

Perché è meglio non usare (praticamente mai) l'analisi dei cluster sulle casistiche in psicologia dello sviluppo

ENRICO TOFFALINI

Dipartimento di Psicologia Generale
Università di Padova

FILIPPO GAMBAROTA

*Dipartimento di Psicologia dello Sviluppo
e della Socializzazione*
Università di Padova

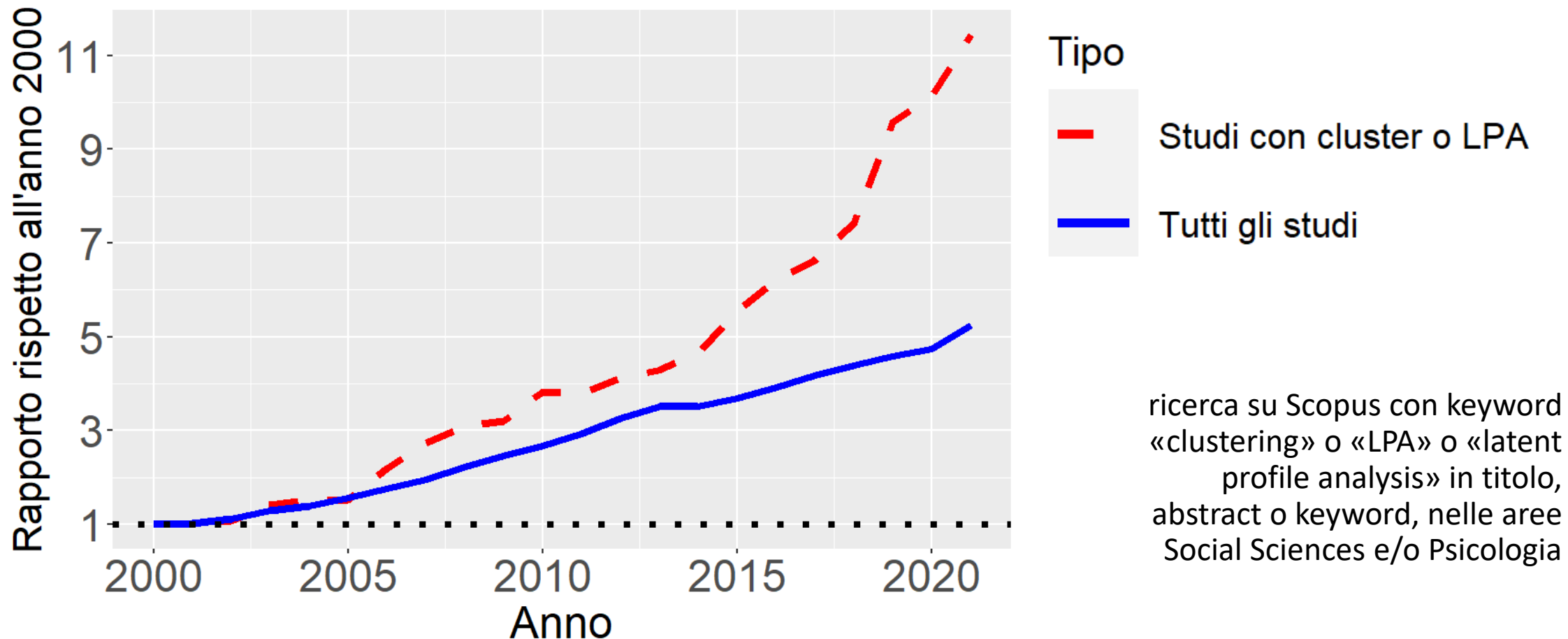


con un grazie al team di lavoro: Paolo Girardi, Ambra Perugini, David Giofrè, Gianmarco Altoè



XXXI Congresso
21/23 sett 2023
| Foggia

Cresce la popolarità dell'analisi dei cluster in psicologia



XXXI Congresso
21/23 sett 2023
Foggia

Enrico Toffalini, DPG, UNIPD
enrico.toffalini@unipd.it

Cresce la popolarità dell'analisi dei cluster in psicologia

Esempio: un'influente review di **Astle e colleghi (2022)** raccomanda l'uso del *clustering* tra i metodi *data driven* nello studio dei disturbi del neurosviluppo, per «scoprire» i «veri» raggruppamenti *transdiagnostici* delle casistiche... nell'ottica di mettere (giustamente!) in discussione le tradizionali categorie diagnostiche



Annual Research Review | [Open Access](#) |  

Annual Research Review: The transdiagnostic revolution in neurodevelopmental disorders

Duncan E. Astle , Joni Holmes, Rogier Kievit, Susan E. Gathercole

First published: 23 July 2021 | <https://doi.org/10.1111/jcpp.13481> | Citations: 48



Volume 63, Issue 4
April 2022
Pages 397-417

Advertisement





XXXI Congresso
21/23 sett 2023
| Foggia

Enrico Toffalini, DPG, UNIPD
enrico.toffalini@unipd.it

Ma di cosa stiamo parlando?

