

Package ‘PSinference’

December 10, 2024

Type Package

Title Inference for Released Plug-in Sampling Single Synthetic Dataset

Version 0.2.2

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](https://doi.org/10.1007/s13571-019-00215-9)>.

License GPL (>= 2)

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

Imports MASS, stats

Encoding UTF-8

RoxygenNote 7.2.3

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

NeedsCompilation no

Author Ricardo Moura [aut, cre, cph] (<<https://orcid.org/0000-0002-3003-9235>>),
Mina Norouzirad [aut] (<<https://orcid.org/0000-0003-0311-6888>>),
Vitor Augusto [aut] (<<https://orcid.org/0009-0004-9003-3558>>),
Miguel Fonseca [ctb] (<<https://orcid.org/0000-0002-0162-8372>>),
FCT, I.P. [fnd] (under the scope of the projects UIDB/00297/2020 and
UIDP/00297/2020 (NovaMath))

Repository CRAN

Date/Publication 2024-12-10 14:50:05 UTC

Contents

canodist	2
GVdist	4
Inndist	5

partition	7
simSynthData	8
Sphdist	9

Index	11
--------------	-----------

canodist	<i>Canonical Empirical Distribution</i>
----------	---

Description

This function calculates the empirical distribution of the pivotal random variable that can be used to perform inferential procedures for the regression of one subset of variables on the other based on the released Single Synthetic data generated under Plug-in Sampling, assuming that the original dataset is normally distributed.

Usage

```
canodist(part, nsample, pvariates, iterations)
```

Arguments

part	Number of variables in the first subset.
nsample	Sample size.
pvariates	Number of variables.
iterations	Number of iterations for simulating values from the distribution and finding the quantiles. Default is 10000.

Details

We define

$$T_4^*|\Delta = \frac{(|\mathbf{S}_{12}^*(\mathbf{S}_{22}^*)^{-1} - \Delta|\mathbf{S}_{22}^*(\mathbf{S}_{12}^*)(\mathbf{S}_{22}^*)^{-1} - \Delta)^\top|}{|\mathbf{S}_{11.2}^*|}$$

where $\mathbf{S}^* = \sum_{i=1}^n (v_i - \bar{v})(v_i - \bar{v})^\top$, v_i is the i th observation of the synthetic dataset, considering \mathbf{S}^* partitioned as

$$\mathbf{S}^* = \begin{bmatrix} \mathbf{S}_{11}^* & \mathbf{S}_{12}^* \\ \mathbf{S}_{21}^* & \mathbf{S}_{22}^* \end{bmatrix}.$$

For $\Delta = \Sigma_{12}\Sigma_{22}^{-1}$, where Σ is partitioned the same way as \mathbf{S}^* its distribution is stochastic equivalent to

$$\frac{|\Omega_{12}\Omega_{22}^{-1}\Omega_{21}|}{|\Omega_{11} - \Omega_{12}\Omega_{22}^{-1}\Omega_{21}|}$$

where $\Omega \sim \mathcal{W}_p(n-1, \frac{\mathbf{W}}{n-1})$, $\mathbf{W} \sim \mathcal{W}_p(n-1, \mathbf{I}_p)$ and Ω partitioned in the same way as \mathbf{S}^* . To test $\mathcal{H}_0 : \Delta = \Delta_0$, compute the value of T_4^* , \widetilde{T}_4^* , with the observed values and reject the null hypothesis if $\widetilde{T}_4^* > t_{4,1-\alpha}^*$ for α -significance level, where $t_{4,\gamma}^*$ is the γ th percentile of T_4^* .

Value

a vector of length iterations that recorded the empirical distribution's values.

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.

