Package 'highmean'

October 14, 2016

Type Package

Title Two-Sample Tests for High-Dimensional Mean Vectors

Version 3.0

Date 2016-10-15

Author Lifeng Lin and Wei Pan

Maintainer Lifeng Lin <linl@umn.edu>

Depends R (>= 1.9.0), mvtnorm (>= 1.0-0)

Imports MASS, mnormt

Description

Provides various tests for comparing high-dimensional mean vectors in two sample populations.

License GPL (>= 2)

NeedsCompilation no

Repository CRAN

Date/Publication 2016-10-14 00:35:09

R topics documented:

highmean-package *Two-Sample Tests for High-Dimensional Mean Vectors*

Description

Provides various tests for comparing high-dimensional mean vectors in two sample populations.

Details

Several two-sample tests for high-dimensional mean vectors have been proposed recently; see, e.g., Bai and Saranadasa (1996), Srivastava and Du (2008), Chen and Qin (2010), Cai et al (2014), and Chen et al (2014). However, these tests are powerful only against certain and limited alternative hypotheses. In practice, since the true alternative hypothesis is unknown, it is unclear how to choose one of these tests to yield high power. Accordingly, Pan et al (2014) and Xu et al (2016) proposed an adaptive test that may maintain high power across a wide range of situations; the asymptotic property of this test was also studied. This package provides functions to calculate p-values of the foregoing tests, using their asymptotic properties and the empirical (permutation or parametric bootstrap resampling) technique.

Author(s)

Lifeng Lin and Wei Pan

Maintainer: Lifeng Lin <linl@umn.edu>

References

Bai ZD and Saranadasa H (1996). "Effect of high dimension: by an example of a two sample problem." *Statistica Sinica*, **6**(2), 311–329.

Cai TT, Liu W, and Xia Y (2014). "Two-sample test of high dimensional means under dependence." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(2), 349–372.

Chen SX and Qin YL (2010). "A two-sample test for high-dimensional data with applications to gene-set testing." *The Annals of Statistics*, **38**(2), 808–835.

Chen SX, Li J, and Zhong PS (2014). "Two-Sample Tests for High Dimensional Means with Thresholding and Data Transformation." arXiv preprint arXiv:1410.2848.

Pan W, Kim J, Zhang Y, Shen X, and Wei P (2014). "A powerful and adaptive association test for rare variants." *Genetics*, **197**(4), 1081–1095.

Srivastava MS and Du M (2008). "A test for the mean vector with fewer observations than the dimension." *Journal of Multivariate Analysis*, **99**(3), 386–402.

Xu G, Lin L, Wei P, and Pan W (2016). "An adaptive two-sample test for high-dimensional means." *Biometrika*, **103**(3), 609–624.

apval_aSPU *Asymptotics-Based p-values of the SPU and aSPU Tests*

Description

Calculates p-values of the sum-of-powers (SPU) and adaptive SPU (aSPU) tests based on the asymptotic distributions of the test statistics (Xu et al, 2016).

Usage

```
apval_aSPU(sam1, sam2, pow = c(1:6, Inf), eq.cov = TRUE, cov.est,
           cov1.est, cov2.est, bandwidth, bandwidth1, bandwidth2,
           cv.fold = 5, norm = "F")
```

Arguments

sam1	an n1 by p matrix from sample population 1. Each row represents a p-dimensional sample.
sam2	an n2 by p matrix from sample population 2. Each row represents a p-dimensional sample.
pow	a numeric vector indicating the candidate powers γ in the SPU tests. It should contain Inf and both odd and even integers. The default is c(1:6, Inf).
eq.cov	a logical value. The default is TRUE, indicating that the two sample populations have same covariance; otherwise, the covariances are assumed to be different.
cov.est	a consistent estimate of the common covariance matrix when eq.cov is TRUE. This can be obtained from various apporoaches (e.g., banding, tapering, and thresholding; see Pourahmadi 2013). If not specified, this function uses a banding approach proposed by Bickel and Levina (2008) to estimate the covariance matrix.
cov1.est	a consistent estimate of the covariance matrix of sample population 1 when eq.cov is FALSE. It is similar with the argument cov.est.
cov2.est	a consistent estimate of the covariance matrix of sample population 2 when eq.cov is FALSE. It is similar with the argument cov.est.
bandwidth	a vector of nonnegative integers indicating the candidate bandwidths to be used in the banding approach (Bickel and Levina, 2008) for estimating the common covariance when eq.cov is TRUE. This argument is effective only if cov.est is not provided. The default is a vector containing 50 candidate bandwidths chosen from {0, 1, 2, ..., p}.
bandwidth1	similar with the argument bandwidth; it is used to specify candidate bandwidths for estimating the covariance of sample population 1 when eq.cov is FALSE.
bandwidth2	similar with the argument bandwidth; it is used to specify candidate bandwidths for estimating the covariance of sample population 2 when eq.cov is FALSE.
cv.fold	an integer greater than or equal to 2 indicating the fold of cross-validation. The default is 5. See page 211 in Bickel and Levina (2008).