- *Clustering*: insieme di metodi esplorativi che raggruppano «oggetti» simili tra loro, distribuiti in uno spazio (di solito) multidimensionale
- Classico uso in psicologia: «svelare» l'esistenza di G sottopopolazioni *finora non rilevate* di individui in una popolazione più ampia, usando k indicatori

Esempio: $N = 41$ partecipanti,
«scopro» $G = 3$ cluster
basandomi su $v = 2$ dimensioni

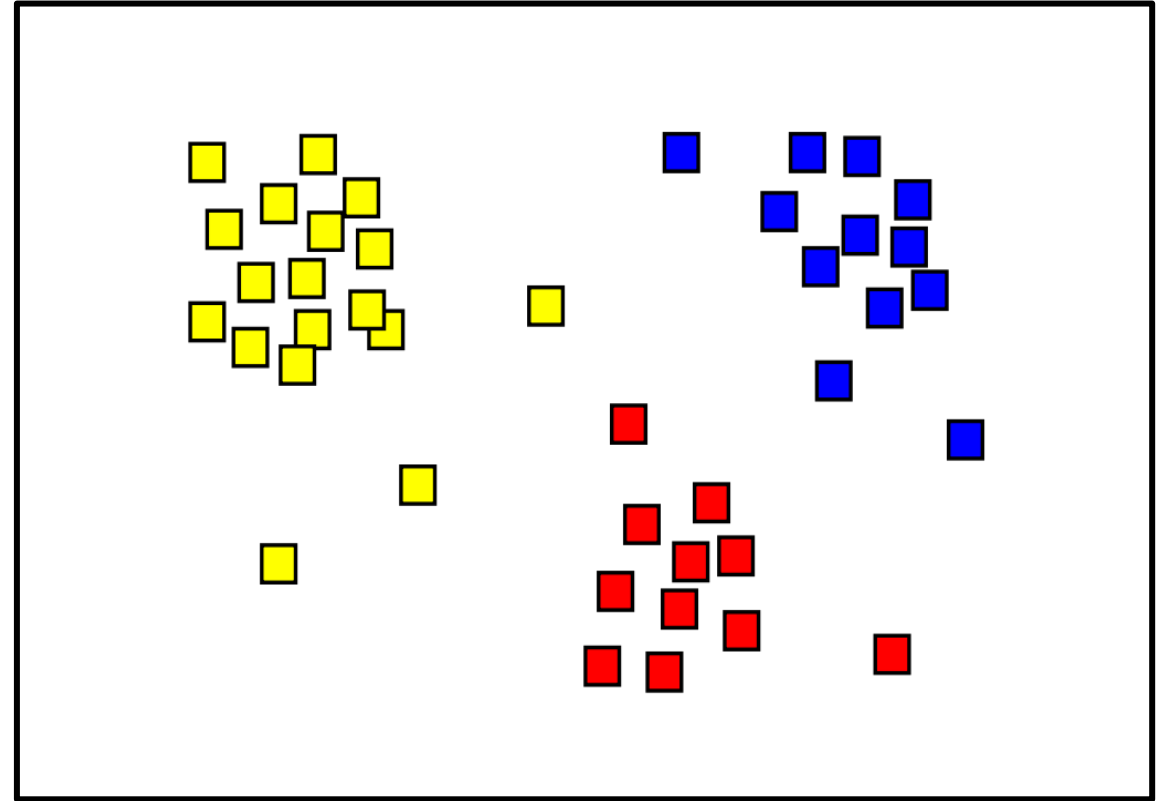


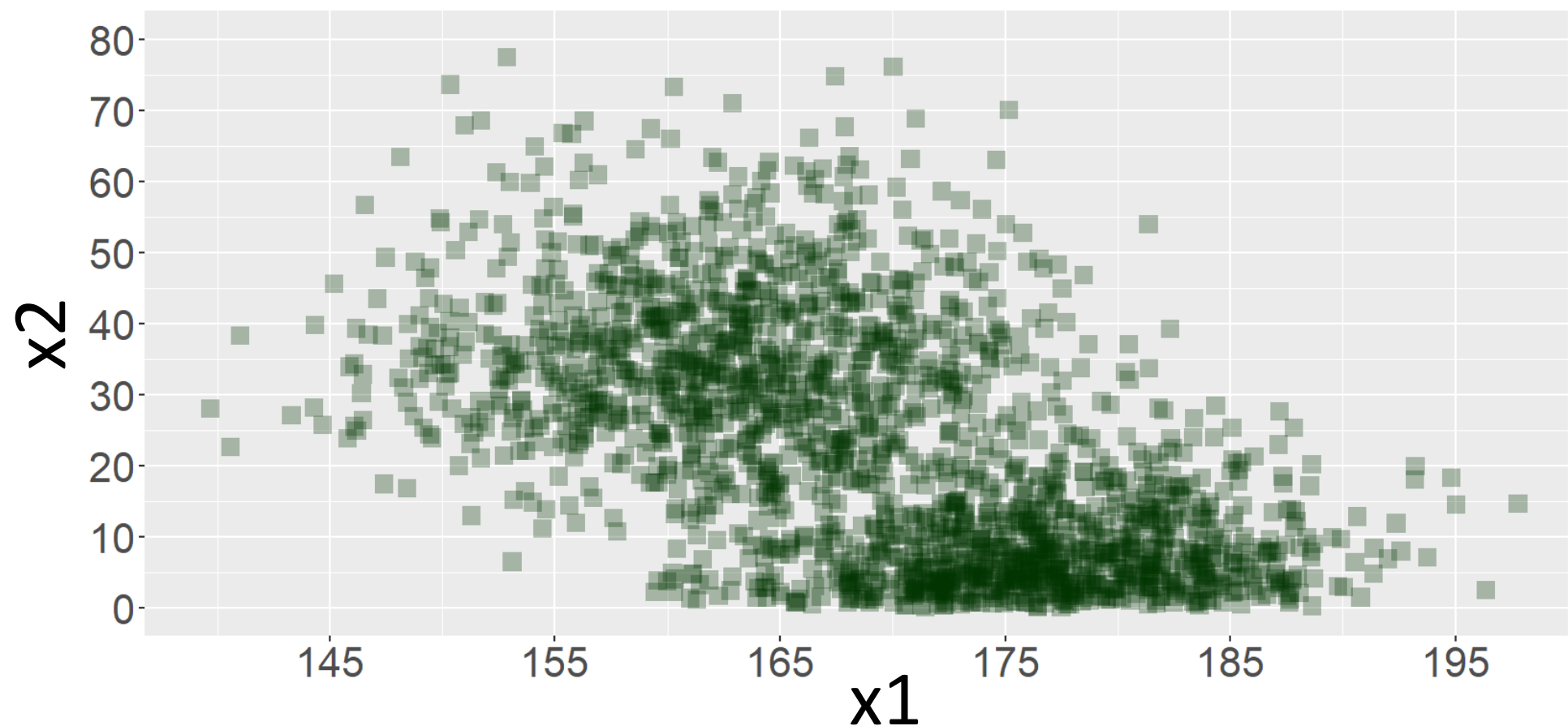
immagine da Wikipedia
https://en.wikipedia.org/wiki/Cluster_analysis



