Package 'PSinference'

July 13, 2026

Type Package

Title Inference for Released Plug-in Sampling Synthetic Dataset

Version 1.0.0

Maintainer Ricardo Moura <rp.moura@fct.unl.pt>

Description Considering the singly imputed synthetic data generated via plug-in sampling under the multivariate normal model, draws inference procedures including the generalized variance, the sphericity test, the test for independence between two subsets of variables, and the test for the regression of one set of variables on the other. For more details see Klein et al. (2021) <doi:10.1007/s13571-019-00215-9>.

License GPL-3

Encoding UTF-8

LazyData true

Depends R (>= 4.1.0)

Imports MASS

Suggests knitr, rmarkdown, testthat, ggplot2, methods

VignetteBuilder knitr

URL https://github.com/ricardomourarpm/PSinference

BugReports https://github.com/ricardomourarpm/PSinference/issues

Config/roxygen2/version 8.0.0

NeedsCompilation no

Author Vítor Augusto [aut] (ORCID: <https://orcid.org/0009-0004-9003-3558>),
Mina Norouzirad [aut] (ORCID: <https://orcid.org/0000-0003-0311-6888>),
Miguel Fonseca [ctb] (ORCID: <https://orcid.org/0000-0002-0162-8372>),
Ricardo Moura [aut, cre, cph] (ORCID:
<https://orcid.org/0000-0002-3003-9235>),
FCT, I.P. [fnd] (under the scope of the projects UID/00297/2025 and
UID/PRR/00297/2025 (NovaMath))

Repository CRAN

Date/Publication 2026-07-13 16:10:02 UTC

Contents

`brittany_soil_ps` *Brittany Loess-Soil Physicochemical Properties*

Description

Standardized physicochemical measurements from 37 cultivated fields on loess-silt parent material in Brittany, France, derived from Morvan et al. (2023). The dataset provides a realistic statistical disclosure control (SDC) example with $p = 6$ variables arranged in two natural blocks of size $p_1 = p_2 = 3$.

The first block contains standard agronomic indicators that are treated as public variables. The second block contains farm-management indicators that are treated as sensitive variables and are therefore suitable for illustrating plug-in sampling (PS) synthetic data releases.

Usage

```
brittany_soil_ps
```

Format

A numeric matrix with 37 rows and 6 columns. All variables are standardized to mean zero and standard deviation one:

pH_water Soil pH measured in water suspension.

pH_KCl Soil pH measured in 1 M KCl suspension.

log_CEC_Metson Log cation exchange capacity, Metson method (log meq per 100 g soil).

log_Organic_C Log soil organic carbon (log g/kg).

log_Total_N Log total soil nitrogen (log g/kg).

log_P_Olsen Log Olsen-P available phosphorus (log g P2O5/kg).

Details

The full dataset from Morvan et al. (2023) contains 137 fields across three parent-material types. This version retains only the most frequent parent-material group, loess-silt.

Variables with approximately log-normal distributions, namely cation exchange capacity, organic carbon, total nitrogen, and Olsen-P, were log-transformed. Mahalanobis outliers were removed using the $\chi^2_{6,0.975}$ cutoff. The retained variables were then standardized.

For illustrative SDC purposes, the variables are partitioned into two blocks:

Public block `pH_water`, `pH_KCl`, and `log_CEC_Metson`.

Sensitive block `log_Organic_C`, `log_Total_N`, and `log_P_Olsen`.

This block structure is useful for illustrating independence, regression, and covariance-based inference when one block is released as observed data and the other block is released through PS synthetic draws.

Normality was assessed using the diagnostics implemented in `mvn_test`. None of the retained diagnostics rejected multivariate normality at the 5% level.

Source

Morvan, T., Lambert, Y., Germain, P., Lemercier, B., Moreira, M., and Beff, L. (2023). A dataset of physico-chemical properties, extractable organic N, N mineralisation and physical organic matter fractionation of soils. *Data in Brief*, **51**, 109776. doi:10.1016/j.dib.2023.109776

Data repository, licensed under CC BY 4.0: doi:10.57745/DGIPGR

See Also

`mvn_test`, `simSynthData`, `ps_test`, `independence_test`, `regression_test`

Examples

```
data(brittany_soil_ps)

dim(brittany_soil_ps)
colnames(brittany_soil_ps)

# Public and sensitive blocks
public_block <- c("pH_water", "pH_KCl", "log_CEC_Metson")
sensitive_block <- c("log_Organic_C", "log_Total_N", "log_P_Olsen")


# Check multivariate normality
mvn_test(brittany_soil_ps, hz_nsim = 500, plot = FALSE)

# Generate three PS synthetic releases
set.seed(1)
V3 <- simSynthData(brittany_soil_ps, M = 3)

# Test independence between the public and sensitive blocks
independence_test(
  V3,
  M = 3,
  group_a = public_block,
  group_b = sensitive_block,
  iterations = 500L
)
```

`canodist` *Simulate the Canonical Regression Null Distribution*

Description

Simulates the null distribution of the regression pivotal statistic $T_4^\star$ under plug-in sampling for $M \geq 1$ releases.

The simulation uses the same compound Wishart structure as `Inddist`:

$$\Omega_1 \sim \mathcal{W}_p(n-1, I_p), \qquad (n-1)\Omega_2 \mid \Omega_1 \sim \mathcal{W}_p(Mn-1, \Omega_1).$$

Usage

```
canodist(part, nsample, pvariates, iterations = 10000L, M = 1L)
```

Arguments

`part`	Size of the first variable block, p_1. Must satisfy $p_1 \leq p_2$.
`nsample`	Original sample size n.

`pvariates`	Total number of variables, p.
`iterations`	Number of Monte Carlo draws. The default is `10000L`.
`M`	Number of synthetic releases. The default is `1L`.

Value

A numeric vector of length `iterations`.

See Also

`regression_test`

Examples

```
set.seed(1)


nd1 <- canodist(
  part = 2, nsample = 50, pvariates = 4,
  M = 1, iterations = 1000L
)
stats::quantile(nd1, probs = 0.95)

nd5 <- canodist(
  part = 2, nsample = 50, pvariates = 4,
  M = 5, iterations = 1000L
)
stats::quantile(nd5, probs = 0.95)

canodist(part = 2, nsample = 50, pvariates = 4, M = 5) |> quantile(0.95)
```

`GVdist` *Simulate the Generalized Variance Null Distribution*

Description

Simulates the null distribution of the generalized variance pivotal statistic $T_1^\star$ under plug-in sampling.