`norm` a character string indicating the type of matrix norm for the calculation of risk function in cross-validation. This argument will be passed to the `norm` function. The default is the Frobenius norm (`"F"`).

Details

Suppose that the two groups of p-dimensional independent and identically distributed samples $\{X_{1i}\}_{i=1}^{n_1}$ and $\{X_{2j}\}_{j=1}^{n_2}$ are observed; we consider high-dimensional data with $p \gg n := n_1+n_2-2$. Assume that the covariances of the two sample populations are $\Sigma_1 = (\sigma_{1,ij})$ and $\Sigma_2 = (\sigma_{2,ij})$. The primary object is to test $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$. Let $\bar{X}_k$ be the sample mean for group $k = 1, 2$. For a vector v, we denote $v^{(i)}$ as its ith element.

For any $1 \leq \gamma < \infty$, the sum-of-powers (SPU) test statistic is defined as:

$$L(\gamma) = \sum_{i=1}^{p} (\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^\gamma.$$

For $\gamma = \infty$,

$$L(\infty) = \max_{i=1,\ldots,p} (\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^2 / (\sigma_{1,ii}/n_1 + \sigma_{2,ii}/n_2).$$

The adaptive SPU (aSPU) test combines the SPU tests and improve the test power:

$$T_{aSPU} = \min_{\gamma \in \Gamma} P_{SPU(\gamma)},$$

where $P_{SPU(\gamma)}$ is the p-value of SPU(γ) test, and Γ is a candidate set of γ's. Note that T_{aSPU} is no longer a genuine p-value. The asymptotic properties of the SPU and aSPU tests are studied in Xu et al (2016).

Value

A list including the following elements:

`sam.info` the basic information about the two groups of samples, including the samples sizes and dimension.

`pow` the powers γ used for the SPU tests.

`opt.bw` the optimal bandwidth determined by the cross-validation when `eq.cov` was `TRUE` and `cov.est` was not specified.

`opt.bw1` the optimal bandwidth determined by the cross-validation when `eq.cov` was `FALSE` and `cov1.est` was not specified.

`opt.bw2` the optimal bandwidth determined by the cross-validation when `eq.cov` was `FALSE` and `cov2.est` was not specified.

`spu.stat` the observed SPU test statistics.

`spu.e` the asymptotic means of SPU test statistics with finite γ under the null hypothesis.

`spu.var` the asymptotic variances of SPU test statistics with finite γ under the null hypothesis.

`spu.corr.odd` the asymptotic correlations between SPU test statistics with odd γ.

`spu.corr.even`	the asymptotic correlations between SPU test statistics with even γ.
`cov.assumption`	the equality assumption on the covariances of the two sample populations; this was specified by the argument `eq.cov`.
`method`	this output reminds users that the p-values are obtained using the asymptotic distributions of test statistics.
`pval`	the p-values of the SPU tests and the aSPU test.

References

Bickel PJ and Levina E (2008). "Regularized estimation of large covariance matrices." *The Annals of Statistics*, **36**(1), 199–227.

Pan W, Kim J, Zhang Y, Shen X, and Wei P (2014). "A powerful and adaptive association test for rare variants." *Genetics*, **197**(4), 1081–1095.

Pourahmadi M (2013). *High-Dimensional Covariance Estimation*. John Wiley & Sons, Hoboken, NJ.

Xu G, Lin L, Wei P, and Pan W (2016). "An adaptive two-sample test for high-dimensional means." *Biometrika*, **103**(3), 609–624.