XXXI Congresso
21/23 sett 2023
| Foggia

Enrico Toffalini, DPG, UNIPD
enrico.toffalini@unipd.it

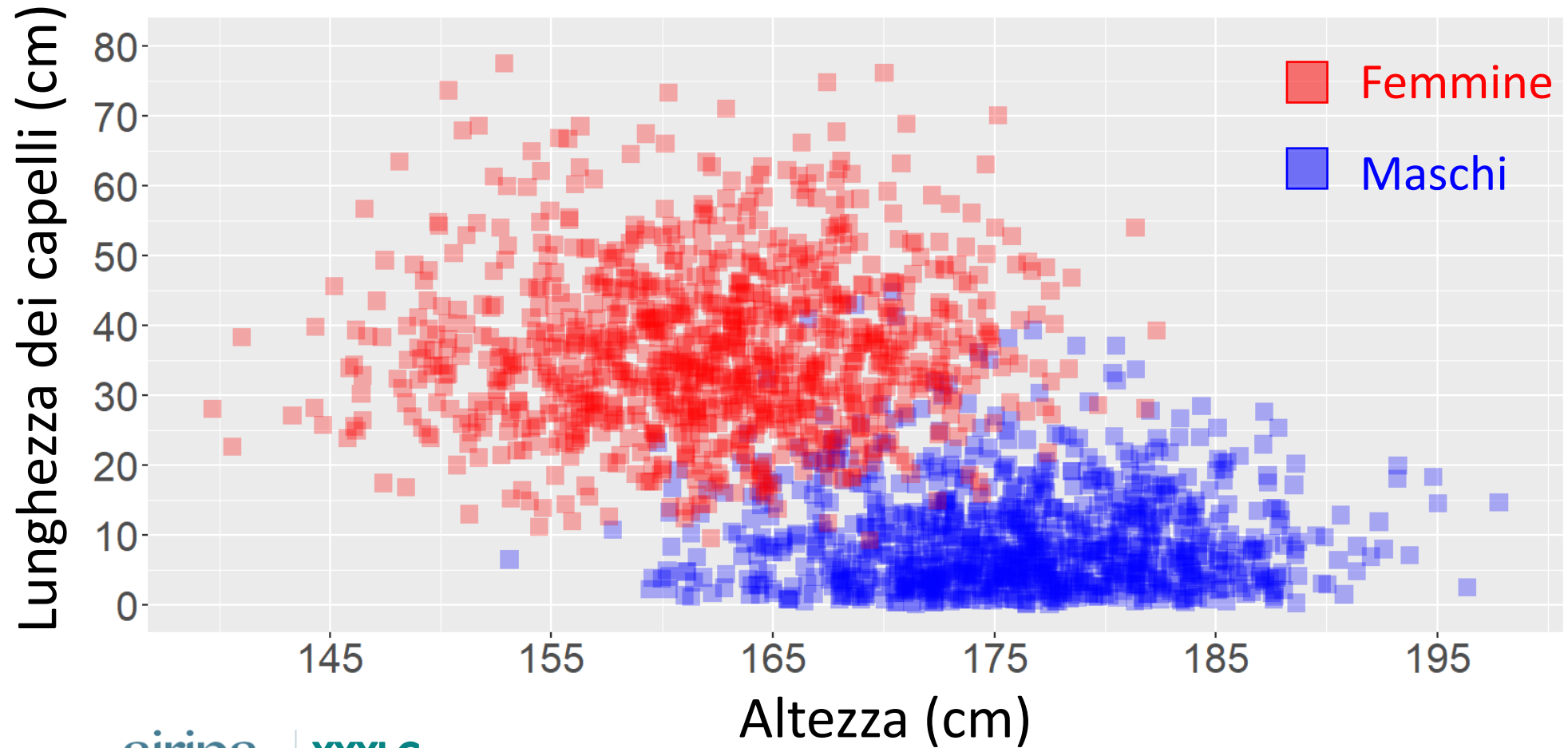
Si vedono, qui, i cluster?



