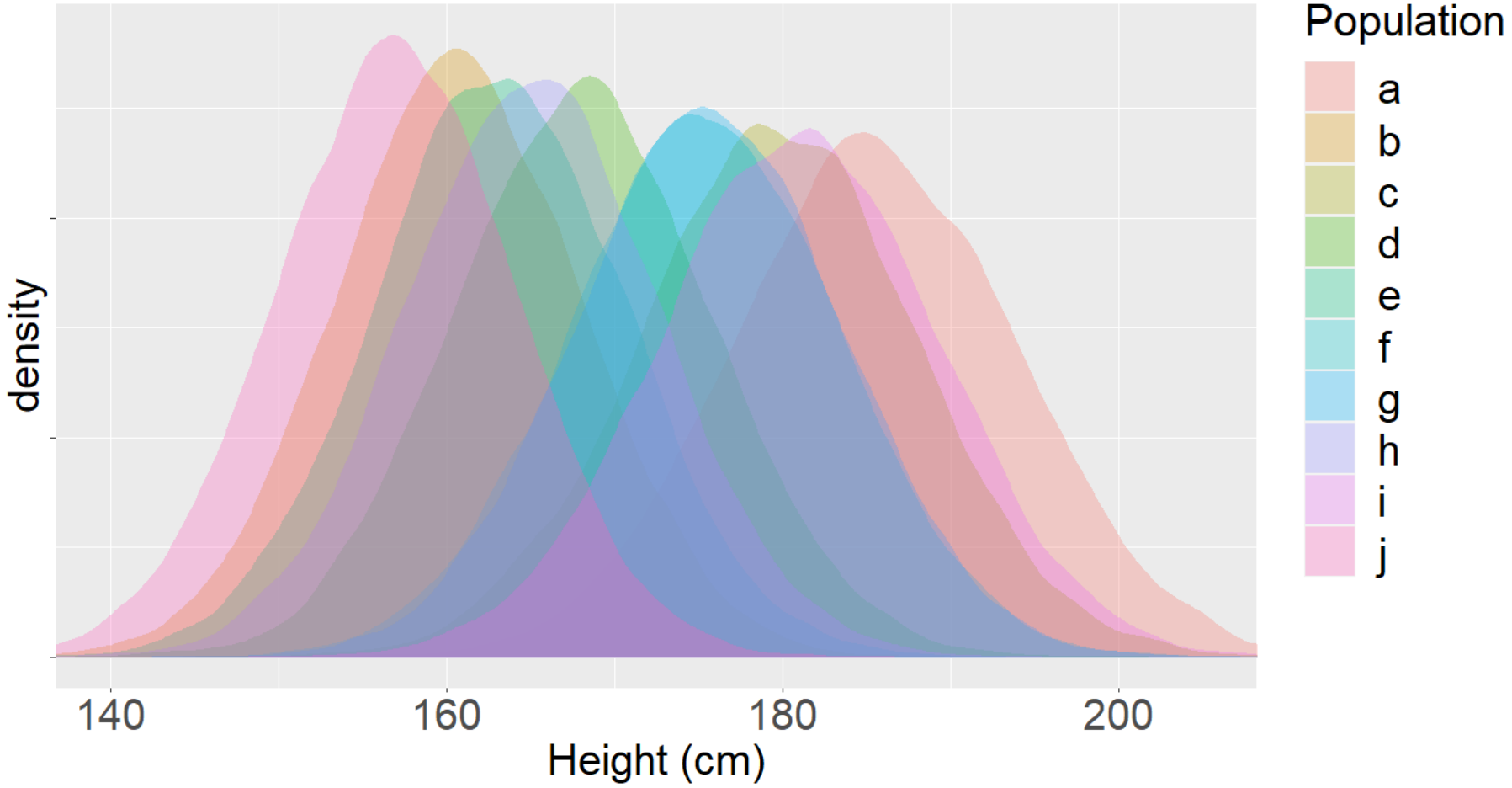
EXAMPLE 1

The difference is so large (about 5 SDs) that it speaks for itself even after losing info on population membership



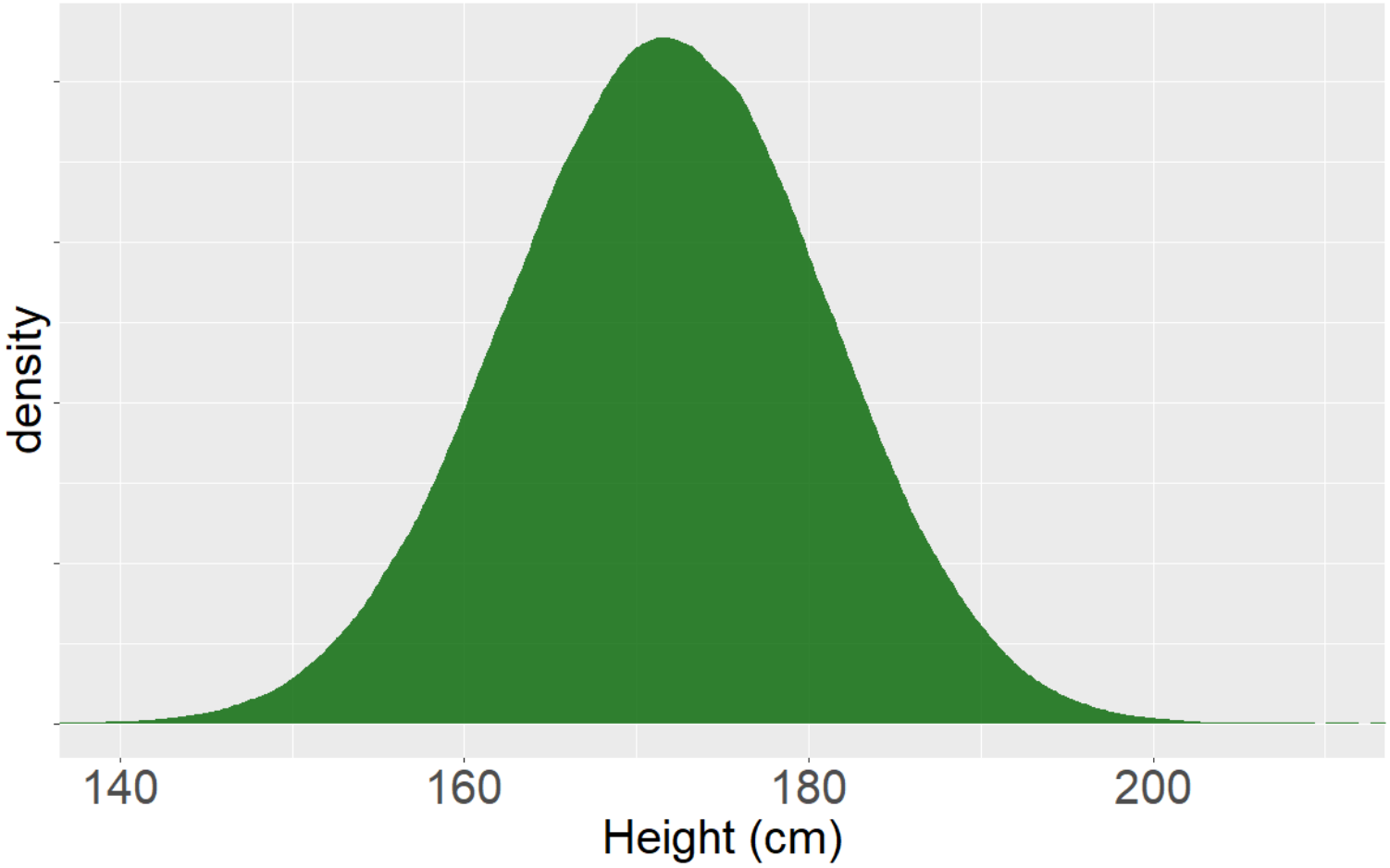
EXAMPLE 1

But what happens if, instead of just 2 extremely different populations you have a mix of many, just not carefully selected?