Under the multiple-release stacking result,

$$(\mathbf{n}-\mathbf{1})\mathbf{S}_{\mathrm{M}}^{\star} \mid \mathbf{S} \sim \mathcal{W}_p\left(Mn-1,\ \mathbf{S}\right),$$

the Bartlett decomposition gives

$$T_1^\star = (n-1)^p \frac{|\mathbf{S}^\star|}{|\Sigma|} \ \overset{d}{=}\ \left(\prod_{j=1}^{p} A_j\right)\left(\prod_{j=1}^{p} B_j\right),$$

where $A_j \sim \chi^2_{Mn-j}$ comes from the synthetic Wishart distribution with degrees of freedom $Mn-1$, and $B_j \sim \chi^2_{n-j}$ comes from the original-sample Wishart distribution with degrees of freedom $n-1$. All $2p$ variables are mutually independent.

For $M = 1$, A_j and B_j have the same χ^2_{n-j} distribution, recovering the single-release result of Klein et al. (2021).

Usage

```
GVdist(nsample, pvariates, iterations = 10000L, M = 1L)
```

Arguments

`nsample`	Original sample size n.
`pvariates`	Number of variables p.
`iterations`	Number of Monte Carlo draws. The default is `10000L`.
`M`	Number of synthetic releases. The default is `1L`. The effective sample size is $N = Mn$.

Value

A numeric vector of length `iterations` containing draws from the null distribution of $T_1^\star$.

References

Klein, M., Moura, R., and Sinha, B. (2021). Multivariate normal inference based on singly imputed synthetic data under plug-in sampling. *Sankhya B*, **83**, 273–287. doi:10.1007/s13571019002159

See Also

`gv_test`

Examples

```
set.seed(1)


# Single release
nd1 <- GVdist(nsample = 50, pvariates = 4, M = 1, iterations = 1000L)
stats::quantile(nd1, probs = c(0.025, 0.975))

# Five releases
nd5 <- GVdist(nsample = 50, pvariates = 4, M = 5, iterations = 1000L)
stats::quantile(nd5, probs = c(0.025, 0.975))
```

gv_ci *Generalized Variance Confidence Interval*

Description

Backward-compatible alias for gv_test.

Usage

```
gv_ci(V, M = 1L, Sigma, alpha = 0.05, iterations = 10000L, null_dist = NULL)
```

Arguments

V	Stacked synthetic data set, given as an $Mn \times p$ numeric matrix, as returned by simSynthData.
M	Positive integer giving the number of synthetic releases. The default is 1L.
Sigma	A $p \times p$ positive-definite matrix specifying the null covariance matrix Σ_0. Typically, this is cov(X), where X is the original data matrix.
alpha	Significance level. The default is 0.05.
iterations	Monte Carlo sample size used to approximate the null distribution. The default is 10000L.
null_dist	Optional numeric vector containing a precomputed null distribution. If supplied, iterations is ignored.

gv_test *Generalized Variance Test and Confidence Interval*

Description

Tests $H_0 : |\Sigma| = |\Sigma_0|$ and computes a $(1-\alpha)$-level confidence interval for the generalized variance $|\Sigma|$, based on M released plug-in sampling synthetic data sets stacked into V. Setting M = 1 recovers the single-release procedure of Klein et al. (2021).

Usage

```
gv_test(V, M = 1L, Sigma, alpha = 0.05, iterations = 10000L, null_dist = NULL)
```

Arguments

V	Stacked synthetic data set, given as an $Mn \times p$ numeric matrix, as returned by simSynthData.
M	Positive integer giving the number of synthetic releases. The default is 1L.
Sigma	A $p \times p$ positive-definite matrix specifying the null covariance matrix Σ_0. Typically, this is cov(X), where X is the original data matrix.
alpha	Significance level. The default is 0.05.
iterations	Monte Carlo sample size used to approximate the null distribution. The default is 10000L.
null_dist	Optional numeric vector containing a precomputed null distribution. If supplied, iterations is ignored.

Value

An object of class ps_test with component conf.int giving the exact $(1-\alpha)$ confidence interval for $|\Sigma|$. The usual S3 methods, including print, summary, and plot, are available.

References

Klein, M., Moura, R., and Sinha, B. (2021). Multivariate normal inference based on singly imputed synthetic data under plug-in sampling. *Sankhya B*, **83**, 273–287. doi:10.1007/s13571019002159

See Also

GVdist, ps_test, simSynthData

Examples

```
data(brittany_soil_ps)

set.seed(1)
V1 <- simSynthData(brittany_soil_ps)


res <- gv_test(V1,
  M = 1, Sigma = cov(brittany_soil_ps),
  iterations = 1000L
)
print(res)
plot(res)

set.seed(1)
V5 <- simSynthData(brittany_soil_ps, M = 5)
res5 <- gv_test(V5,
  M = 5, Sigma = cov(brittany_soil_ps),
  iterations = 1000L
)

print(res5)
```

```
plot(res5)
```

`Inddist` *Simulate the Independence Null Distribution*

Description

Simulates the null distribution of the independence pivotal statistic $T_3^\star$ under plug-in sampling for $M \geq 1$ releases.

Under the stacking result, the compound Wishart representation gives

$$T_3^\star \overset{d}{=} \frac{|\Omega_2|}{|\Omega_{2,11}||\Omega_{2,22}|},$$

where $\Omega_1 \sim \mathcal{W}_p(n-1, I_p)$ and $(n-1)\Omega_2 \mid \Omega_1 \sim \mathcal{W}_p(Mn-1,\, \Omega_1)$.

Usage

```
Inddist(part, nsample, pvariates, iterations = 10000L, M = 1L)
```

Arguments

`part`	Size of the first variable block, p_1.
`nsample`	Original sample size n.
`pvariates`	Total number of variables, p.
`iterations`	Number of Monte Carlo draws. The default is `10000L`.
`M`	Number of synthetic releases. The default is `1L`.

Value

A numeric vector of length `iterations`.

See Also

`independence_test`

Examples

```
set.seed(1)


nd1 <- Inddist(
  part = 2, nsample = 50, pvariates = 4,
  M = 1, iterations = 1000L
)
stats::quantile(nd1, probs = 0.05)
```

```
nd5 <- Inddist(
  part = 2, nsample = 50, pvariates = 4,
  M = 5, iterations = 1000L
)
stats::quantile(nd5, probs = 0.05)
```

`independence_test` *Independence Test*

Description

Tests $H_0 : \Sigma_{12} = \mathbf{0}$, that is, independence between two subsets of variables, based on M released plug-in sampling synthetic data sets stacked into `V`. Setting `M = 1` recovers the single-release procedure of Klein et al. (2021).