XXXI Congresso
21/23 sett 2023
| Foggia

Enrico Toffalini, DPG, UNIPD
enrico.toffalini@unipd.it

Si vedono, qui, i cluster?



XXXI Congresso
21/23 sett 2023
| Foggia

Enrico Toffalini, DPG, UNIPD
enrico.toffalini@unipd.it

Clustering e inferenza statistica

Usiamo spesso metodi di clustering per fare inferenza statistica, ad esempio sono famosi:

- ***k-means clustering*** con Duda-Hart test ($p < 0.05$) per discriminare uno vs più cluster, e il valore di Silhouette per stabilire il numero preciso;
- ***mixture models*** (*model-based*; raccomandato) con indice BIC per selezionare la soluzione (modello) migliore e stabilire il numero di cluster/componenti;

Questi sono tra i metodi più frequenti emersi da una review di 191 studi con clustering pubblicati negli scorsi 5 anni (Toffalini et al., 2022)



XXXI Congresso
21/23 sett 2023
| Foggia

Enrico Toffalini, DPG, UNIPD
enrico.toffalini@unipd.it

1° problema in psicologia: la *potenza*

Nel caso visto prima il clustering funzionava egregiamente... **MA...**
N = 2000 osservazioni; Cohen's $d = 2.00$ su altezza; Cohen's $d = 2.72$ su lunghezza capelli

- eh sì... per farvi «vedere» i cluster, ho dovuto ricorrere a un esempio non-psicologico
- Il clustering difficilmente funziona se non abbiamo Cohen's $d \geq 0.80$ su diversi indicatori, possibilmente non correlati (es. Tein et al., 2013; Toffalini et al., 2022)
- **MA** in psicologia effetti così ampi sono comunque rari, ed è quasi irragionevole supporli tra sottopopolazioni *non ancora scoperte*



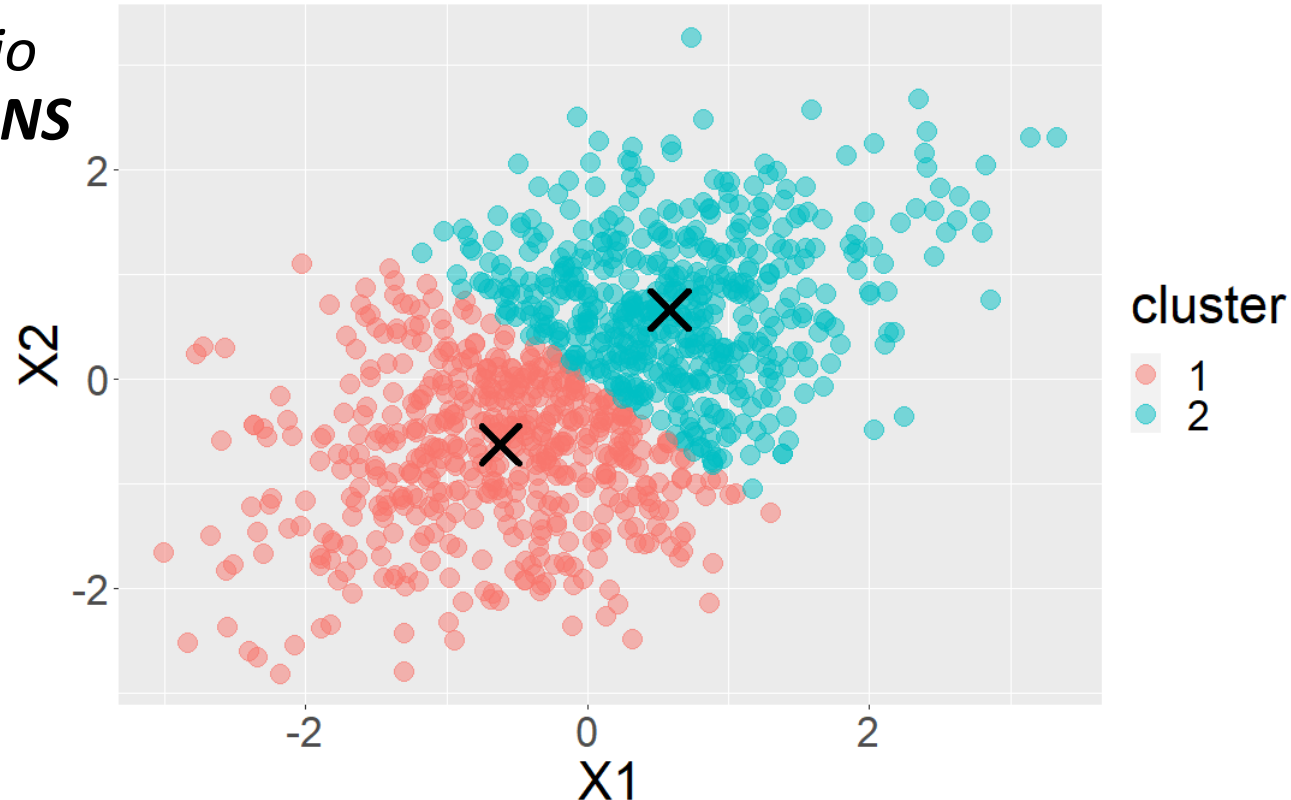
XXXI Congresso
21/23 sett 2023
| Foggia

Enrico Toffalini, DPG, UNIPD
enrico.toffalini@unipd.it

2° problema in psicologia: i *falsi positivi*

Metodi non-model based come *k-means*, *PAM*, *hierarchical clustering*, sono estremamente sensibili alla correlazione tra indicatori