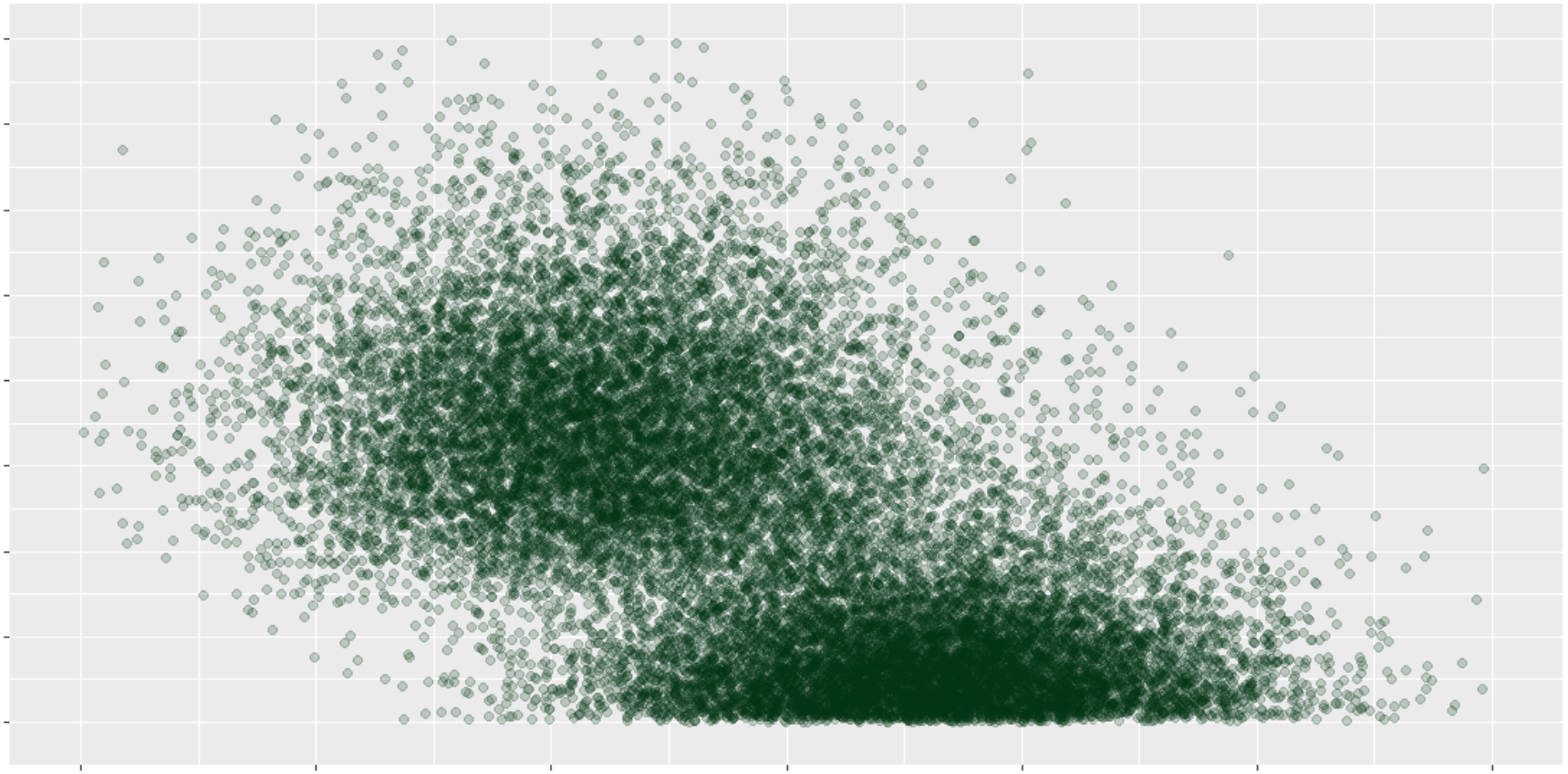
EXAMPLE 1

Once combined, they appear just as one larger population
Once lost, info on membership appears irretrievable



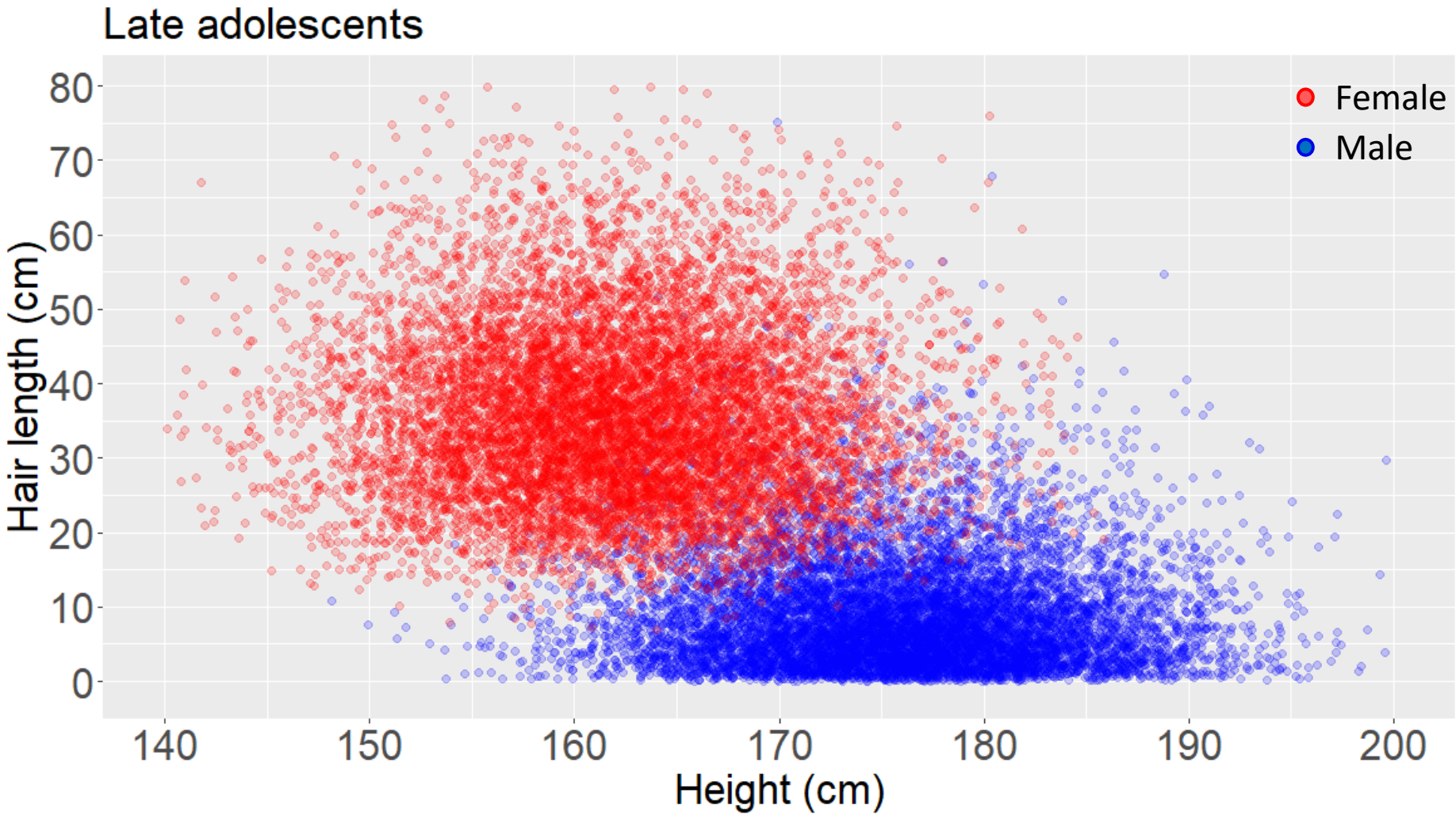
EXAMPLE 2

Now let's take a *bidimensional* case…



It's easy to spot **two sub-populations** within this **bivariate distribution!**