Examples

```
# generate original data
library(MASS)
n_sample = 100
p = 4
mu <- c(1,2,3,4)
Sigma = matrix(c(1, 0.5, 0.1, 0.7,
                 0.5, 2, 0.4, 0.9,
                 0.1, 0.4, 3, 0.2,
                 0.7, 0.9, 0.2, 4), nr = 4, nc = 4, byrow = TRUE)

df = mvrnorm(n_sample, mu = mu, Sigma = Sigma)
# generate synthetic data
df_s = simSynthData(df)
#Decompose Sigma and Sstar
part = 2
Sigma_12 = partition(Sigma,nrows = part, ncol = part)[[2]]
Sigma_22 = partition(Sigma,nrows = part, ncol = part)[[4]]
Delta0 = Sigma_12 %*% solve(Sigma_22)

Sstar = cov(df_s)
Sstar_11 = partition(Sstar,nrows = part, ncol = part)[[1]]
Sstar_12 = partition(Sstar,nrows = part, ncol = part)[[2]]
Sstar_21 = partition(Sstar,nrows = part, ncol = part)[[3]]
Sstar_22 = partition(Sstar,nrows = part, ncol = part)[[4]]

DeltaEst = Sstar_12 %*% solve(Sstar_22)
Sstar11_2 = Sstar_11 - Sstar_12 %*% solve(Sstar_22) %*% Sstar_21

T4_obs = det((DeltaEst-Delta0)%*%Sstar_22%*t(DeltaEst-Delta0))/det(Sstar11_2)

T4 <- canodist(part = part, nsample = n_sample, pvariates = p, iterations = 10000)
q95 <- quantile(T4, 0.95)

T4_obs > q95 #False means that we don't have statistical evidences to reject Delta0
print(T4_obs)
print(q95)
# When the observed value is smaller than the 95% quantile,
# we don't have statistical evidences to reject the Sphericity property.
#
# Note that the value is very close to zero
```

GVdist

*Generalized Variance Empirical Distribution***Description**

This function calculates the empirical distribution of the pivotal random variable that can be used to perform inferential procedures for the Generalized Variance of the released Single Synthetic dataset generated under Plug-in Sampling, assuming that the original distribution is normally distributed.

Usage

```
GVdist(nsamples, pvariables, iterations = 10000)
```

Arguments

nsamples	Sample size.
pvariables	Number of variables.
iterations	Number of iterations for simulating values from the distribution and finding the quantiles. Default is 10000.

Details

We define

$$T_1^* = (n - 1) \frac{|\mathbf{S}^*|}{|\boldsymbol{\Sigma}|},$$

where $\mathbf{S}^* = \sum_{i=1}^n (v_i - \bar{v})(v_i - \bar{v})^\top$, $\boldsymbol{\Sigma}$ is the population covariance matrix and v_i is the i th observation of the synthetic dataset. Its distribution is stochastic equivalent to

$$\prod_{i=1}^n \chi_{n-i}^2 \prod_{i=1}^p \chi_{n-i}^2$$

where χ_{n-i}^2 are all independent chi-square random variables. The $(1 - \alpha)$ level confidence interval for $|\boldsymbol{\Sigma}|$ is given by

$$\left(\frac{(n-1)^p |\tilde{\mathbf{S}}^*|}{t_{1,1-\alpha/2}^*}, \frac{(n-1)^p |\tilde{\mathbf{S}}^*|}{t_{1,\alpha/2}^*} \right)$$

where $\tilde{\mathbf{S}}^*$ is the observed value of \mathbf{S}^* and $t_{1,\gamma}^*$ is the γ th percentile of T_1 .

Value

a vector of length `iterations` that recorded the empirical distribution's values.

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.

Examples

```

# Original data creation
library(MASS)
mu <- c(1,2,3,4)
Sigma <- matrix(c(1, 0.5, 0.5, 0.5,
                 0.5, 1, 0.5, 0.5,
                 0.5, 0.5, 1, 0.5,
                 0.5, 0.5, 0.5, 1), nrow = 4, ncol = 4, byrow = TRUE)

seed = 1
n_sample = 100
# Create original simulated dataset
df = mvrnorm(n_sample, mu = mu, Sigma = Sigma)

# Synthetic data created

df_s = simSynthData(df)

# Gather the 0.025 and 0.975 quantiles and construct confident interval for sigma^2
# Check that sigma^2 is inside in both cases
p = dim(df_s)[2]

T <- GVdist(100, p, 10000)
q975 <- quantile(T, 0.975)
q025 <- quantile(T, 0.025)

left <- (n_sample-1)^p * det(cov(df_s)*(n_sample-1))/q975
right <- (n_sample-1)^p * det(cov(df_s)*(n_sample-1))/q025

cat(left,right,'\n')
print(det(Sigma))
# The synthetic value is inside the confidence interval of GV

```

Description

This function calculates the empirical distribution of the pivotal random variable that can be used to perform inferential procedures and test the independence of two subsets of variables based on the released Single Synthetic data generated under Plug-in Sampling, assuming that the original dataset is normally distributed.