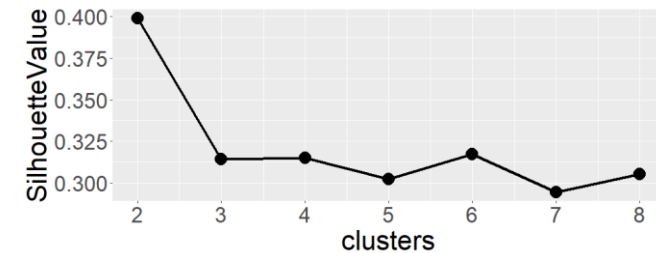
Esempio
K-MEANS



Simulati $N = 1000$ casi, normalmente distribuiti, su $v = 2$ indicatori correlati $r = 0.50$

k-means (fpc::pamk) rileva come soluzione ottimale $G = 2$ **cluster che non esistono**

Duda-Hart test: $p < 0.001$; $dh = 0.51$



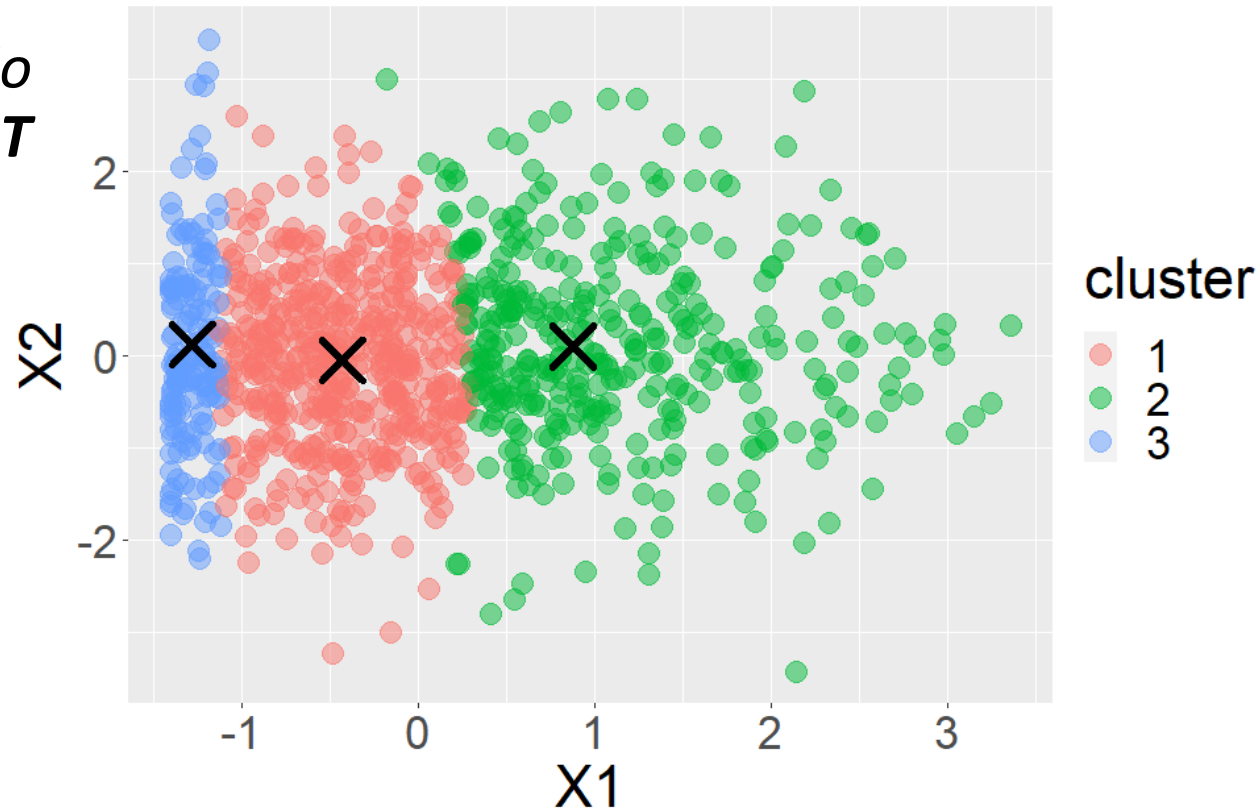
XXXI Congresso
21/23 sett 2023
| Foggia

Enrico Toffalini, DPG, UNIPD
enrico.toffalini@unipd.it

2° problema in psicologia: i *falsi positivi*

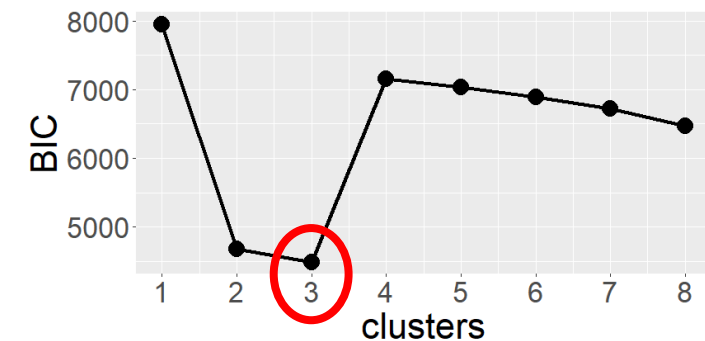
Metodi *model-based* sono sensibili a violazione assunzione distribuzioni