See Also

`cpval_aSPU`, `epval_aSPU`

Examples

```
library(MASS)
set.seed(1234)
n1 <- n2 <- 50
p <- 200
mu1 <- rep(0, p)
mu2 <- mu1
mu2[1:10] <- 0.2
true.cov <- 0.4^(abs(outer(1:p, 1:p, "-"))) # AR1 covariance
sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov)
sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov)
# use true covariance matrix
apval_aSPU(sam1, sam2, cov.est = true.cov)
# fix bandwidth as 10
apval_aSPU(sam1, sam2, bandwidth = 10)
# use the optimal bandwidth from a candidate set
#apval_aSPU(sam1, sam2, bandwidth = 0:20)

# the two sample populations have different covariances
#true.cov1 <- 0.2^(abs(outer(1:p, 1:p, "-")))
#true.cov2 <- 0.6^(abs(outer(1:p, 1:p, "-")))
#sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov1)
#sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov2)
#apval_aSPU(sam1, sam2, eq.cov = FALSE,
# bandwidth1 = 10, bandwidth2 = 10)
```

apval_Bai1996 *Asymptotics-Based p-value of the Test Proposed by Bai and Saranadasa (1996)*

Description

Calculates p-value of the test for testing equality of two-sample high-dimensional mean vectors proposed by Bai and Saranadasa (1996) based on the asymptotic distribution of the test statistic.

Usage

```
apval_Bai1996(sam1, sam2)
```

Arguments

sam1	an n1 by p matrix from sample population 1. Each row represents a p-dimensional sample.
sam2	an n2 by p matrix from sample population 2. Each row represents a p-dimensional sample.

Details

Suppose that the two groups of p-dimensional independent and identically distributed samples $\{X_{1i}\}_{i=1}^{n_1}$ and $\{X_{2j}\}_{j=1}^{n_2}$ are observed; we consider high-dimensional data with $p \gg n := n_1 + n_2 - 2$. Assume that the two groups share a common covariance matrix. The primary object is to test $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$. Let $\bar{X}_k$ be the sample mean for group $k = 1, 2$. Also, let $S = n^{-1} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (X_{ki} - \bar{X}_k)(X_{ki} - \bar{X}_k)^T$ be the pooled sample covariance matrix from the two groups.

Bai and Saranadasa (1996) proposed the following test statistic:

$$T_{BS} = \frac{(n_1^{-1} + n_2^{-1})^{-1}(\bar{X}_1 - \bar{X}_2)^T(\bar{X}_1 - \bar{X}_2) - trS}{\sqrt{2n(n+1)(n-1)^{-1}(n+2)^{-1}[trS^2 - n^{-1}(trS)^2]}},$$

and its asymptotic distribution is normal under the null hypothesis.

Value

A list including the following elements:

sam.info	the basic information about the two groups of samples, including the samples sizes and dimension.
cov.assumption	this output reminds users that the two sample populations have a common covariance matrix.
method	this output reminds users that the p-values are obtained using the asymptotic distributions of test statistics.
pval	the p-value of the test proposed by Bai and Saranadasa (1996).

Note

The asymptotic distribution of the test statistic was derived under normality assumption in Bai and Saranadasa (1996). Also, this function assumes that the two sample populations have a common covariance matrix.

References

Bai ZD and Saranadasa H (1996). "Effect of high dimension: by an example of a two sample problem." *Statistica Sinica*, **6**(2), 311–329.

See Also

epval_Bai1996

Examples

```
library(MASS)
set.seed(1234)
n1 <- n2 <- 50
p <- 200
mu1 <- rep(0, p)
mu2 <- mu1
mu2[1:10] <- 0.2
true.cov <- 0.4^(abs(outer(1:p, 1:p, "-"))) # AR1 covariance
sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov)
sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov)
apval_Bai1996(sam1, sam2)
```

---

apval_Cai2014 *Asymptotics-Based p-value of the Test Proposed by Cai et al (2014)*

Description

Calculates p-value of the test for testing equality of two-sample high-dimensional mean vectors proposed by Cai et al (2014) based on the asymptotic distribution of the test statistic.

Usage

```
apval_Cai2014(sam1, sam2, eq.cov = TRUE)
```

Arguments

sam1	an n1 by p matrix from sample population 1. Each row represents a p-dimensional sample.
sam2	an n2 by p matrix from sample population 2. Each row represents a p-dimensional sample.
eq.cov	a logical value. The default is TRUE, indicating that the two sample populations have same covariance; otherwise, the covariances are assumed to be different.

Details

Suppose that the two groups of p-dimensional independent and identically distributed samples $\{X_{1i}\}_{i=1}^{n_1}$ and $\{X_{2j}\}_{j=1}^{n_2}$ are observed; we consider high-dimensional data with $p \gg n := n_1+n_2-2$. Assume that the covariances of the two sample populations are $\Sigma_1 = (\sigma_{1,ij})$ and $\Sigma_2 = (\sigma_{2,ij})$. The primary object is to test $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$. Let $\bar{X}_k$ be the sample mean for group $k = 1, 2$. For a vector v, we denote $v^{(i)}$ as its ith element.