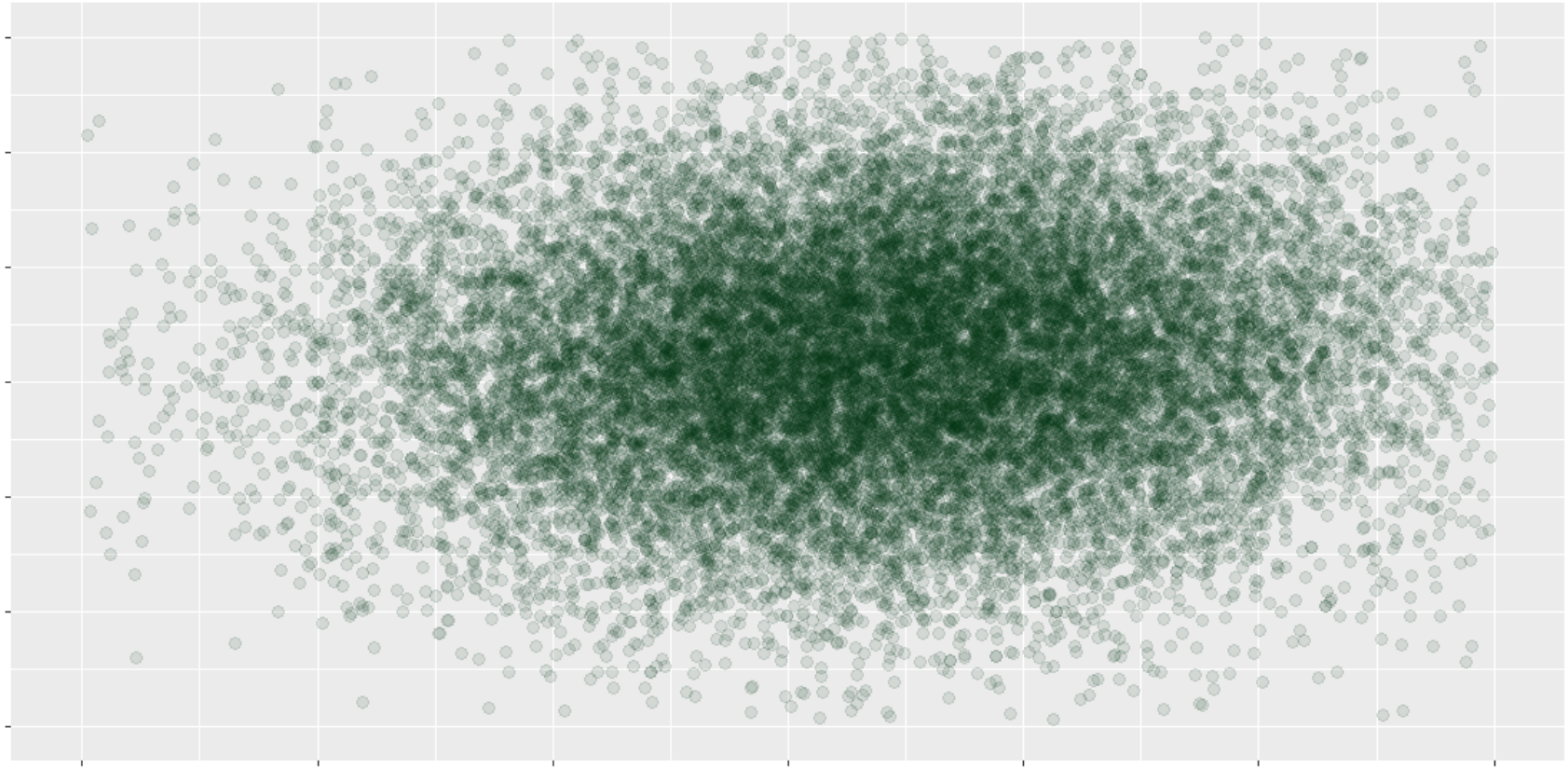
EXAMPLE 2



- Girls are shorter and have longer hair than boys, on average
- Within cluster, height and hair length are uncorrelated

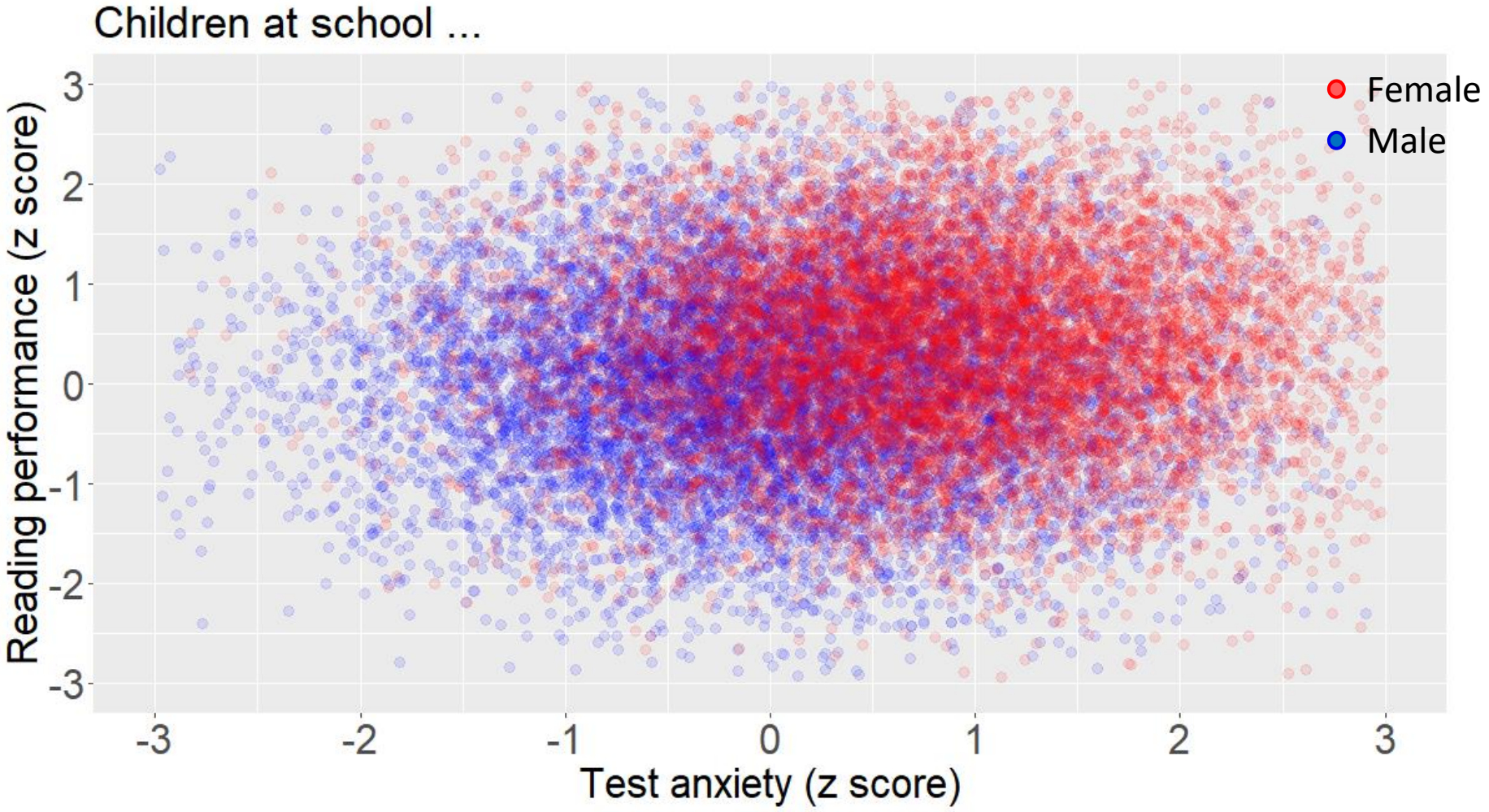
EXAMPLE 3

Can you see any **sub-population** here?