Esempio
MCLUST



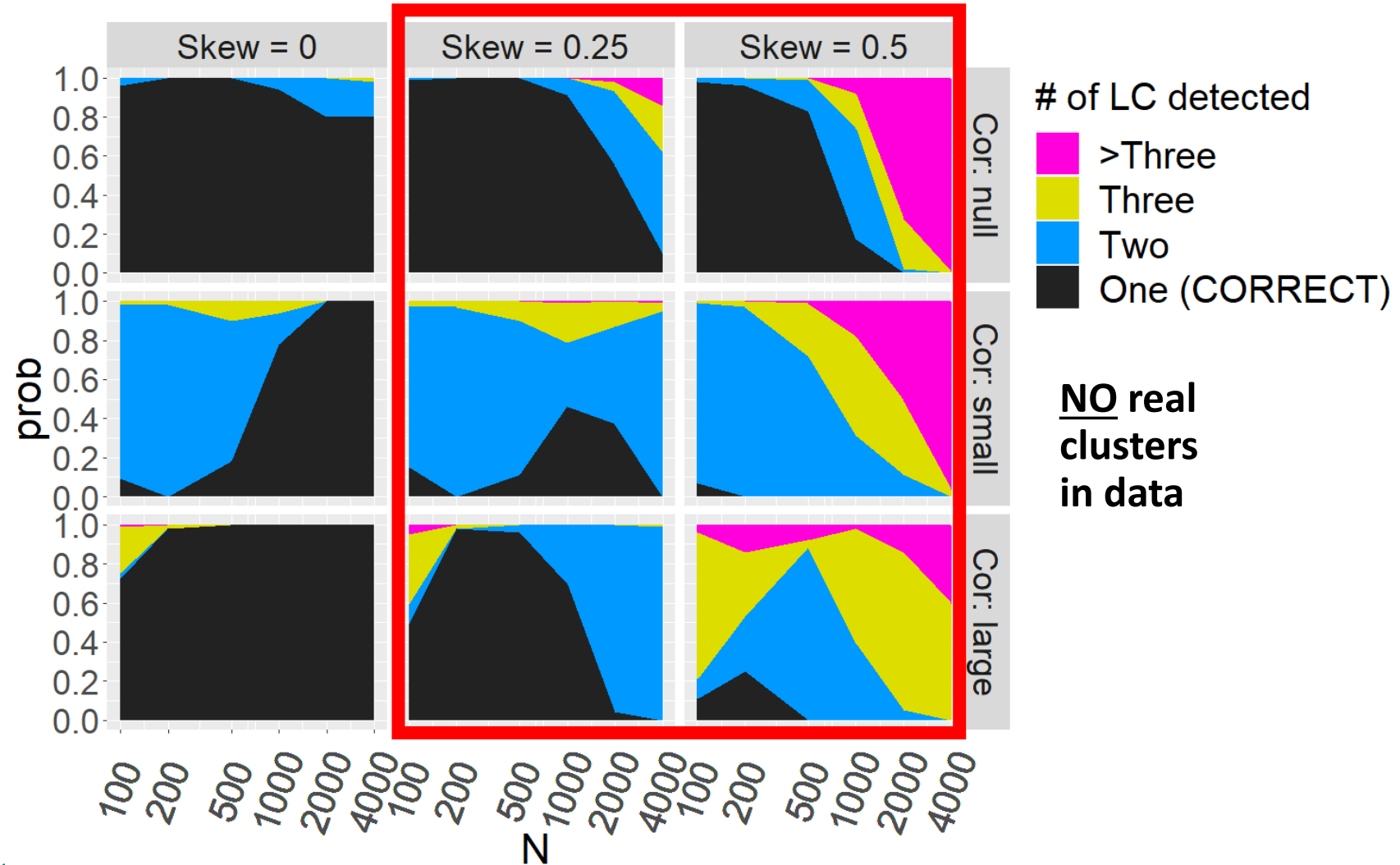
Simulati $N = 1000$ casi, normalmente distribuiti, su $v = 2$ indicatori non correlati ($r = 0.00$), ma uno dei due ha skewness = 0.70

mclust (`mclust::Mclust`) rileva come soluzione ottimale $G = 3$
cluster che non esistono