Usage

```
Inddist(part, nsample, pvariates, iterations)
```

Arguments

part	Number of variables in the first subset.
nsample	Sample size.
pvariables	Number of variables.
iterations	Number of iterations for simulating values from the distribution and finding the quantiles. Default is 10000.

Details

We define

$$T_3^* = \frac{|\mathbf{S}^*|}{|\mathbf{S}_{11}^*||\mathbf{S}_{22}^*|}$$

where $\mathbf{S}^* = \sum_{i=1}^n (v_i - \bar{v})(v_i - \bar{v})^\top$, v_i is the i th observation of the synthetic dataset, considering \mathbf{S}^* partitioned as

$$\mathbf{S}^* = \begin{bmatrix} \mathbf{S}_{11}^* & \mathbf{S}_{12}^* \\ \mathbf{S}_{21}^* & \mathbf{S}_{22}^* \end{bmatrix}.$$

Under the assumption that $\Sigma_{12} = \mathbf{0}$, its distribution is stochastic equivalent to

$$\frac{|\Omega|}{|\Omega_{11}||\Omega_{22}|}$$

where $\Omega \sim \mathcal{W}_p(n-1, \frac{\mathbf{W}}{n-1})$, $\mathbf{W} \sim \mathcal{W}_p(n-1, \mathbf{I}_p)$ and Ω partitioned in the same way as \mathbf{S}^* . To test $\mathcal{H}_0 : \Sigma_{12} = \mathbf{0}$, compute the value of T_3^* , \widetilde{T}_3^* , with the observed values and reject the null hypothesis if $\widetilde{T}_3^* < t_{3,\alpha}^*$ for α -significance level, where $t_{3,\gamma}^*$ is the γ th percentile of T_3^* .

Value

a vector of length iterations that recorded the empirical distribution's values.

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.

Examples

```
#generate original data with two independent subsets of variables
library(MASS)
n_sample = 100
p = 4
mu <- c(1,2,3,4)
Sigma = matrix(c(1, 0.5, 0, 0,
                0.5, 2, 0, 0,
                0, 0, 3, 0.2,
                0, 0, 0.2, 4), nr = 4, nc = 4, byrow = TRUE)
df = mvrnorm(n_sample, mu = mu, Sigma = Sigma)
# generate synthetic data
df_s = simSynthData(df)
```

```
#Decompose Sstar in 4 parts
part = 2

Sstar = cov(df_s)
Sstar_11 = partition(Sstar,nrows = part, ncol = part)[[1]]
Sstar_12 = partition(Sstar,nrows = part, ncol = part)[[2]]
Sstar_21 = partition(Sstar,nrows = part, ncol = part)[[3]]
Sstar_22 = partition(Sstar,nrows = part, ncol = part)[[4]]

#Compute observed T3_star
T3_obs = det(Sstar)/(det(Sstar_11)*det(Sstar_22))

alpha = 0.05

# collect the quantile from the distribution assuming independence between the two subsets
T3 <- Inddist(part = part, nsample = n_sample, pvariates = p, iterations = 10000)
q5 <- quantile(T3, alpha)

T3_obs < q5 #False means that we don't have statistical evidences to reject independence
print(T3_obs)
print(q5)
# Note that the value of the observed T3_obs is close to one as expected
```

partition

Split a matrix into blocks

Description

This function split a matrix into a list of blocks (either by rows and columns).

Usage

```
partition(Matrix, nrows, ncols)
```

Arguments

Matrix a matrix to split .
nrows positive integer indicating the number of rows blocks.
ncols positive integer indicating the number of columns blocks.

Value

a list of partitioned submatrices

Examples

```
Mat = matrix(c(1,0.5,0,0,
              0.5,2,0,0,
              0,0,3,0.2,
              0, 0, 0.2,4), nrow = 4, ncol = 4, byrow = TRUE)
partition(Matrix = Mat, nrows = 2, ncols = 2)
```

 simSynthData

Plug-in Sampling Single Synthetic Dataset Generation

Description

This function is used to generate a single synthetic version of the original data via Plug-in Sampling.