I don’t

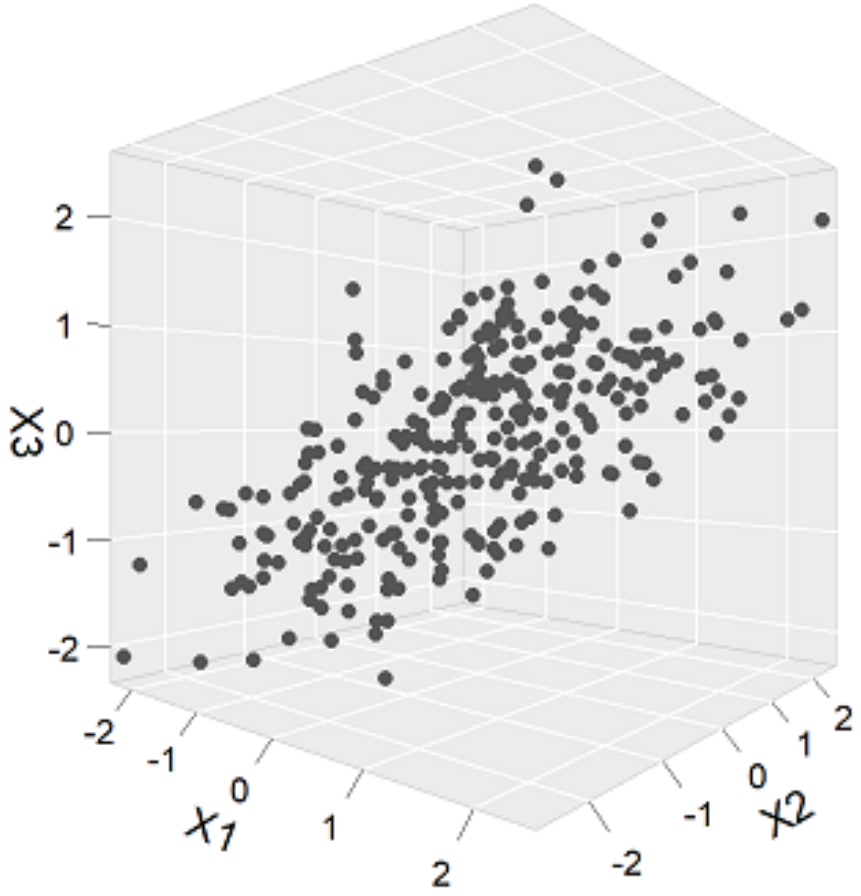
EXAMPLE 3



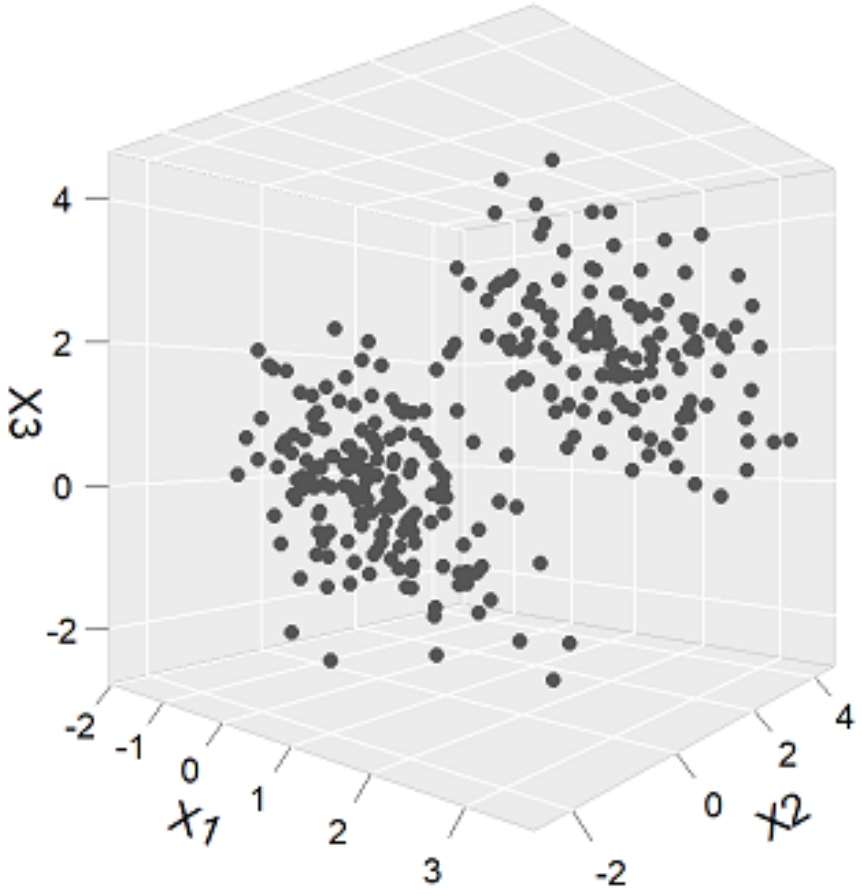
- Females outperform males in reading comprehension but are also more test anxious
- There are average differences, but just not large enough to emerge from unsupervised learning (perhaps unless you have millions of observations)