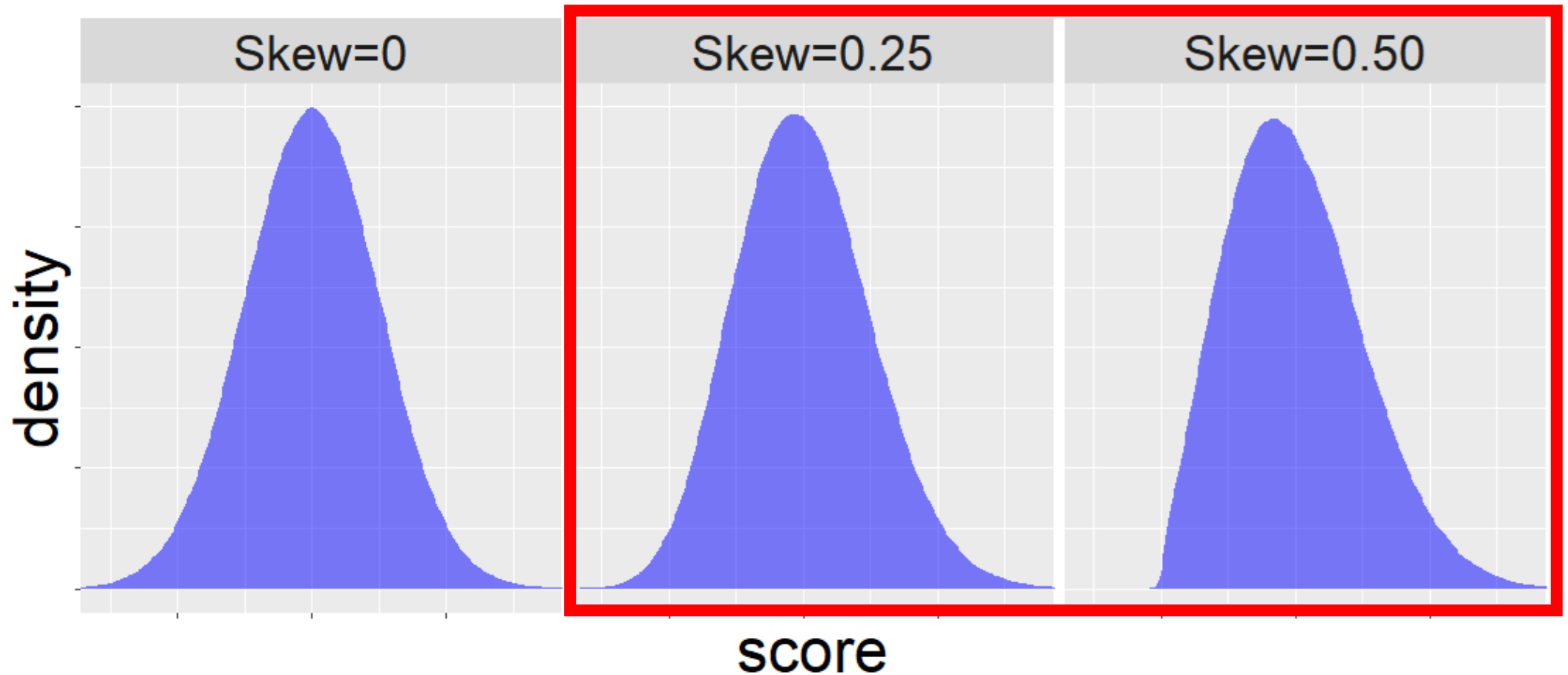


2° problema in psicologia: i *falsi positivi*

Metodi model-based
come il Gaussian
mixture model sono
estremamente sensibili
a violazione delle
assunzioni sulle
distribuzioni, tanto più
quanto più grandi sono
i campioni