Cai et al (2014) proposed the following test statistic:

$$T_{CLX} = \max_{i=1,\ldots,p} (\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^2/(\sigma_{1,ii}/n_1 + \sigma_{2,ii}/n_2),$$

This test statistic follows an extreme value distribution under the null hypothesis.

Value

A list including the following elements:

`sam.info`	the basic information about the two groups of samples, including the samples sizes and dimension.
`cov.assumption`	the equality assumption on the covariances of the two sample populations; this was specified by the argument `eq.cov`.
`method`	this output reminds users that the p-values are obtained using the asymptotic distributions of test statistics.
`pval`	the p-value of the test proposed by Cai et al (2014).

Note

This function does not transform the data with their precision matrix (see Cai et al, 2014). To calculate the p-value of the test statisic with transformation, users can use transformed samples for `sam1` and `sam2`.

References

Cai TT, Liu W, and Xia Y (2014). "Two-sample test of high dimensional means under dependence." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(2), 349–372.

See Also

`epval_Cai2014`

Examples

```
library(MASS)
set.seed(1234)
n1 <- n2 <- 50
p <- 200
mu1 <- rep(0, p)
mu2 <- mu1
mu2[1:10] <- 0.2
true.cov <- 0.4^(abs(outer(1:p, 1:p, "-"))) # AR1 covariance
```

```
sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov)
sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov)
apval_Cai2014(sam1, sam2)

# the two sample populations have different covariances
true.cov1 <- 0.2^(abs(outer(1:p, 1:p, "-")))
true.cov2 <- 0.6^(abs(outer(1:p, 1:p, "-")))
sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov1)
sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov2)
apval_Cai2014(sam1, sam2, eq.cov = FALSE)
```

`apval_Chen2010` *Asymptotics-Based p-value of the Test Proposed by Chen and Qin (2010)*

Description

Calculates p-value of the test for testing equality of two-sample high-dimensional mean vectors proposed by Chen and Qin (2010) based on the asymptotic distribution of the test statistic.

Usage

```
apval_Chen2010(sam1, sam2, eq.cov = TRUE)
```

Arguments

`sam1`	an n1 by p matrix from sample population 1. Each row represents a p-dimensional sample.
`sam2`	an n2 by p matrix from sample population 2. Each row represents a p-dimensional sample.
`eq.cov`	a logical value. The default is `TRUE`, indicating that the two sample populations have same covariance; otherwise, the covariances are assumed to be different.

Details

Suppose that the two groups of p-dimensional independent and identically distributed samples $\{X_{1i}\}_{i=1}^{n_1}$ and $\{X_{2j}\}_{j=1}^{n_2}$ are observed; we consider high-dimensional data with $p \gg n := n_1 + n_2 - 2$. The primary object is to test $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$. Let $\bar{X}_k$ be the sample mean for group $k = 1, 2$.

Chen and Qin (2010) proposed the following test statistic:

$$T_{CQ} = \frac{\sum_{i \neq j}^{n_1} X_{1i}^T X_{1j}}{n_1(n_1-1)} + \frac{\sum_{i \neq j}^{n_2} X_{2i}^T X_{2j}}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} X_{1i}^T X_{2j}}{n_1 n_2},$$

and its asymptotic distribution is normal under the null hypothesis.

Value

A list including the following elements:

`sam.info`	the basic information about the two groups of samples, including the samples sizes and dimension.
`cov.assumption`	the equality assumption on the covariances of the two sample populations; this was specified by the argument `eq.cov`.
`method`	this output reminds users that the p-values are obtained using the asymptotic distributions of test statistics.
`pval`	the p-value of the test proposed by Chen and Qin (2010).

References

Chen SX and Qin YL (2010). "A two-sample test for high-dimensional data with applications to gene-set testing." *The Annals of Statistics*, **38**(2), 808–835.

See Also

`epval_Chen2010`

Examples

```
library(MASS)
set.seed(1234)
n1 <- n2 <- 50
p <- 200
mu1 <- rep(0, p)
mu2 <- mu1
mu2[1:10] <- 0.2
true.cov <- 0.4^(abs(outer(1:p, 1:p, "-"))) # AR1 covariance
sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov)
sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov)
apval_Chen2010(sam1, sam2)

# the two sample populations have different covariances
true.cov1 <- 0.2^(abs(outer(1:p, 1:p, "-")))
true.cov2 <- 0.6^(abs(outer(1:p, 1:p, "-")))
sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov1)
sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov2)
apval_Chen2010(sam1, sam2, eq.cov = FALSE)
```

---

apval_Chen2014 *Asymptotics-Based p-value of the Test Proposed by Chen et al (2014)*

Description

Calculates p-value of the test for testing equality of two-sample high-dimensional mean vectors proposed by Chen et al (2014) based on the asymptotic distribution of the test statistic.