EXAMPLE 4

(1) Only one population emerges from the multivariate space



(2) Two clusters emerge from the multivariate space

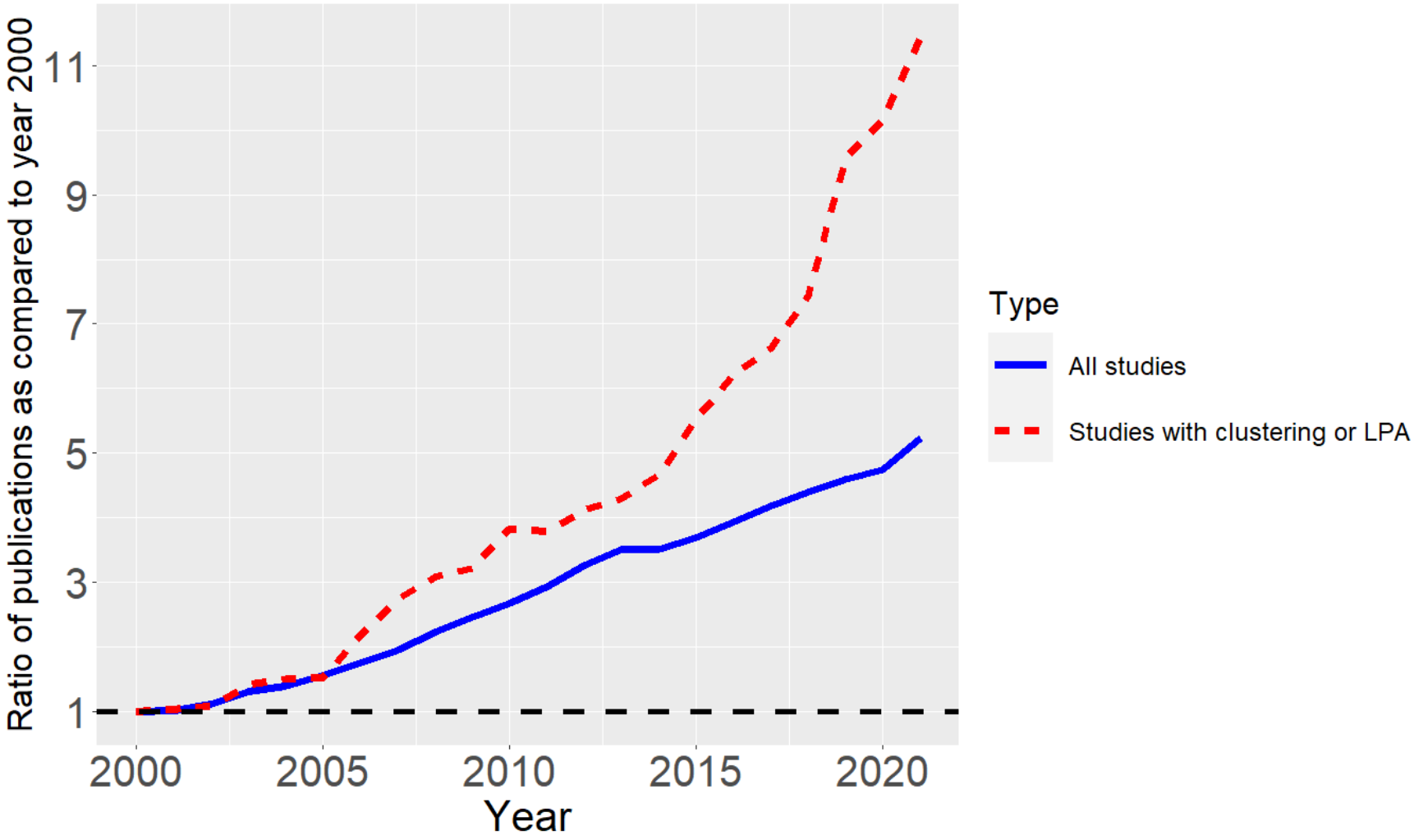


***Clustering* and *Latent profile analysis* are unsupervised machine learning methods** that may help you discover previously unobserved sub-populations within larger populations, as in the previous examples

But under what conditions?

- Large enough sample size is an obvious factor
- Large enough separation(s) / effect size(s) is obviously another one
- Other relevant factors: *number of dimensions / indicators, correlations, distribution of indicators*

Tremendous growth of publications mentioning clustering or LPA in psychology!



SCOPUS - Keywords: "clustering" or "LPA" or "latent profile analysis" in Title, Abstract, Keywords, as compared to all publications – Limited to "Social Sciences" and "Psychology"

Review of 191 studies implementing clustering or LPA methods in psychology, published in 2016-2020 - indexed in Scopus

Table 1. Percentiles of interest for the number of individuals on which clustering was performed, the number of indicators, and the number of clusters identified, across the 191 studies reviewed.

	Percentiles				
	5th	*25th*	*50th*	*75th*	*95th*
N. of individuals on which clustering was performed	66	153	322	589	2,119
N. of variables (indicators) used for clustering	3	4	6	9	19
N. of clusters identified	2	3	3	4	6

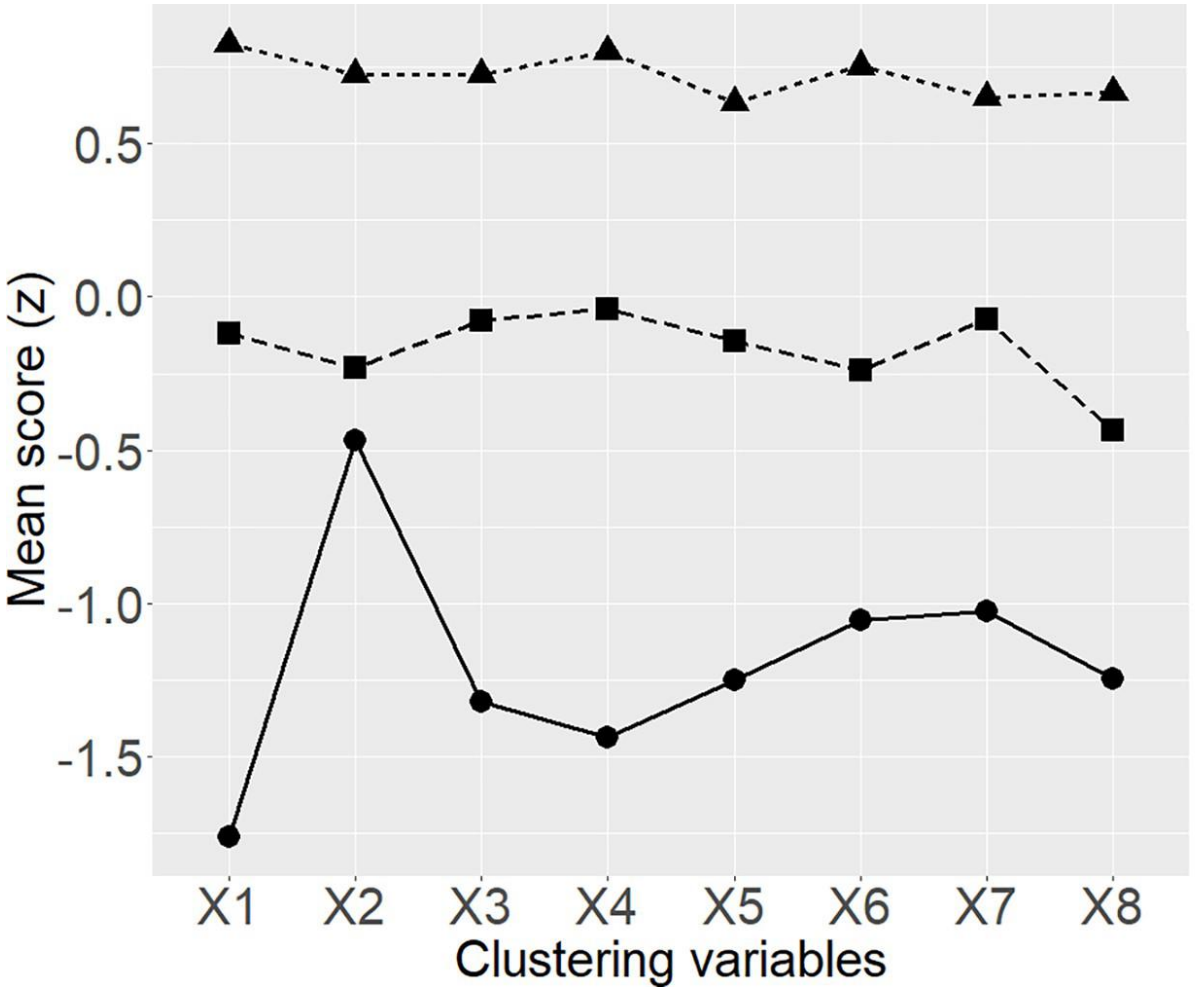
https://doi.org/10.1371/journal.pone.0269584.t001

"median" study

- Median IF was 3.23; Q1 is overrepresented (41%)
- Psychiatry (29%), Psychology developmental (20%), Psychology Clinical (13%)
- Most use LPA (76%), other use hierarchical clustering (11%), or partitioning (9%)
- Dimensionality reduction largely missing or used at most locally (96%)
- Many fail to test the one-cluster solution (34%) OR are unclear (21%)
- Not a single study concludes in favor of the one-cluster solution!
- Almost half of studies show clusters dominated by "high" vs "low" profiles (48%)

A common artifact

Cluster solutions featuring “high” vs “low” profiles emerge frequently from (positively) correlated indicators, even when there are NO real clusters in the data (e.g., because you simulate data and there is no clustering in the generative model!)