The two variable blocks can be specified in exactly one of two ways:

`part` An integer scalar. The first `part` columns form Block 1, and the remaining columns form Block 2. This is the original backward-compatible interface.

`group_a` **and** `group_b` Integer indices or column names identifying the two blocks. Together, they must cover all columns of `V` exactly once. If names are used, `V` must have column names.

Usage

```
independence_test(
  V,
  M = 1L,
  part = NULL,
  group_a = NULL,
  group_b = NULL,
  alpha = 0.05,
  iterations = 10000L,
  null_dist = NULL
)
```

Arguments

`V`	Stacked synthetic data set, given as an $Mn \times p$ numeric matrix.
`M`	Positive integer giving the number of synthetic releases. The default is `1L`.
`part`	Integer scalar giving the size of Block 1. The first `part` columns form Block 1, and the remaining columns form Block 2. Ignored when `group_a` and `group_b` are supplied.
`group_a`	Integer indices or column names identifying Block 1.
`group_b`	Integer indices or column names identifying Block 2. Together with `group_a`, these must cover all columns.
`alpha`	Significance level. The default is `0.05`.

`iterations`	Monte Carlo sample size used to approximate the null distribution. The default is `10000L`.
`null_dist`	Optional numeric vector containing a precomputed null distribution. If supplied, `iterations` is ignored.

Value

An object of class `ps_test`. The null hypothesis string in `$null.value` names the two blocks explicitly.

References

Klein, M., Moura, R., and Sinha, B. (2021). Multivariate normal inference based on singly imputed synthetic data under plug-in sampling. *Sankhya B*, **83**, 273–287. doi:10.1007/s13571019002159

See Also

`Inddist`, `ps_test`

Examples

```
data(brittany_soil_ps)

set.seed(1)
V5 <- simSynthData(brittany_soil_ps, M = 5)


# Integer interface
independence_test(V5, M = 5, part = 2L, iterations = 1000L)

# Named interface
independence_test(
  V5,
  M = 5,
  group_a = c("pH_water", "pH_KCl", "log_CEC_Metson"),
  group_b = c("log_Organic_C", "log_Total_N", "log_P_Olsen"),
  iterations = 1000L
)
```

`is.ps_test` *Test Whether an Object Has Class* `ps_test`

Description

Checks whether an object inherits from class `ps_test`.

Usage

```
is.ps_test(x)
```

Arguments

x	Any R object.

Value

A logical value: TRUE if x inherits from class ps_test, and FALSE otherwise.

Examples

```
data(brittany_soil_ps)

set.seed(1)
V <- simSynthData(brittany_soil_ps, M = 3)


res <- sphericity_test(V, M = 3, iterations = 1000L)
is.ps_test(res)
```

---

mvn_test *Multivariate Normality Assessment*

Description

Assesses multivariate normality of a data set using five complementary approaches: (1) univariate Shapiro-Wilk tests on each variable, (2) Mardia's multivariate skewness test, (3) Mardia's multivariate kurtosis test, (4) the Henze-Zirkler omnibus test, and (5) Royston's multivariate extension of the Shapiro-Wilk test. A visual diagnostic panel shows one histogram with a fitted normal curve for each variable and a chi-square Q-Q plot of squared Mahalanobis distances.

Usage

```
mvn_test(X, alpha = 0.05, plot = TRUE, hz_nsim = 2000L, verbose = TRUE)
```

Arguments

X	A numeric matrix or data frame with dimension $n \times p$.
alpha	Significance level for all tests. The default is 0.05.
plot	Logical. If TRUE, the default, a diagnostic plot panel is produced.
hz_nsim	Integer. Number of Monte Carlo draws used to calibrate the Henze-Zirkler null distribution. The default is 2000L. Increasing this value gives more accurate p-values at the cost of additional computation time.
verbose	Logical. If TRUE, the default, a formatted summary of all test results is printed.

Details

Mardia's skewness test evaluates the null hypothesis of zero multivariate skewness:

$$\kappa = \frac{n}{6} b_{1,p} \sim \chi^2\left(\frac{p(p+1)(p+2)}{6}\right), \qquad b_{1,p} = \frac{1}{n^2}\sum_{a=1}^{n}\sum_{b=1}^{n} d_{ab}^3.$$

Here

$$d_{ab} = (\boldsymbol{x}_a - \bar{\boldsymbol{x}})' S^{-1} (\boldsymbol{x}_b - \bar{\boldsymbol{x}})$$

is a Mahalanobis inner product between observations a and b. The indices a and b run over observations, not variables.

Mardia's kurtosis test evaluates whether the multivariate kurtosis equals $p(p+2)$:

$$z = \frac{b_{2,p} - p(p+2)}{\sqrt{8p(p+2)/n}} \sim N(0,1), \qquad b_{2,p} = \frac{1}{n}\sum_{a=1}^{n} d_{aa}^2.$$

The quantity d_{aa} is the squared Mahalanobis distance of observation a from the sample mean.

The Henze-Zirkler omnibus test is based on a weighted L^2 distance between the empirical and theoretical multivariate normal characteristic functions. The statistic is

$$\mathrm{HZ} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{-\frac{\beta^2}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_S^2} - 2(1+\beta^2)^{-p/2}\frac{1}{n}\sum_{i=1}^{n} e^{-\frac{\beta^2}{2(1+\beta^2)} d_i^2} + (1+2\beta^2)^{-p/2},$$

where

$$\beta = \frac{1}{\sqrt{2}}\left(\frac{2p+1}{4}\right)^{1/(p+4)} n^{1/(p+4)}.$$

The null distribution of HZ is approximated by a log-normal distribution whose parameters are estimated by Monte Carlo simulation of size `hz_nsim` from $\mathcal{N}_p(\mathbf{0}, \boldsymbol{I}_p)$. This test is particularly powerful useful against heavy-tailed and skewed alternatives.

Royston's H test extends the univariate Shapiro-Wilk statistic to the multivariate setting. For each variable, the Shapiro-Wilk p-value p_j is transformed to $Z_j = \Phi^{-1}(1 - p_j)$. The test statistic is

$$H = e^{-1}\sum_{j=1}^{p} Z_j^2 \sim \chi_e^2,$$

where $e = p/\left[1 + (p-1)\hat{\rho}_z\right]$ is an effective degree of freedom parameter that accounts for correlation among the Z_j values. The quantity $\hat{\rho}_z$ estimated from the average squared pairwise correlation $\hat{\rho}_z$ of the original variables.