Usage

```
simSynthData(X, n_imp = dim(X)[1])
```

Arguments

X	matrix or dataframe
n_imp	sample size

Details

Assume that $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ is the original data, assumed to be normally distributed, we compute $\bar{\mathbf{x}}$ as the sample mean and $\hat{\Sigma} = \mathbf{S}/(n-1)$ as the sample covariance matrix, where \mathbf{S} is the sample Wishart matrix. We generate $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_n)$, by drawing

$$\mathbf{v}_i \stackrel{i.i.d.}{\sim} N_p(\bar{\mathbf{x}}, \hat{\Sigma}).$$

Value

a matrix of generated dataset

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.

Examples

```

library(MASS)
n_sample = 1000
mu=c(0,0,0,0)
Sigma=diag(1,4,4)
# Create original simulated dataset
df_o = mvrnorm(n_sample, mu, Sigma)
# Create singly imputed synthetic dataset
df_s = simSynthData(df_o)
#Estimators synthetic
mean_s <- colMeans(df_s)
S_s <- (t(df_s)- mean_s) %*% t(t(df_s)- mean_s)
# careful about this computation
# mean_o is a column vector and if you are thinking as n X p matrices and
# row vectors you should be aware of this.
print(mean_s)
print(S_s/(dim(df_s)[1]-1))

```

Sphdist

*Spherical Empirical Distribution***Description**

This function calculates the empirical distribution of the pivotal random variable that can be used to perform the Sphericity test of the population covariance matrix Σ that is $\Sigma = \sigma^2 \mathbf{I}_p$, based on the released Single Synthetic data generated under Plug-in Sampling, assuming that the original dataset is normally distributed.

Usage

```
Sphdist(nsample, pvariates, iterations)
```

Arguments

nsample	Sample size.
pvariates	Number of variables.
iterations	Number of iterations for simulating values from the distribution and finding the quantiles. Default is 10000.

Details

We define

$$T_2^* = \frac{|\mathbf{S}^*|^{\frac{1}{p}}}{tr(\mathbf{S}^*)/p}$$

where $\mathbf{S}^* = \sum_{i=1}^n (v_i - \bar{v})(v_i - \bar{v})^\top$, v_i is the i th observation of the synthetic dataset. For $\Sigma = \sigma^2 \mathbf{I}_p$, its distribution is stochastic equivalent to

$$\frac{|\Omega_1 \Omega_2|^{\frac{1}{p}}}{tr(\Omega_1 \Omega_2)/p}$$

where Ω_1 and Ω_2 are Wishart random variables, $\Omega_1 \sim \mathcal{W}_p(n-1, \frac{\mathbf{I}_p}{n-1})$ is independent of $\Omega_2 \sim \mathcal{W}_p(n-1, \mathbf{I}_p)$. To test $\mathcal{H}_0 : \Sigma = \sigma^2 \mathbf{I}_p$, compute the observed value of T_2^* , \widetilde{T}_2^* , with the observed values and reject the null hypothesis if $\widetilde{T}_2^* > t_{2,\alpha}^*$ for α -significance level, where $t_{2,\gamma}^*$ is the γ th percentile of T_2^* .

Value

a vector of length `iterations` that recorded the empirical distribution's values.

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.

Examples

```
# Original data created
library(MASS)
mu <- c(1,2,3,4)
Sigma <- matrix(c(1, 0, 0, 0,
                  0, 1, 0, 0,
                  0, 0, 1, 0,
                  0, 0, 0, 1), nrow = 4, ncol = 4, byrow = TRUE)

seed = 1
n_sample = 100
# Create original simulated dataset
df = mvrnorm(n_sample, mu = mu, Sigma = Sigma)

# Synthetic data created

df_s = simSynthData(df)

# Gather the 0.95 quantile

p = dim(df_s)[2]

T_sph <- Sphdist(nsample = n_sample, pvariables = p, iterations = 10000)
q95 <- quantile(T_sph, 0.95)

# Compute the observed value of T from the synthetic dataset
S_star = cov(df_s*(n_sample-1))

T_obs = (det(S_star)^(1/p))/(sum(diag(S_star))/p)

print(q95)
print(T_obs)

#Since the observed value is bigger than the 95% quantile,
#we don't have statistical evidences to reject the Sphericity property.
#
#Note that the value is very close to one
```

Index

canodist, 2

GVdist, 4

Inddist, 5

partition, 7

simSynthData, 8

Sphdist, 9