Usage

```
apval_Chen2014(sam1, sam2, eq.cov = TRUE)
```

Arguments

sam1	an n1 by p matrix from sample population 1. Each row represents a p-dimensional sample.
sam2	an n2 by p matrix from sample population 2. Each row represents a p-dimensional sample.
eq.cov	a logical value. The default is TRUE, indicating that the two sample populations have same covariance; otherwise, the covariances are assumed to be different.

Details

Suppose that the two groups of p-dimensional independent and identically distributed samples $\{X_{1i}\}_{i=1}^{n_1}$ and $\{X_{2j}\}_{j=1}^{n_2}$ are observed; we consider high-dimensional data with $p \gg n := n_1+n_2-2$. Assume that the covariances of the two sample populations are $\Sigma_1 = (\sigma_{1,ij})$ and $\Sigma_2 = (\sigma_{2,ij})$. The primary object is to test $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$. Let $\bar{X}_k$ be the sample mean for group $k = 1, 2$. For a vector v, we denote $v^{(i)}$ as its ith element.

Chen et al (2014) proposed removing estimated zero components in the mean difference through thresholding; they considered

$$T_{CLZ}(s) = \sum_{i=1}^{p} \left\{ \frac{(\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^2}{\sigma_{1,ii}/n_1 + \sigma_{2,ii}/n_2} - 1 \right\} I\left\{ \frac{(\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^2}{\sigma_{1,ii}/n_1 + \sigma_{2,ii}/n_2} > \lambda_p(s) \right\},$$

where the threshold level is $\lambda_p(s) := 2s \log p$ and $I(\cdot)$ is the indicator function. Since an optimal choice of the threshold is unknown, they proposed trying all possible threshold values, then choosing the most significant one as their final test statistic:

$$T_{CLZ} = \max_{s \in (0, 1-\eta)} \{T_{CLZ}(s) - \hat{\mu}_{T_{CLZ}(s),0}\}/\hat{\sigma}_{T_{CLZ}(s),0},$$

where $\hat{\mu}_{T_{CLZ}(s),0}$ and $\hat{\sigma}_{T_{CLZ}(s),0}$ are estimates of the mean and standard deviation of $T_{CLZ}(s)$ under the null hypothesis. They derived its asymptotic null distribution as an extreme value distribution.

Value

A list including the following elements:

sam.info	the basic information about the two groups of samples, including the samples sizes and dimension.
cov.assumption	the equality assumption on the covariances of the two sample populations; this was specified by the argument eq.cov.
method	this output reminds users that the p-values are obtained using the asymptotic distributions of test statistics.
pval	the p-value of the test proposed by Chen et al (2014).

Note

This function does not transform the data with their precision matrix (see Chen et al, 2014). To calculate the p-value of the test statisic with transformation, users can use transformed samples for `sam1` and `sam2`.

References

Chen SX, Li J, and Zhong PS (2014). "Two-Sample Tests for High Dimensional Means with Thresholding and Data Transformation." arXiv preprint arXiv:1410.2848.

See Also

`epval_Chen2014`

Examples

```
library(MASS)
set.seed(1234)
n1 <- n2 <- 50
p <- 200
mu1 <- rep(0, p)
mu2 <- mu1
mu2[1:10] <- 0.2
true.cov <- 0.4^(abs(outer(1:p, 1:p, "-"))) # AR1 covariance
sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov)
sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov)
apval_Chen2014(sam1, sam2)

# the two sample populations have different covariances
true.cov1 <- 0.2^(abs(outer(1:p, 1:p, "-")))
true.cov2 <- 0.6^(abs(outer(1:p, 1:p, "-")))
sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov1)
sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov2)
apval_Chen2014(sam1, sam2, eq.cov = FALSE)
```

---

apval_Sri2008 *Asymptotics-Based p-value of the Test Proposed by Srivastava and Du (2008)*

Description

Calculates p-value of the test for testing equality of two-sample high-dimensional mean vectors proposed by Srivastava and Du (2008) based on the asymptotic distribution of the test statistic.