Diagnostic panel: The diagnostic panel contains $p+1$ plots arranged in a grid. The first p panels show histograms with fitted $\mathcal{N}(\bar{x}_j, s_j^2)$ density curves. The bar color is steel-blue when the Shapiro-Wilk test fails to reject normality and tomato-red when it rejects. The final panel shows the chi-square Q-Q plot of squared Mahalanobis distances.

Value

A list of class `mvn_test`, returned invisibly, with components:

shapiro Data frame of per-variable Shapiro-Wilk statistics and p-values.

mardia_skewness Named list with components `statistic`, `df`, `p.value`, and `decision`.

mardia_kurtosis Named list with components `statistic`, `p.value`, `decision`.

henze_zirkler Named list Named list with components `statistic`, `p.value`, `decision`.

royston Named list Named list with components `statistic`, the Royston H statistic; `df`, the effective degrees of freedom; `p.value`, and `decision`.

mahal_distances Numeric vector of squared Mahalanobis distances.

overall Character string giving the overall conclusion based on all tests.

References

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, **57**, 519–530.

Henze, N. and Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics: Theory and Methods*, **19**, 3595–3617.

Royston, J. P. (1992). Approximating the Shapiro-Wilk W test for non-normality. *Statistics and Computing*, **2**, 117–119.

See Also

`simSynthData`, `brittany_soil_ps`,

Examples

```
data(brittany_soil_ps)
mvn_test(brittany_soil_ps)
```

`original_gv_test` *Classical Generalized Variance Test (Original Data)*

Description

Tests $H_0 : |\Sigma| = |\Sigma_0|$ using the original data matrix `X`. The null value `Sigma0` is typically supplied by the user.

If `Sigma0 = cov(X)`, the null determinant is estimated from the same data used to compute the test statistic. In that case, the determinant ratio is one, the chi-square statistic is zero, and the p-value is one by construction.

Supplying a different `Sigma0` gives a likelihood-ratio chi-square test with Bartlett correction.

The test statistic is

$$\chi^2 = -\Big[(n-1) - \tfrac{2p^2+p+2}{6p}\Big] \log\Big(\tfrac{|\hat{\Sigma}|}{|\Sigma_0|}\Big),$$

referred to a χ^2 distribution with $\frac{1}{2}p(p+1) - 1$ degrees of freedom.

Usage

```
original_gv_test(X, Sigma0 = NULL, alpha = 0.05)
```

Arguments

X	Original data matrix with dimension $n \times p$.
Sigma0	A $p \times p$ positive-definite null covariance matrix. The default is cov(X).
alpha	Significance level. The default is 0.05.

Value

A list with components:

statistic Observed χ^2 statistic.

p.value p-value.

df Degrees of freedom, $p(p+1)/2 - 1$.

det.Sigma.hat Value of $|\hat{\Sigma}|$.

decision Character string: "Reject H0" or "Fail to reject H0".

alpha Significance level used.

n, p Sample size and number of variables.

References

Anderson, T. W. (1984). An Introduction to Multivariate Statistical Analysis (2nd ed.). John Wiley & Sons, New York.

See Also

gv_test

Examples

```
data(brittany_soil_ps)
X <- brittany_soil_ps

## Null = MLE  =>  p-value = 1 by construction
original_gv_test(X)

## Test against a specific null
Sigma0 <- diag(ncol(X))
original_gv_test(X, Sigma0 = Sigma0)
```

`original_independence_test` *Classical Independence Test (Original Data)*

Description

Tests $H_0 : \Sigma_{12} = \mathbf{0}$, corresponding to block independence, using Bartlett's factored-likelihood chi-square approximation applied to the original data matrix `X`.

The Wilks statistic is

$$\Lambda = \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_{11}||\hat{\Sigma}_{22}|},$$

and the test statistic is

$$\chi^2 = -\Big[(n-1) - \tfrac{p+3}{2}\Big] \log \Lambda,$$

referred to a χ^2 distribution with $p_1 p_2$ degrees of freedom.

Usage

```
original_independence_test(
  X,
  part = NULL,
  group_a = NULL,
  group_b = NULL,
  alpha = 0.05
)
```

Arguments

`X`	Original data matrix with dimension $n \times p$.
`part`	Integer scalar. The first `part` columns form Block 1. Ignored when `group_a` and `group_b` are supplied.
`group_a`	Integer indices or column names identifying Block 1.
`group_b`	Integer indices or column names identifying Block 2.
`alpha`	Significance level. The default is `0.05`.

Value

A list with components `statistic` (χ^2), `p.value`, `df`, `Lambda` (Wilks statistic), `decision`, `alpha`, `n`, `p`, `p1`, `p2`, `lbl1`, and `lbl2`.

References

Anderson, T. W. (1984). An Introduction to Multivariate Statistical Analysis (2nd ed.). John Wiley & Sons, New York.

See Also

independence_test

Examples

```
data(brittany_soil_ps)
original_independence_test(brittany_soil_ps,
  group_a = c("pH_water", "pH_KCl"),
  group_b = c(
    "log_CEC_Metson", "log_Organic_C",
    "log_Total_N", "log_P_Olsen"
  )
)
```

---

original_regression_test

Classical Regression Test (Original Data)

Description

Tests $H_0 : \Delta = \Delta_0$ for the population regression matrix $\Delta = \Sigma_{12}\Sigma_{22}^{-1}$, using Rao's F-approximation to Wilks' Λ statistic applied to the original data matrix `X`.

The Wilks statistic is

$$\Lambda = \frac{|\hat{\Sigma}_{11.2(\Delta_0)}|}{|\hat{\Sigma}_{11}|},$$

where

$$\hat{\Sigma}_{11.2(\Delta_0)} = \hat{\Sigma}_{11} - (\hat{\Delta} - \Delta_0)\hat{\Sigma}_{22}(\hat{\Delta} - \Delta_0)^\top$$

is the residual Schur complement under H_0. This equals the ordinary Schur complement $\hat{\Sigma}_{11.2}$ when $\Delta_0 = \hat{\Delta}$, giving $\Lambda = 1$ and $F = 0$ by construction. The default `Delta0 = NULL` tests $H_0 : \Delta = 0$, corresponding to zero regression.