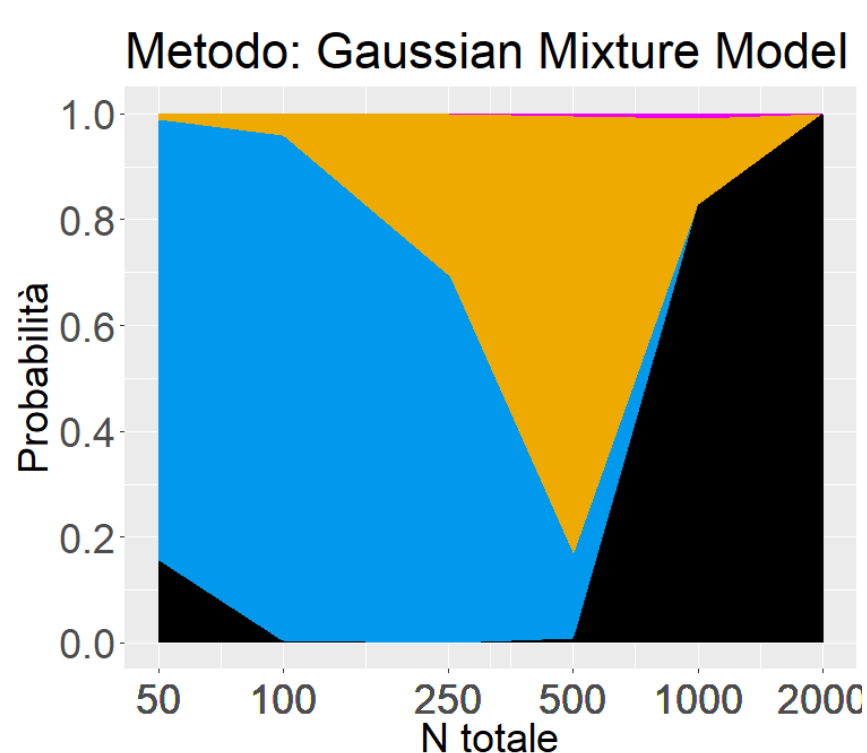


2° problema in psicologia: i *falsi positivi*



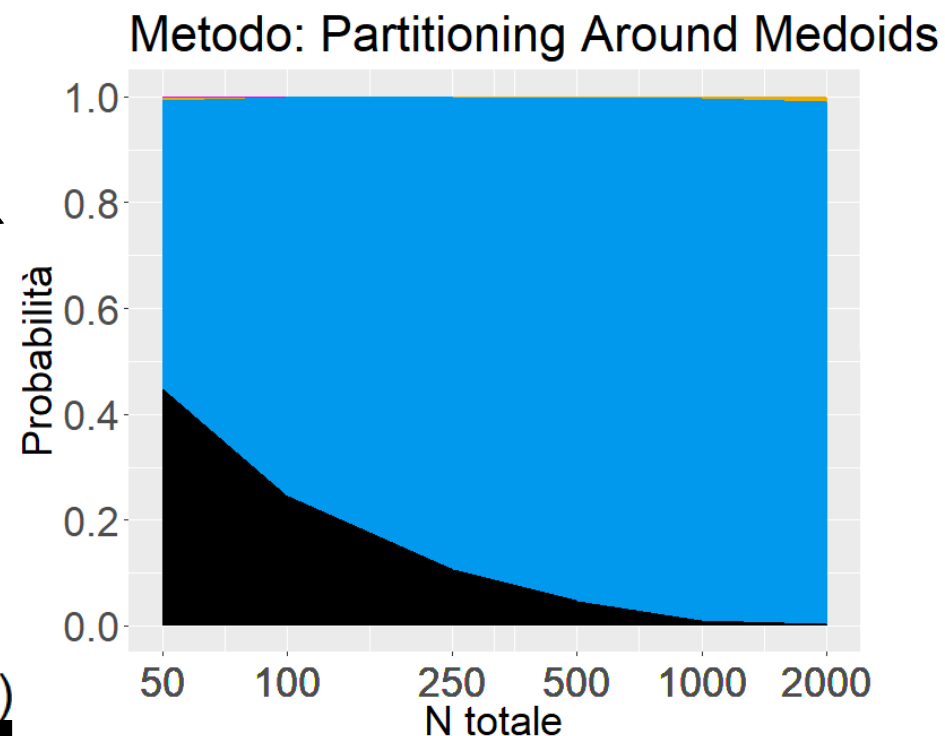
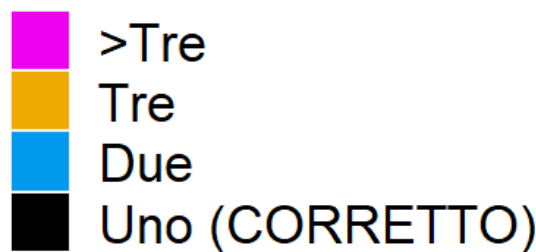
2° problema in psicologia: i *falsi positivi*

QUINDI violazione di assunzioni o talvolta anche solo combinazioni «sfortunate» di parametri portano a trovare più cluster di quanti ne esistano realmente



ES. dati simulati con
 $v = 12$ indicatori,
correlati mediamente
 $r = 0.20$
e UN SOLO VERO
CLUSTER

cluster rilevati



XXXI Congresso
21/23 sett 2023
Foggia

Enrico Toffalini, DPG, UNIPD
enrico.toffalini@unipd.it

In conclusione

- Metodi come *k-means*, *PAM* e *hierarchical* (ma anche model-based con N medio-piccoli) vi fregano quando gli indicatori sono correlati, anche debolmente... e in psicologia lo sono quasi sempre
- Metodi *model based* vi fregano se le distribuzioni non rispettano le assunzioni... e in psicologia succede quasi sempre
- I rischi di falsi positivi sono spesso tanto più gravi quanto maggiore è l'N! ... e comunque con effetti piccoli dovete avere N grandi, il che in psicologia vale quasi sempre



XXXI Congresso
21/23 sett 2023
| Foggia

Enrico Toffalini, DPG, UNIPD
enrico.toffalini@unipd.it

Siete i soliti disfattisti! Diteci cosa dobbiamo fare

- Prendete i vostri dati e valutate voi stessi *a priori* i rischi, applicando diversi metodi di clustering
- Come? Simuli dati sotto H_0 e sotto H_1 e valuti i rischi via Monte Carlo
- Cioè cosa simulo? Dati multivariati con una certa distribuzione (es. Gaussiana), N soggetti, v dimensioni, date correlazioni, skewness, curtosi, ... (molto utile `semTools::mvrnonnorm`)

Seguirà tutorial (*stay tuned*)



XXXI Congresso
21/23 sett 2023
| Foggia

Enrico Toffalini, DPG, UNIPD
enrico.toffalini@unipd.it



STATISTICS IS ART

Grazie per l'attenzione



ENRICO TOFFALINI
enrico.toffalini@unipd.it



FILIPPO GAMBAROTA
filippo.gambarota@unipd.it

con un grazie al team di lavoro: Paolo Girardi, Ambra Perugini, David Giofrè, Gianmarco Altoè



XXXI Congresso
21/23 sett 2023
| Foggia

Enrico Toffalini, DPG, UNIPD
enrico.toffalini@unipd.it