Usage

```
apval_Sri2008(sam1, sam2)
```

Arguments

sam1	an n1 by p matrix from sample population 1. Each row represents a p-dimensional sample.
sam2	an n2 by p matrix from sample population 2. Each row represents a p-dimensional sample.

Details

Suppose that the two groups of p-dimensional independent and identically distributed samples $\{X_{1i}\}_{i=1}^{n_1}$ and $\{X_{2j}\}_{j=1}^{n_2}$ are observed; we consider high-dimensional data with $p \gg n := n_1 + n_2 - 2$. Assume that the two groups share a common covariance matrix. The primary object is to test $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$. Let $\bar{X}_k$ be the sample mean for group $k = 1, 2$. Also, let $S = n^{-1} \sum_{k=1}^{2} \sum_{i=1}^{n_k} (X_{ki} - \bar{X}_k)(X_{ki} - \bar{X}_k)^T$ be the pooled sample covariance matrix from the two groups.

Srivastava and Du (2008) proposed the following test statistic:

$$T_{SD} = \frac{(n_1^{-1} + n_2^{-1})^{-1}(\bar{X}_1 - \bar{X}_2)^T D_S^{-1}(\bar{X}_1 - \bar{X}_2) - (n-2)^{-1}np}{\sqrt{2(trR^2 - p^2 n^{-1})c_{p,n}}},$$

where $D_S = diag(s_{11}, s_{22}, ..., s_{pp})$, s_{ii}'s are the diagonal elements of S, $R = D_S^{-1/2} S D_S^{-1/2}$ is the sample correlation matrix and $c_{p,n} = 1 + trR^2 p^{-3/2}$. This test statistic follows normal distribution under the null hypothesis.

Value

A list including the following elements:

sam.info	the basic information about the two groups of samples, including the samples sizes and dimension.
cov.assumption	this output reminds users that the two sample populations have a common covariance matrix.
method	this output reminds users that the p-values are obtained using the asymptotic distributions of test statistics.
pval	the p-value of the test proposed by Srivastava and Du (2008).

Note

The asymptotic distribution of the test statistic was derived under normality assumption in Bai and Saranadasa (1996). Also, this function assumes that the two sample populations have a common covariance matrix.

References

Srivastava MS and Du M (2008). "A test for the mean vector with fewer observations than the dimension." *Journal of Multivariate Analysis*, **99**(3), 386–402.

See Also

epval_Sri2008

Examples

```
library(MASS)
set.seed(1234)
n1 <- n2 <- 50
p <- 200
mu1 <- rep(0, p)
mu2 <- mu1
mu2[1:10] <- 0.2
true.cov <- 0.4^(abs(outer(1:p, 1:p, "-"))) # AR1 covariance
sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov)
sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov)
apval_Sri2008(sam1, sam2)
```

---

cpval_aSPU *Permutation-And-Asymptotics-Based p-values of the SPU and aSPU Tests*

Description

Calculates p-values of the sum-of-powers (SPU) and adaptive SPU (aSPU) tests based on the combination of permutation method and asymptotic distributions of the test statistics (Xu et al, 2016).

Usage

```
cpval_aSPU(sam1, sam2, pow = c(1:6, Inf), n.iter = 1000, seeds)
```

Arguments

sam1	an n1 by p matrix from sample population 1. Each row represents a p-dimensional sample.
sam2	an n2 by p matrix from sample population 2. Each row represents a p-dimensional sample.
pow	a numeric vector indicating the candidate powers γ in the SPU tests. It should contain `Inf` and both odd and even integers. The default is `c(1:6, Inf)`.
n.iter	a numeric integer indicating the number of permutation iterations for calculating the means, variances, covariances of SPU test statistics' asymptotic distributions. The default is 1,000.
seeds	a vector of seeds for each permutation iteration; this is optional.

Details

Suppose that the two groups of p-dimensional independent and identically distributed samples $\{X_{1i}\}_{i=1}^{n_1}$ and $\{X_{2j}\}_{j=1}^{n_2}$ are observed; we consider high-dimensional data with $p \gg n := n_1+n_2-2$. Assume that the covariances of the two sample populations are $\Sigma_1 = (\sigma_{1,ij})$ and $\Sigma_2 = (\sigma_{2,ij})$. The primary object is to test $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$. Let $\bar{X}_k$ be the sample mean for group $k = 1, 2$. For a vector v, we denote $v^{(i)}$ as its ith element.