Usage

```
original_regression_test(
  X,
  part = NULL,
  Delta0 = NULL,
  response = NULL,
  predictors = NULL,
  alpha = 0.05
)
```

Arguments

X	Original data matrix with dimension $n \times p$.
part	Integer scalar giving the size of the response block (Block 1). The first part columns form the response block. Ignored when response and predictors are supplied. Must satisfy $p_1 \leq p_2$.
Delta0	A $p_1 \times p_2$ null regression matrix. The default NULL sets $\Delta_0 = 0$.
response	Integer or character vector identifying the response block.
predictors	Integer or character vector identifying the predictor block.
alpha	Significance level. The default is 0.05.

Value

A list with components statistic (F), p.value, df1, df2, Lambda, Delta.hat ($\hat{\Delta} = \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}$), decision, alpha, n, p, p1, p2, lbl1, and lbl2.

References

Anderson, T. W. (1984). An Introduction to Multivariate Statistical Analysis (2nd ed.). John Wiley & Sons, New York.

Rao, C. R. (1951). An asymptotic expansion of the distribution of Wilks' criterion. *Bulletin of the International Statistical Institute*, **33**, 177–180.

See Also

regression_test

Examples

```
data(brittany_soil_ps)
X <- brittany_soil_ps

## Test H0: Delta = 0 (default)
original_regression_test(X,
  response = c("pH_water", "pH_KCl"),
  predictors = c(
    "log_CEC_Metson", "log_Organic_C",
    "log_Total_N", "log_P_Olsen"
  )
)

## Test H0: Delta = Delta_hat (MLE) => F = 0, p = 1 by construction
blk <- partition(cov(X), part1 = c("pH_water", "pH_KCl"))
Delta0 <- blk$B %*% solve(blk$D)
original_regression_test(X,
  response = c("pH_water", "pH_KCl"),
  predictors = c(
    "log_CEC_Metson", "log_Organic_C",
    "log_Total_N", "log_P_Olsen"
  ),
```

```
  Delta0 = Delta0
)
```

`original_sphericity_test` *Classical Sphericity Test (Original Data)*

Description

Tests $H_0 : \Sigma = \sigma^2 I_p$ using the Bartlett-Box chi-square approximation applied to the original data matrix `X`.

The Mauchly statistic is

$$W = \frac{|\hat{\Sigma}|}{(\mathrm{tr}(\hat{\Sigma})/p)^p},$$

and the test statistic is

$$\chi^2 = -\left[(n-1) - \tfrac{2p^2+p+2}{6p}\right] \log W,$$

referred to a χ^2 distribution with $\frac{1}{2}p(p+1) - 1$ degrees of freedom.

Usage

```
original_sphericity_test(X, alpha = 0.05)
```

Arguments

`X`	Original data matrix with dimension $n \times p$.
`alpha`	Significance level. The default is `0.05`.

Value

A list with components `statistic` (χ^2), `p.value`, `df`, `W` (Mauchly statistic), `sigma2.hat` (plug-in estimate $\hat{\sigma}^2 = \mathrm{tr}(\hat{\Sigma})/p$), `decision`, `alpha`, `n`, and `p`.

References

Anderson, T. W. (1984). An Introduction to Multivariate Statistical Analysis (2nd ed.). John Wiley & Sons, New York.

Bartlett, M. S. (1954). A note on the multiplying factors for various chi-square approximations. *Journal of the Royal Statistical Society: Series B*, **16**, 296–298.

See Also

`sphericity_test`

Examples

```
data(brittany_soil_ps)
original_sphericity_test(brittany_soil_ps)
```

`partition` *Partition a Matrix into Four Blocks*

Description

Splits a numeric matrix $\mathbf{M}$ into four submatrices according to a two-group partition of its rows and columns:

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix}.$$

The blocks are:

A Rows in the first group and columns in the first group.

B Rows in the first group and columns in the second group.

C Rows in the second group and columns in the first group.

D Rows in the second group and columns in the second group.

Two interfaces are available:

Integer interface Supply `nrows` and `ncols`. The first `nrows` rows and the first `ncols` columns form the first block.

Name interface Supply `part1` as a character vector of row and column names. This interface is intended for square matrices, such as covariance or correlation matrices. The matrix is reordered so that the `part1` rows and columns appear first. The optional `part2` argument is checked against the complement of `part1`.

Usage

```
partition(Matrix, nrows = NULL, ncols = NULL, part1 = NULL, part2 = NULL)
```

Arguments

`Matrix`	A numeric matrix. When using the name interface, it must be a square matrix with matching row and column names.
`nrows`	Integer giving the number of rows in the first row block. Ignored when `part1` is supplied.
`ncols`	Integer giving the number of columns in the first column block. Ignored when `part1` is supplied.
`part1`	Character vector of names forming the first group. Used for both row and column reordering in the name interface.
`part2`	Optional character vector naming the second group. If supplied, it is checked against the complement of `part1`.

Value

A named list of class `"ps_partition"` with elements `A`, `B`, `C`, and `D`. Numeric indexing with `[[1]]` through `[[4]]` also works.

Examples

```
M <- matrix(
  1:16,
  4,
  4,
  dimnames = list(c("A", "B", "C", "D"), c("A", "B", "C", "D"))
)

# Integer interface
b <- partition(M, nrows = 2, ncols = 2)
b$A

# Name interface
b2 <- partition(M, part1 = c("A", "B"))
b2$A
b2$D

# Covariance matrix example
data(brittany_soil_ps)
b3 <- partition(
  cov(brittany_soil_ps),
  part1 = c("log_Organic_C", "log_Total_N", "log_P_Olsen"),
  part2 = c("pH_water", "pH_KCl", "log_CEC_Metson")
)
b3$A
b3$D
```

---

plot.mvn_test *Plot Method for* mvn_test *Objects*

Description

Re-draws the chi-square Q-Q diagnostic. For the full histogram panel, call `mvn_test(X, plot = TRUE)` on the original data directly.

Usage

```
## S3 method for class 'mvn_test'
plot(x, ...)
```

Arguments

x	An object of class mvn_test.
...	Further arguments (currently ignored).

Value

Invisibly returns x.

`plot.ps_test` *Plot a* `ps_test` *Object*

Description

Produces a density plot of the simulated null distribution with the observed test statistic and critical value(s) marked. The rejection region is shaded.

The x-axis always includes both the null distribution and the observed statistic. For the generalized variance and regression tests, a log10 scale is used automatically because these statistics may span several orders of magnitude. Key information is placed below the plot as text so it does not overlap the density curve. The function is multi-panel aware: inside `par(mfrow = ...)`, it uses compact in-plot annotations and does not modify the outer margins.