EXAMPLE – Optimal Gaussian mixture model solution selected via BIC after model-based clustering; N = 200 simulated observations from ONE multivariate Gaussian distribution with 8 dimensions, correlated r = 0.30

2 to 4 clusters as optimal solutions 100% of times over 1,000 iterations under the above conditions

Simulation study - Method

Data simulated from multivariate Gaussian

$$Y \sim N_p(0, \Sigma(\rho)), \text{ with 1 group}$$

$$Y|G = g \sim N_p(d\, I_2(g), \Sigma(\rho)), \text{ with 2 groups}$$

Parameters of simulation

either **1** or **2 true clusters** ($I_2(g)$ is 1 when g = 2; 0 otherwise)

p = **3**, **6**, or **12** indicators

ρ = **0.00** (orthogonal indicators), **0.20** (plausible modest, e.g., social/personality scales), **0.50** (plausible strong, e.g., cognitive/intelligence scales)

Cohen's d = **0** (1 true cluster), **0.40** (modest, plausible in psychology), **0.80** (large, unlikely for genuine discovery of previously unobserved sub-populations), **1.50** (very large, implausible in psychology unless reflect diagnostic criteria)

500 iterations for each condition

Methods tested

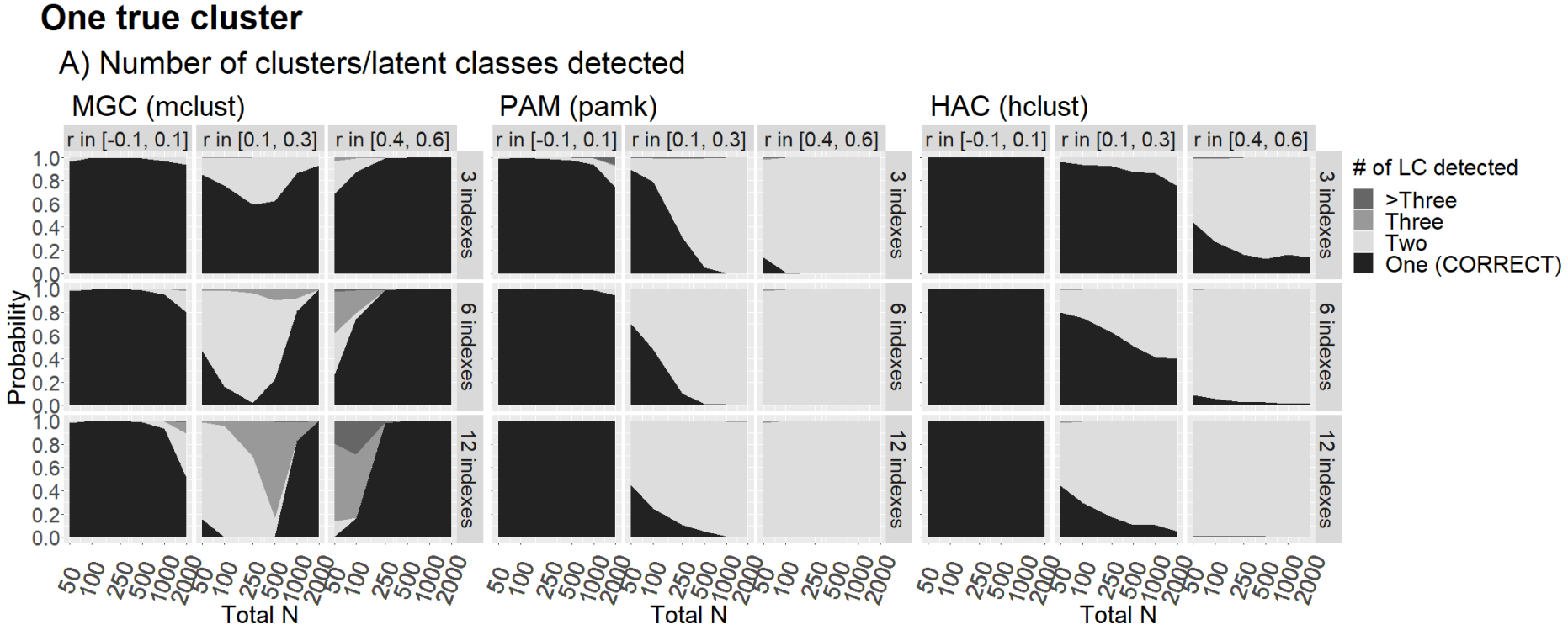
- Model-based clustering based on Gaussian mixture models (MGC), estimated via expectation-maximization model; BIC for model selection («mclust» in R)
- Partition algorithm Around Medoids (PAM*; close but more robust than k-means)
- Hierarchical Agglomerative Clustering (HAC*)

* for PAM and HAC: Duda-Hart test ($p<.05$) for one vs multiple clusters; average Silhouette profile index for choosing number of clusters if ≥ 2

Outputs considered

- Number of cluster selected as best solution by the algorithm
- Rand index for classification accuracy

Scenario 1: ONE true cluster (no sub-populations)

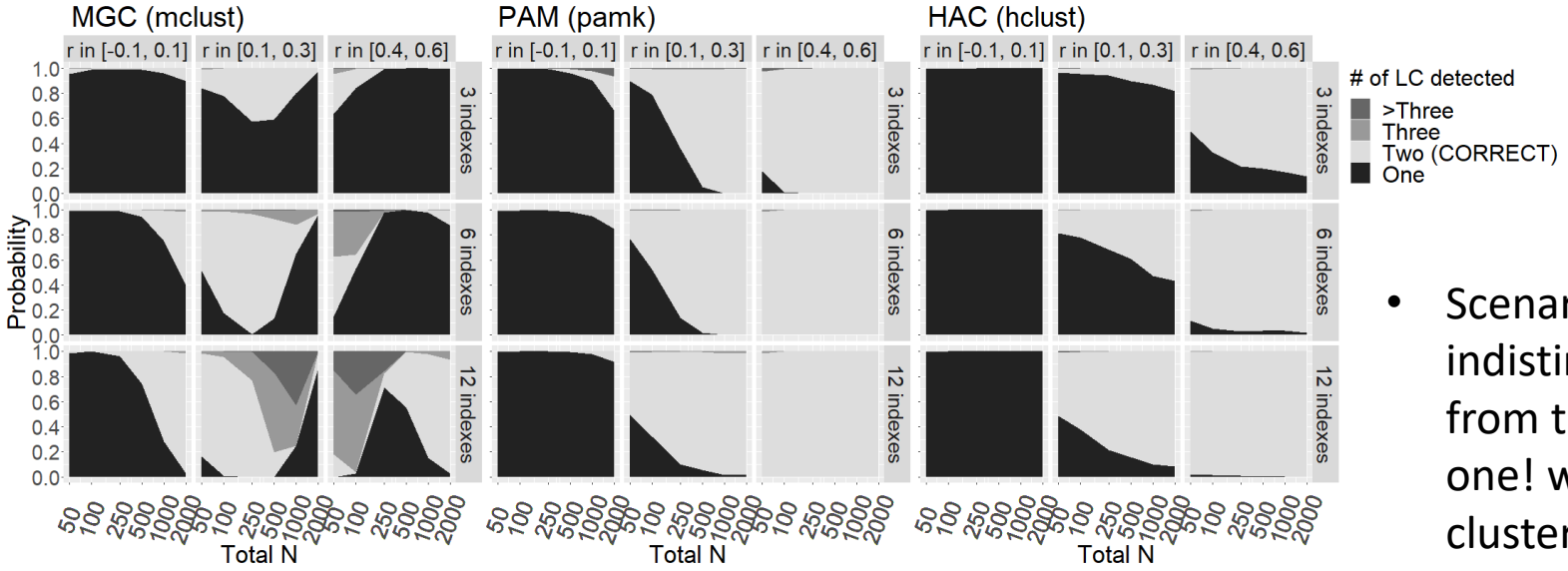


- High risk of ***false positives*** if indicators are correlated! (more indicators = more risk)
- Risk ***NOT*** avoided with a ***MGC*** model with medium sample sizes (N = 100-500), especially with many indicators!

