

Package ‘matchednull’

July 21, 2026

Title Matched-Null Tests for Cluster-Count Claims

Version 0.1.0

Description Builds Gaussian-copula matched nulls: synthetic twins of a dataset that preserve every marginal distribution and the full correlation matrix while containing no cluster structure by construction. A reported number of clusters or ``types" can then be tested against what the data's own margins and covariance already produce, using any clustering pipeline. Implements the matched-null procedure of Meng (2026) ``Types Without Taxa" <<https://osf.io/2ekcg>>.

License MIT + file LICENSE

Encoding UTF-8

RoxygenNote 7.3.3

Imports stats

Suggests mclust, testthat (>= 3.0.0), knitr, rmarkdown

VignetteBuilder knitr

Config/testthat/edition 3

URL <https://github.com/haomeng797-ship-it/matchednull>

BugReports <https://github.com/haomeng797-ship-it/matchednull/issues>

NeedsCompilation no

Author Miura Meng [aut, cre] (ORCID: <<https://orcid.org/0009-0004-1522-1997>>)

Maintainer Miura Meng <meng10@upenn.edu>

Repository CRAN

Date/Publication 2026-07-21 09:30:13 UTC

Contents

copula_null	2
matched_null_test	3
Index	5

copula_null

Draw one Gaussian-copula matched null ("null twin") of a dataset

Description

Builds a synthetic twin of x that preserves every marginal distribution exactly and the correlation matrix to within sampling error, while containing no cluster structure by construction: all dependence in the twin is Gaussian. Clustering found in real data but not in its twins must come from structure beyond the margins and covariance; clustering found equally in both was never more than the data's shape.

Usage

```
copula_null(x, ridge = 1e-06)
```

Arguments

<code>x</code>	A numeric matrix or data frame (rows = observations, columns = variables), complete cases only, at least two rows and two columns.
<code>ridge</code>	Small value added to the diagonal of the correlation matrix only if it is not positive definite (default 1e-6).

Details

The construction is a rank reordering in the tradition of Iman and Conover (1982): draw a Gaussian sample with the data's correlation matrix, then replace each column with the sorted real values laid down in the rank order of the Gaussian column. Every real value is reused exactly once per column, which is why the margins match exactly; the rank correlation is matched exactly and the Pearson correlation to within sampling error.

Value

A numeric matrix of the same dimensions as x : one matched-null draw. Each column contains exactly the values of the corresponding column of x , rearranged.

References

Iman, R. L., & Conover, W. J. (1982). A distribution-free approach to inducing rank correlation among input variables. *Communications in Statistics - Simulation and Computation*, 11(3), 311-334.

Examples

```
set.seed(1)
x <- matrix(rnorm(200 * 3), 200, 3) %%% chol(matrix(c(1, .5, .3,
                                                         .5, 1, .4,
                                                         .3, .4, 1), 3, 3))

twin <- copula_null(x)
```

```
# margins identical:
all(sort(twin[, 1]) == sort(x[, 1]))
# correlations close:
round(cor(x) - cor(twin), 2)
```

matched_null_test	<i>Test a cluster count against Gaussian-copula matched nulls</i>
-------------------	---

Description

Runs the user's own clustering pipeline on the real data and on R matched-null twins of it (see [copula_null\(\)](#)), and asks whether the real result exceeds what the twins, which share the data's margins and covariance but contain no cluster structure, produce. The pipeline is supplied as a function, so any procedure that returns a number, a BIC-selected mixture, a k-means heuristic, a published typology's own workflow, can be tested unchanged.

Usage

```
matched_null_test(
  x,
  cluster_fn,
  R = 200,
  probs = c(0.025, 0.975),
  ridge = 1e-06
)
```

Arguments

<code>x</code>	A numeric matrix or data frame, complete cases only.
<code>cluster_fn</code>	A function taking a data matrix and returning a single number: the statistic to be tested, typically the selected number of clusters (but any scalar summary of clustering strength works).
<code>R</code>	Number of matched-null twins to draw (default 200).
<code>probs</code>	Lower and upper quantiles of the null distribution used for the interval verdict (default <code>c(.025, .975)</code>).
<code>ridge</code>	Passed to copula_null() .

Value

An object of class "matched_null_test": a list with the real statistic (`real`), the null draws (`null`), the null interval (`interval`), a one-sided Monte Carlo p-value for exceedance (`p_exceed`), and the interval verdict (`within`). Reproducibility is the caller's: set a seed before calling.

Examples

```
if (requireNamespace("mclust", quietly = TRUE)) {  
  suppressPackageStartupMessages(library(mclust))  
  # a pipeline: how many components does BIC select?  
  pick_k <- function(d) Mclust(d, G = 1:5, verbose = FALSE)$G  
  set.seed(42)  
  x <- matrix(rnorm(500 * 4), 500, 4) # typeless data  
  matched_null_test(x, pick_k, R = 30)  
}
```

Index

`copula_null`, [2](#)

`copula_null()`, [3](#)

`matched_null_test`, [3](#)