Usage

```
## S3 method for class 'ps_test'
plot(
  x,
  main = NULL,
  shade_col = grDevices::adjustcolor("tomato", 0.45),
  dist_col = grDevices::adjustcolor("steelblue", 0.22),
  stat_col = "firebrick",
  crit_col = "steelblue4",
  ...
)
```

Arguments

`x`	An object of class `ps_test`.
`main`	Optional title string. If `NULL`, a title is generated automatically.
`shade_col`	Color for the rejection-region shading.
`dist_col`	Color for the null-distribution density fill.
`stat_col`	Color for the observed-statistic line.
`crit_col`	Color for the critical-value line(s).
`...`	Further arguments passed to `plot()`.

Value

Invisibly returns `x`.

Examples

```
data(brittany_soil_ps)
V <- simSynthData(brittany_soil_ps, M = 3)
plot(sphericity_test(V, M = 3))
```

print.mvn_test *Print Method for* mvn_test *Objects*

Description

Print the result of the test.

Usage

```
## S3 method for class 'mvn_test'
print(x, ...)
```

Arguments

x	An object of class mvn_test.
...	Further arguments (currently ignored).

Value

Invisibly returns x.

print.original_test *Print an* original_test *Object*

Description

Prints a concise summary of a classical test result.

Usage

```
## S3 method for class 'original_test'
print(x, ...)
```

Arguments

x	An object of class original_test.
...	Further arguments, currently ignored.

Value

Invisibly returns x.

print.ps_test *Print a* ps_test *Object*

Description

Prints a concise, human-readable summary of the test result stored in a ps_test object.

Usage

```
## S3 method for class 'ps_test'
print(x, ...)
```

Arguments

x	An object of class ps_test.
...	Further arguments, currently ignored.

Value

Invisibly returns x.

Examples

```
data(brittany_soil_ps)

set.seed(1)
V <- simSynthData(brittany_soil_ps, M = 3)


res <- sphericity_test(V, M = 3, iterations = 1000L)
print(res)
```

print.ps_utility *Print Method for* ps_utility *Objects*

Description

Prints a formatted summary of the five headline utility measures returned by utility_measures.

Usage

```
## S3 method for class 'ps_utility'
print(x, ...)
```

Arguments

x	An object of class ps_utility.
...	Further arguments, currently ignored.

Value

Invisibly returns x.

---

ps_test *Unified Wrapper for PS Inference Tests*

Description

Dispatches to the appropriate exact inferential procedure based on the test argument and can optionally produce the diagnostic plot immediately. This is the main entry point for users who prefer a single function instead of calling the four individual test functions directly.

Usage

```
ps_test(
  V,
  M = 1L,
  test = c("gv", "sphericity", "independence", "regression"),
  plot = FALSE,
  ...
)
```

Arguments

V	Stacked synthetic data set, given as an $Mn \times p$ numeric matrix, as returned by simSynthData.
M	Positive integer giving the number of synthetic releases. The default is 1L. Setting M = 1 recovers the single-release procedures of Klein et al. (2021).
test	Character string specifying the test. One of "gv" for generalized variance, "sphericity", "independence", or "regression".
plot	Logical. If TRUE, plot() is called on the result before it is returned, so the null-distribution diagnostic is displayed automatically. The default is FALSE.
...	Additional arguments passed to the corresponding test function or, when plot = TRUE, to plot.ps_test(). Arguments intended for the test function, such as part, Sigma, Delta0, alpha, or iterations, and graphical arguments, such as main, shade_col, dist_col, stat_col, and crit_col, are separated automatically.

Value

An object of class ps_test-class, invisibly when plot = TRUE.

See Also

gv_test, sphericity_test, independence_test, regression_test

Examples

```
data(brittany_soil_ps)

set.seed(1)
V <- simSynthData(brittany_soil_ps, M = 3)


# Run and print only
ps_test(V, M = 3, test = "sphericity", iterations = 1000L)

# Run and plot in one call
ps_test(V,
  M = 3, test = "sphericity",
  iterations = 1000L, plot = TRUE
)

# Independence with named blocks
ps_test(
  V,
  M = 3,
  test = "independence",
  group_a = c("log_Organic_C", "log_Total_N", "log_P_Olsen"),
  group_b = c("pH_water", "pH_KCl", "log_CEC_Metson"),
  iterations = 1000L,
  plot = TRUE
)

# Generalized variance with a reference covariance matrix
ps_test(
  V,
  M = 3,
  test = "gv",
  Sigma = cov(brittany_soil_ps),
  iterations = 1000L,
  plot = TRUE
)
```

---

ps_test-class *S3 Class for PS Inference Test Results*

Description

The ps_test class is the unified output object returned by the inferential functions in **PSinference**. It stores the test result, the simulated null distribution, and relevant metadata, and provides print, summary, and plot methods for convenient inspection and reporting.

Slots

statistic Numeric. Observed value of the test statistic.

p.value Numeric. Monte Carlo p-value.

alpha Numeric. Significance level used.

decision Character. `"Reject H0"` or `"Fail to reject H0"`.

null.dist Numeric vector. Simulated null distribution.

test Character. One of `"gv"`, `"sphericity"`, `"independence"`, `"regression"`.

n Integer. Original sample size.

M Integer. Number of synthetic releases.

N Integer. Effective sample size $N = Mn$.

p Integer. Number of variables.

conf.int Numeric vector of length 2 or `NULL`. Confidence interval (generalized variance only).

sigma2.hat Numeric or `NULL`. Plug-in estimator of σ^2 (sphericity only).

Delta.hat Matrix or `NULL`. Plug-in estimator of Δ, used for the regression test.

lbl1 Character or `NULL`. Label for the first variable block, used by block-based tests.

lbl2 Character or `NULL`. Label for the second variable block, used by block-based tests.

iterations Integer. Number of Monte Carlo iterations used to calibrate the null distribution.

`regression_test` *Regression Test*

Description

Tests $H_0 : \Delta = \Delta_0$ for the population regression matrix $\Delta = \Sigma_{12}\Sigma_{22}^{-1}$, based on M released plug-in sampling synthetic data sets stacked into `V`. The test requires $p_1 \leq p_2$. Setting `M = 1` recovers the single-release procedure of Klein et al. (2021).

The two variable blocks can be specified in exactly one of two ways:

`part` An integer scalar. The first `part` columns form the response block, and the remaining columns form the predictor block. This is the original backward-compatible interface.

`response` **and** `predictors` Integer indices or column names identifying the response and predictor blocks. Together, they must cover all columns of `V` exactly once.