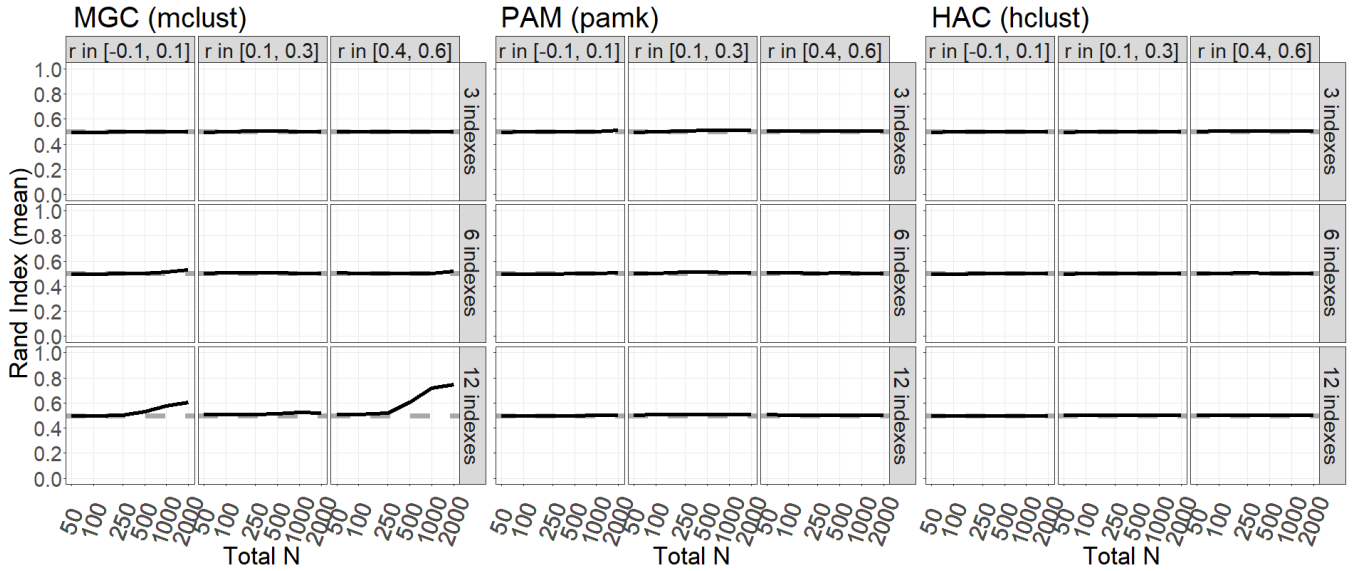
Scenario 2: TWO true cluster, small (plausible) separations

Two true clusters, Cohen's d in [0.2, 0.6]