For any $1 \leq \gamma < \infty$, the sum-of-powers (SPU) test statistic is defined as:

$$L(\gamma) = \sum_{i=1}^{p} (\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^\gamma.$$

For $\gamma = \infty$,

$$L(\infty) = \max_{i=1,\ldots,p} (\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^2 / (\sigma_{1,ii}/n_1 + \sigma_{2,ii}/n_2).$$

The adaptive SPU (aSPU) test combines the SPU tests and improve the test power:

$$T_{aSPU} = \min_{\gamma \in \Gamma} P_{SPU(\gamma)},$$

where $P_{SPU(\gamma)}$ is the p-value of SPU(γ) test, and Γ is a candidate set of γ's. Note that T_{aSPU} is no longer a genuine p-value.

The asymptotic properties of the SPU and aSPU tests are studied in Xu et al (2016). When using the theoretical means, variances, and covarainces of $L(\gamma)$ to calculate the p-values of SPU and aSPU tests ($1 \leq \gamma < \infty$), the high-dimensional covariance matrix of the samples needs to be consistently estimated; such estimation is usually time-consuming.

Alternatively, assuming that the two sample groups have same covariance, the permutation method can be applied to efficiently estimate the means, variances, and covarainces of $L(\gamma)$'s asymptotic distributions, which then yield the p-values of SPU and aSPU tests based on the combination of permutation method and asymptotic distributions.

Value

A list including the following elements:

`sam.info`	the basic information about the two groups of samples, including the samples sizes and dimension.
`pow`	the powers γ used for the SPU tests.
`spu.stat`	the observed SPU test statistics.
`spu.e`	the asymptotic means of SPU test statistics with finite γ under the null hypothesis.
`spu.var`	the asymptotic variances of SPU test statistics with finite γ under the null hypothesis.
`spu.corr.odd`	the asymptotic correlations between SPU test statistics with odd γ.
`spu.corr.even`	the asymptotic correlations between SPU test statistics with even γ.
`cov.assumption`	the equality assumption on the covariances of the two sample populations; this reminders users that `cpval_aSPU()` assumes that the two sample groups have same covariance.
`method`	this output reminds users that the p-values are obtained using the asymptotic distributions of test statistics.
`pval`	the p-values of the SPU tests and the aSPU test.

Note

The permutation technique assumes that the distributions of the two sample populations are the same under the null hypothesis.

References

Bickel PJ and Levina E (2008). "Regularized estimation of large covariance matrices." *The Annals of Statistics*, **36**(1), 199–227.

Pan W, Kim J, Zhang Y, Shen X, and Wei P (2014). "A powerful and adaptive association test for rare variants." *Genetics*, **197**(4), 1081–1095.

Pourahmadi M (2013). *High-Dimensional Covariance Estimation*. John Wiley & Sons, Hoboken, NJ.

Xu G, Lin L, Wei P, and Pan W (2016). "An adaptive two-sample test for high-dimensional means." *Biometrika*, **103**(3), 609–624.

See Also

apval_aSPU, epval_aSPU

Examples

```
library(MASS)
set.seed(1234)
n1 <- n2 <- 50
p <- 200
mu1 <- rep(0, p)
mu2 <- mu1
mu2[1:10] <- 0.2
true.cov <- 0.4^(abs(outer(1:p, 1:p, "-"))) # AR1 covariance
sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov)
sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov)
cpval_aSPU(sam1, sam2, n.iter = 100)
```

---

epval_aSPU *Empirical Permutation- or Resampling-Based p-values of the SPU and aSPU Tests*

Description

Calculates p-values of the sum-of-powers (SPU) and adaptive SPU (aSPU) tests based on permutation or parametric bootstrap resampling.

Usage

```
epval_aSPU(sam1, sam2, pow = c(1:6, Inf), eq.cov = TRUE, n.iter = 1000, cov1.est,
           cov2.est, bandwidth1, bandwidth2, cv.fold = 5, norm = "F", seeds)
```