Usage

```
regression_test(
  V,
  M = 1L,
  part = NULL,
  Delta0 = NULL,
  response = NULL,
```

```
  predictors = NULL,
  alpha = 0.05,
  iterations = 10000L,
  null_dist = NULL
)
```

Arguments

V	Stacked synthetic data set, given as an $Mn \times p$ numeric matrix.
M	Positive integer giving the number of synthetic releases. The default is 1L.
part	Integer scalar giving the size of the response block. The first part columns form the response block, and the remaining columns form the predictor block. Must satisfy $p_1 \leq p_2$. Ignored when response and predictors are supplied.
Delta0	A $p_1 \times p_2$ matrix giving the null value Δ_0. The default is the zero matrix, corresponding to a test of zero regression.
response	Integer or character vector identifying the response block.
predictors	Integer or character vector identifying the predictor block. Together with response, these must cover all columns.
alpha	Significance level. The default is 0.05.
iterations	Monte Carlo sample size used to approximate the null distribution. The default is 10000L.
null_dist	Optional numeric vector containing a precomputed null distribution. If supplied, iterations is ignored.

Value

An object of class ps_test. Component Delta.hat gives the plug-in slope estimator $\hat{\Delta} = S_{12}^{\star}(S_{22}^{\star})^{-1}$. The null hypothesis string in $null.value names both blocks.

References

Klein, M., Moura, R., and Sinha, B. (2021). Multivariate normal inference based on singly imputed synthetic data under plug-in sampling. *Sankhya B*, **83**, 273–287. doi:10.1007/s13571019002159

See Also

canodist, ps_test

Examples

```
data(brittany_soil_ps)

set.seed(1)
V5 <- simSynthData(brittany_soil_ps, M = 5)


# Integer interface: zero regression
regression_test(V5, M = 5, part = 2L, iterations = 1000L)
```

```
# Named interface with Delta0 estimated from the original data
S0 <- cov(brittany_soil_ps)
response <- c("pH_water", "pH_KCl", "log_CEC_Metson")
predictors <- c("log_Organic_C", "log_Total_N", "log_P_Olsen")
b <- partition(S0,
  part1 = response,
  part2 = predictors
)
Delta0 <- b$B %*% solve(b$D)

regression_test(
  V5,
  M = 5,
  response = response,
  predictors = predictors,
  Delta0 = Delta0,
  iterations = 1000L
)
```

`simSynthData` *Generate Plug-in Sampling Synthetic Data Sets*

Description

Generates $M \geq 1$ independent fully synthetic data sets from an original numeric matrix `X` using the plug-in sampling (PS) mechanism under a multivariate normal model. The synthetic observations are returned as a single stacked $Mn \times p$ matrix.

The unknown population parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are replaced by the sample mean $\bar{\mathbf{x}}$ and sample covariance matrix $\hat{\boldsymbol{\Sigma}}$. Then Mn synthetic observations are drawn independently from $\mathcal{N}_p(\bar{\mathbf{x}}, \hat{\boldsymbol{\Sigma}})$.

Setting `M = 1` produces a single synthetic data set of size n, corresponding to the classical single-release PS procedure of Klein et al. (2021). Setting `M > 1` produces the stacked data set $\mathbf{V}_{\text{complete}}$ used by the multiple-release procedures:

$$\mathbf{V}_{\text{complete}} = \begin{pmatrix} \mathbf{V}_1 \\ \vdots \\ \mathbf{V}_M \end{pmatrix} \in \mathbb{R}^{Mn \times p}.$$

Usage

```
simSynthData(X, M = 1L)
```

Arguments

`X`	A numeric matrix or data frame containing the original confidential observations. Rows are observations and columns are variables. The input must have dimension $n \times p$ with $n > p$.
`M`	A positive integer giving the number of independent synthetic releases to generate. The default is `1L`. The returned matrix has Mn rows.

Details

The stacked representation is statistically justified because all Mn rows are conditionally independent and identically distributed given the original data. Thus, the stacked sufficient statistic $\mathbf{S}_{\mathrm{M}}^{\star}$ satisfies

$$(n-1)\mathbf{S}_{\mathrm{M}}^{\star} \mid \mathbf{S} \sim \mathcal{W}_p(\mathbf{S},\ Mn-1)\,.$$

Value

An $Mn \times p$ numeric matrix. Column names are preserved from `X`. For `M = 1`, row names are preserved from `X` when available. For `M > 1`, row names encode the release index and observation index using the form `"release_j.obs_i"`.

References

Klein, M., Moura, R., and Sinha, B. (2021). Multivariate normal inference based on singly imputed synthetic data under plug-in sampling. *Sankhya B*, **83**, 273–287. doi:10.1007/s13571019002159

See Also

`ps_test`, `gv_test`, `sphericity_test`, `independence_test`, `regression_test`

Examples

```
data(brittany_soil_ps)

# Single release: M = 1
set.seed(1)
V1 <- simSynthData(brittany_soil_ps)
dim(V1)

# Five releases stacked row-wise
set.seed(1)
V5 <- simSynthData(brittany_soil_ps, M = 5)
dim(V5)
```

`Sphdist` *Simulate the Sphericity Null Distribution*

Description

Simulates the null distribution of the sphericity pivotal statistic $T_2^{\star}$ under plug-in sampling for $M \geq 1$ releases.

Under the stacking result, the compound Wishart representation gives

$$T_2^{\star} \overset{d}{=} \frac{|W_1 W_2|^{1/p}}{\mathrm{tr}(W_1 W_2)/p},$$

where $W_1 \sim \mathcal{W}_p((n-1)^{-1} I_p,\, n-1)$ and $W_2 \sim \mathcal{W}_p(I_p,\, Mn-1)$ independently.

For $M = 1$, both Wishart distributions have degrees of freedom $n-1$, recovering the single-release result of Klein et al. (2021).

Usage

```
Sphdist(nsample, pvariates, iterations = 10000L, M = 1L)
```

Arguments

`nsample`	Original sample size n.
`pvariates`	Number of variables p.
`iterations`	Number of Monte Carlo draws. The default is `10000L`.
`M`	Number of synthetic releases. The default is `1L`.

Value

A numeric vector of length `iterations`.

See Also

`sphericity_test`

Examples

```
set.seed(1)


nd1 <- Sphdist(nsample = 50, pvariates = 4, M = 1, iterations = 1000L)
stats::quantile(nd1, probs = 0.05)

nd5 <- Sphdist(nsample = 50, pvariates = 4, M = 5, iterations = 1000L)
stats::quantile(nd5, probs = 0.05)
```

---

sphericity_test *Sphericity Test*

Description

Tests $H_0 : \Sigma = \sigma^2 I_p$, that is, all variables uncorrelated with equal variance. The test is based on M released plug-in sampling synthetic data sets stacked into `V`. The test is left-tailed. Setting `M = 1` recovers the single-release procedure of Klein et al. (2021).