A) Number of clusters/latent classes detected

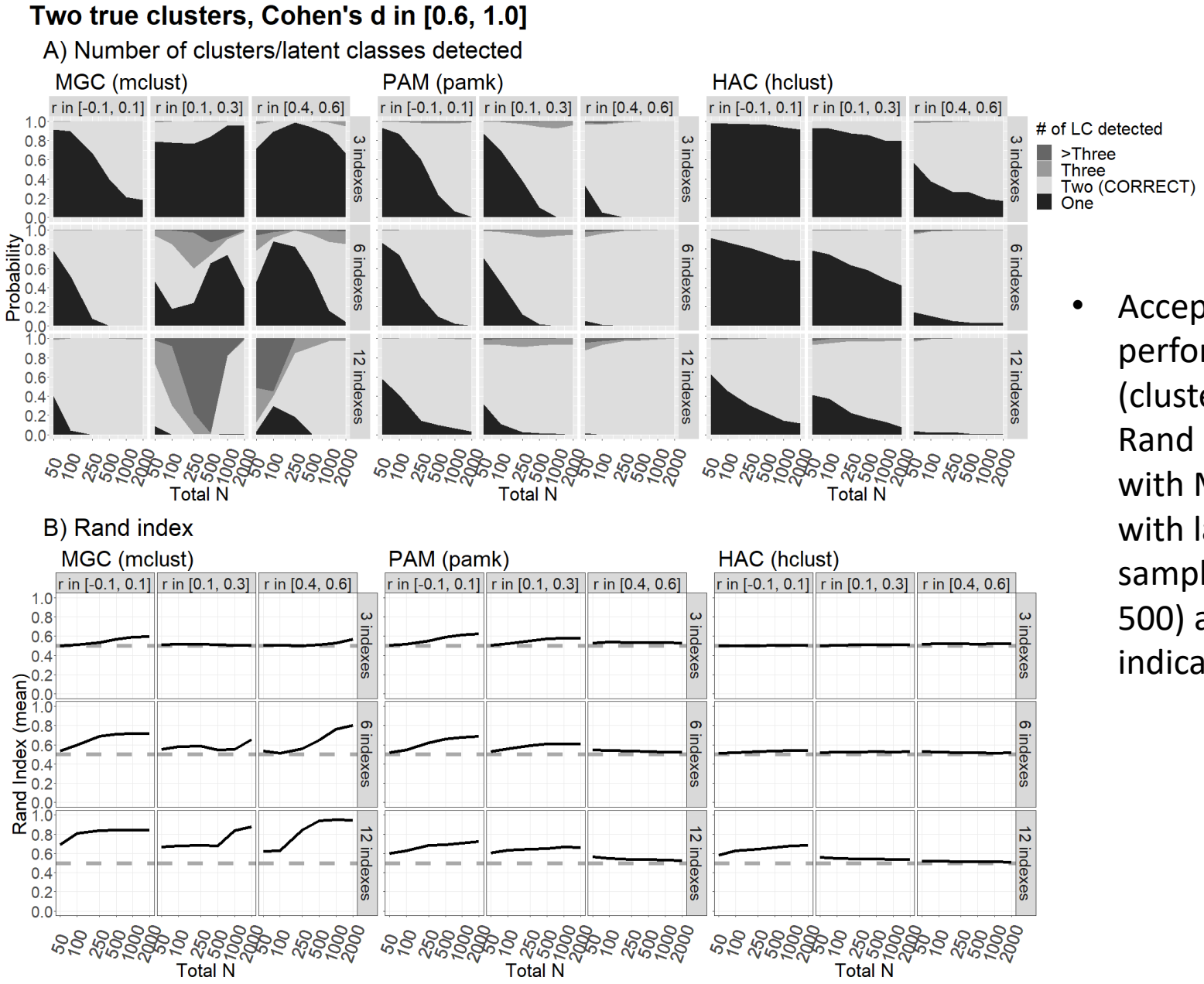


B) Rand index



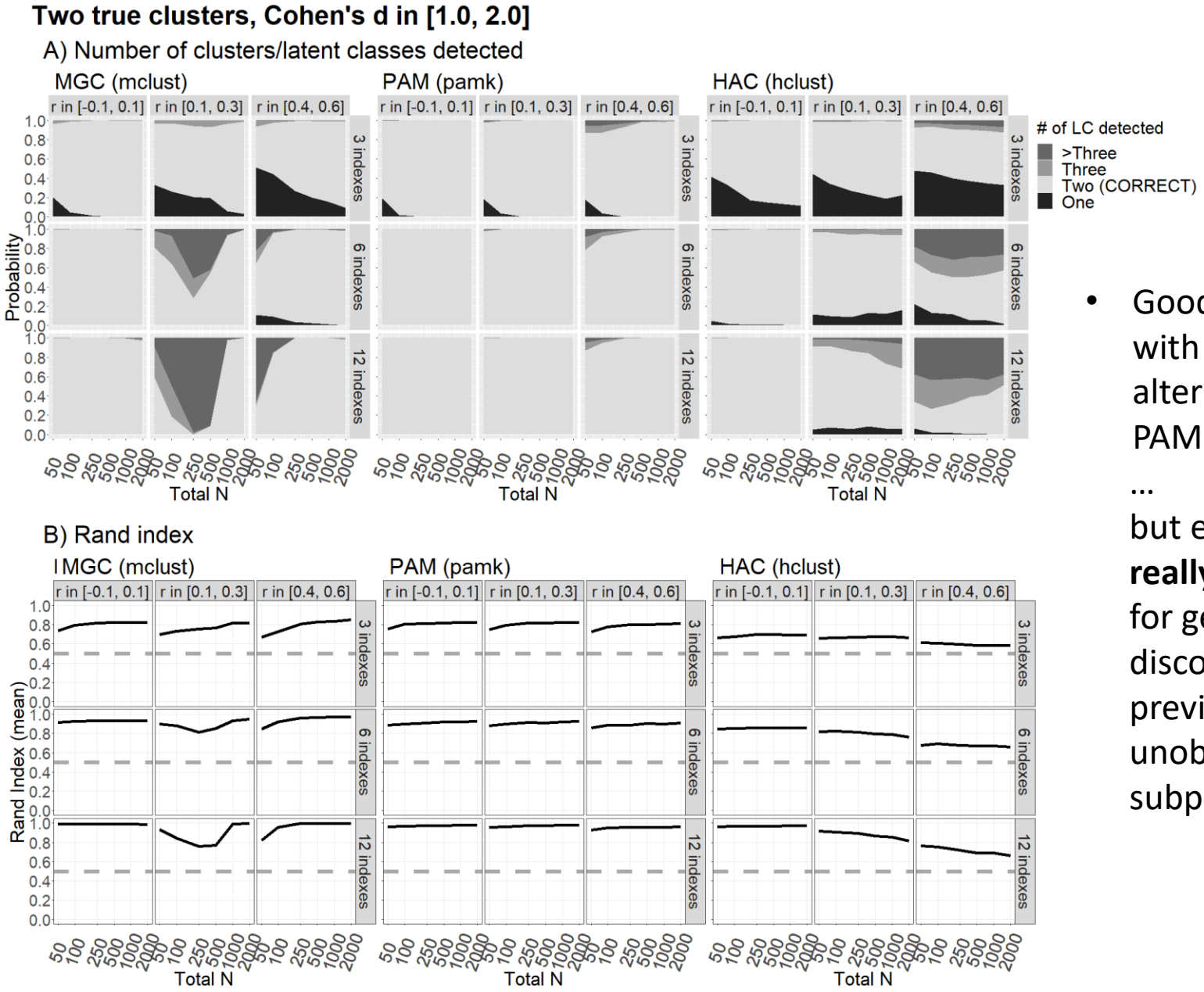
- Scenario virtually indistinguishable from the previous one! with ANY clustering method
- Even when the correct number of clusters is detected, accuracy of classifications is extremely low

Scenario 3: TWO true cluster, large (dubious!) separations



- Acceptable performance (clusters detected + Rand index) only with MGC models, with large enough sample sizes (N > 500) and many indicators

Scenario 4: TWO true cluster, large (implausible) separations



- Good performance with ***most*** alternatives (but PAM might be ideal)

…

but effect sizes are **really implausible** for genuine new discovery of previously unobserved subpopulations!

Conclusions

- Under a reasoned set of scenarios plausible for the cognitive research, *none* of the methods adequately discriminates between one vs two true clusters
- High risk of incorrectly detecting multiple clusters where none exist, when indicators are correlated… even with MGC, which should model covariance matrices – might be typical of real psychological research scenarios!
- It is hard for researchers to be in a condition to achieve a valid unsupervised clustering for inferential purposes with a view to classifying individuals
- Do you *really* need clustering?!

Entia non sunt multiplicanda sine necessitate

for more details

PLOS ONE

OPEN ACCESS PEER-REVIEWED RESEARCH ARTICLE

Entia Non Sunt Multiplicanda ... Shall I look for clusters in my cognitive data?

Enrico Toffalini, Paolo Girardi, David Giofrè, Gianmarco Altoè

Published: June 30, 2022 • https://doi.org/10.1371/journal.pone.0269584

On related issues (i.e., applied cases for the Ockham's razor) , see also:

The Journal of Child Psychology and Psychiatry

Journal of Child Psychology and Psychiatry **:* (2021), pp **–** doi:10.1111/jcpp.13397

No evidence for a core deficit in developmental dyscalculia or mathematical learning disabilities

Irene C. Mammarella,[1] Enrico Toffalini,[2] Sara Caviola,[1,3] Lincoln Colling,[4] and Denes Szűcs[5]

[1]Department of Developmental and Social Psychology, University of Padova, Padova, Italy; [2]Department of General Psychology, University of Padova, Padova, Italy; [3]School of Psychology, University of Leeds, Leeds, UK; [4]School of Psychology, University of Sussex, Sussex, UK; [5]Department of Psychology, University of Cambridge, Cambridge, UK

SCIENTIFIC STUDIES OF READING

https://doi.org/10.1080/10888438.2021.1897596

Routledge Taylor & Francis Group

Inferring the Performance of Children with Dyslexia from that of the General Population: The Case of Associative Phonological Working Memory

Barbara Carretti, Cesare Cornoldi, Arianna Antonello, Laura Di Criscienzo, and Enrico Toffalini

Department of General Psychology, University of Padua, Padua, Italy

ABSTRACT

The study examines whether the average performance of the population with dyslexia in a working memory measure can be inferred dimensionally from the characteristics of the typical population. Specifically, we focused on Associative Phonological Working Memory (APWM), an ability that we predicted being impaired in dyslexia due to the relationship of reading with both associative learning and working memory (WM). Study 1a revealed a linear relationship between APWM and reading ability in 438 typically-developing (TD) children, after accounting for fluid intelligence and phonological WM. In Study 1b a simulation procedure was used to calculate the deficit in APWM expected in children with dyslexia, based on the set of correlations found in Study 1a. This prediction was compared with the actual performance of 26 children with dyslexia. A deficit in APWM was confirmed, and its extent was in line with that simulated from the correlational structure seen in the TD population.

Associazione Italiana di Psicologia | SEZIONE SPERIMENTALE

THANK YOU

FOR LISTENING

ENRICO TOFFALINI DPG, University of Padova

PAOLO GIRARDI DSS + DPSS, University of Padova

DAVID GIOFRÈ DISFOR, University of Genoa

GIANMARCO ALTOÈ DPSS, University of Padova

ΣΨ PSICOSTAT

STATISTICS IS ART

SESSION: Methods-3