Usage

```
sphericity_test(V, M = 1L, alpha = 0.05, iterations = 10000L, null_dist = NULL)
```

Arguments

V	Stacked synthetic data set, given as an $Mn \times p$ numeric matrix.
M	Positive integer giving the number of synthetic releases. The default is 1L.
alpha	Significance level. The default is 0.05.
iterations	Monte Carlo sample size used to approximate the null distribution. The default is 10000L.
null_dist	Optional numeric vector containing a precomputed null distribution. If supplied, iterations is ignored.

Value

An object of class ps_test. Component sigma2.hat gives the plug-in estimator $\hat{\sigma}^2 = \mathrm{tr}(S^\star)/(p(N-1))$ under H_0.

References

Klein, M., Moura, R., and Sinha, B. (2021). Multivariate normal inference based on singly imputed synthetic data under plug-in sampling. *Sankhya B*, **83**, 273–287. doi:10.1007/s13571019002159

See Also

Sphdist, ps_test

Examples

```
data(brittany_soil_ps)

set.seed(1)
V5 <- simSynthData(brittany_soil_ps, M = 5)


res <- sphericity_test(V5, M = 5, iterations = 1000L)
print(res)
plot(res)
```

---

summary.ps_test *Summarize a* ps_test *Object*

Description

Prints a detailed summary including the null-distribution quantiles and a comparison with the observed statistic.

Usage

```
## S3 method for class 'ps_test'
summary(object, ...)
```

Arguments

object	An object of class ps_test.
...	Further arguments, currently ignored.

Value

Invisibly returns object.

Examples

```
data(brittany_soil_ps)
V <- simSynthData(brittany_soil_ps, M = 3)
res <- sphericity_test(V, M = 3)
summary(res)
```

---

utility_measures *Utility Measures for Plug-in Sampling Synthetic Data*

Description

Computes five complementary utility measures quantifying how well plug-in sampling (PS) synthetic data preserve the statistical properties of the original confidential data. The reported headline measures are Frobenius distance, mean standardized mean difference, variance ratio range, propensity score MSE (pMSE) ratio, and mean confidence interval overlap. For $M > 1$ releases, per-release statistics are computed and averaged where appropriate.

Usage

```
utility_measures(X, V, M = 1L, alpha = 0.05, verbose = TRUE)
```

Arguments

X	A numeric matrix or data frame with dimension $n \times p$, containing the original confidential data.
V	A numeric matrix or data frame with dimension $Mn \times p$, containing the stacked PS synthetic data returned by simSynthData.
M	Positive integer giving the number of synthetic releases. The default is 1L.
alpha	Significance level used for confidence interval overlap. The default is 0.05.
verbose	Logical. If TRUE, the default, a formatted summary is printed.

Details

Frobenius distance: The Frobenius distance measures covariance matrix preservation:

$$d_F = \|\hat{\boldsymbol{\Sigma}}_X - \bar{\hat{\boldsymbol{\Sigma}}}_V\|_F.$$

For $M > 1$, $\bar{\hat{\boldsymbol{\Sigma}}}_V$ is the average per-release sample covariance matrix.

Mean standardized mean difference: The mean standardized mean difference is

$$\frac{1}{p}\sum_{j=1}^{p}\frac{|\bar{v}_j - \bar{x}_j|}{s_{x_j}}.$$

Values below 0.10 indicate negligible marginal mean differences.

Variance ratio range: The variance ratio range is

$$\left[\min_j \frac{s^2_{v_j}}{s^2_{x_j}}, \max_j \frac{s^2_{v_j}}{s^2_{x_j}}\right].$$

Values close to 1 indicate good variance preservation.

Propensity score MSE ratio: The original data, labeled 0, and the synthetic data, labeled 1, are combined. A logistic classifier is fitted to distinguish original from synthetic records. The pMSE is

$$\text{pMSE} = \frac{1}{n+Mn}\sum_{i=1}^{n+Mn}(\hat{p}_i - c)^2, \qquad c = \frac{Mn}{n+Mn}.$$

The expected value under a correctly specified synthesis model is

$$\text{pMSE}_{\text{null}} = \frac{c(1-c)(p+1)}{n+Mn}.$$

The pMSE ratio is

$$\text{pMSE}/\text{pMSE}_{\text{null}}.$$

Values near 1 indicate good utility, whereas values well below 1 indicate that the synthetic data blend in well with the original data.

Mean confidence interval overlap: For each variable j, a $(1-\alpha)$ confidence interval (CI) is computed from the original data and from the synthetic data. The overlap coefficient is

$$J_j = \frac{\max\left[0,\ \min(u_j^X, u_j^V) - \max(l_j^X, l_j^V)\right]}{\max(u_j^X - l_j^X,\ u_j^V - l_j^V)}.$$

where $[l_j^X, u_j^X]$ and $[l_j^V, u_j^V]$ are the CIs from original and synthetic data. The reported measure is the average of J_j across all variables.

Value

A list of class `ps_utility`, returned invisibly, with components:

frobenius Numeric. Frobenius distance $\|\hat{\boldsymbol{\Sigma}}_X - \bar{\hat{\boldsymbol{\Sigma}}}_V\|_F$.

smd_mean Numeric. Mean standardized mean difference, $p^{-1}\sum_j |\bar{v}_j - \bar{x}_j|/s_{x_j}$.

var_ratio_range Named numeric vector with the minimum and maximum per-variable variance ratios.

pmse_ratio Numeric. Ratio `pmse / pmse_null`.

ci_overlap_mean Numeric. Mean confidence interval overlap across all variables.

M Integer. Number of synthetic releases.

n Integer. Original sample size.

p Integer. Number of variables.

alpha Numeric. Significance level used.

References

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, **60**, 224–232.

Snoke, J., Raab, G. M., Nowok, B., Dibben, C., and Slavkovic, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A*, **181**, 663–688.

Woo, M.-J., Reiter, J. P., Oganian, A., and Karr, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality*, **1**, 111–124.

See Also

`simSynthData`, `ps_test`

Examples

```
data(brittany_soil_ps)

set.seed(1)
V3 <- simSynthData(brittany_soil_ps, M = 3)

utility_measures(brittany_soil_ps, V3, M = 3)
```

Index