Arguments

sam1	an n1 by p matrix from sample population 1. Each row represents a p-dimensional sample.
sam2	an n2 by p matrix from sample population 2. Each row represents a p-dimensional sample.
pow	a numeric vector indicating the candidate values for the power γ in SPU tests. It should contain Inf and both odd and even integers. The default is c(1:6, Inf).
eq.cov	a logical value. The default is TRUE, indicating that the two sample populations have same covariance; otherwise, the covariances are assumed to be different. If eq.cov is TRUE, the permutation method is used to calculate p-values; otherwise, the parametric bootstrap resampling is used.
n.iter	a numeric integer indicating the number of permutation/resampling iterations. The default is 1,000.
cov1.est	This and the following arguments are only effective when eq.cov = FALSE and the parametric bootstrap resampling is used to calculate p-values. This argument specifies a consistent estimate of the covariance matrix of sample population 1 when eq.cov is FALSE. This can be obtained from various apporoaches (e.g., banding, tapering, and thresholding; see Pourahmadi 2013). If not specified, this function uses a banding approach proposed by Bickel and Levina (2008) to estimate the covariance matrix.
cov2.est	a consistent estimate of the covariance matrix of sample population 2 when eq.cov is FALSE. It is similar with the argument cov1.est.
bandwidth1	a vector of nonnegative integers indicating the candidate bandwidths to be used in the banding approach (Bickel and Levina, 2008) for estimating the covariance of sample population 1 when eq.cov is FALSE. This argument is effective when cov1.est is not provided. The default is a vector containing 50 candidate bandwidths chosen from {0, 1, 2, ..., p}.
bandwidth2	similar with the argument bandwidth1; it is used to specify candidate bandwidths for estimating the covariance of sample population 2 when eq.cov is FALSE.
cv.fold	an integer greater than or equal to 2 indicating the fold of cross-validation. The default is 5. See page 211 in Bickel and Levina (2008).
norm	a character string indicating the type of matrix norm for the calculation of risk function in cross-validation. This argument will be passed to the norm function. The default is the Frobenius norm ("F").
seeds	a vector of seeds for each permutation or parametric bootstrap resampling iteration; this is optional.

Details

See the details in apval_aSPU.

Value

A list including the following elements:

sam.info	the basic information about the two groups of samples, including the samples sizes and dimension.
pow	the powers γ used for the SPU tests.
opt.bw1	the optimal bandwidth determined by the cross-validation when eq.cov was FALSE and cov1.est was not specified.
opt.bw2	the optimal bandwidth determined by the cross-validation when eq.cov was FALSE and cov2.est was not specified.
cov.assumption	the equality assumption on the covariances of the two sample populations; this was specified by the argument eq.cov.
method	this output reminds users that the p-values are obtained using permutation or parametric bootstrap resampling.
pval	the p-values of the SPU tests and the aSPU test.

References

Bickel PJ and Levina E (2008). "Regularized estimation of large covariance matrices." *The Annals of Statistics*, **36**(1), 199–227.

Pan W, Kim J, Zhang Y, Shen X, and Wei P (2014). "A powerful and adaptive association test for rare variants." *Genetics*, **197**(4), 1081–1095.

Pourahmadi M (2013). *High-Dimensional Covariance Estimation*. John Wiley & Sons, Hoboken, NJ.

Xu G, Lin L, Wei P, and Pan W (2016). "An adaptive two-sample test for high-dimensional means." *Biometrika*, **103**(3), 609–624.

See Also

apval_aSPU, cpval_aSPU

Examples

```
library(MASS)
set.seed(1234)
n1 <- n2 <- 50
p <- 200
mu1 <- rep(0, p)
mu2 <- mu1
mu2[1:10] <- 0.2
true.cov <- 0.4^(abs(outer(1:p, 1:p, "-"))) # AR1 covariance
sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov)
sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov)
# increase n.iter to reduce Monte Carlo error
epval_aSPU(sam1, sam2, n.iter = 10)

# the two sample populations have different covariances
#true.cov1 <- 0.2^(abs(outer(1:p, 1:p, "-")))
#true.cov2 <- 0.6^(abs(outer(1:p, 1:p, "-")))
#sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov1)
#sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov2)
```

```
# increase n.iter to reduce Monte Carlo error
#epval_aSPU(sam1, sam2, eq.cov = FALSE, n.iter = 10,
# bandwidth1 = 10, bandwidth2 = 10)
```

`epval_Bai1996` *Empirical Permutation-Based p-value of the Test Proposed by Bai and Saranadasa (1996)*

Description

Calculates p-value of the test for testing equality of two-sample high-dimensional mean vectors proposed by Bai and Saranadasa (1996) based on permutation.

Usage

```
epval_Bai1996(sam1, sam2, n.iter = 1000, seeds)
```

Arguments

`sam1`	an n1 by p matrix from sample population 1. Each row represents a p-dimensional sample.
`sam2`	an n2 by p matrix from sample population 2. Each row represents a p-dimensional sample.
`n.iter`	a numeric integer indicating the number of permutation iterations. The default is 1,000.
`seeds`	a vector of seeds for each permutation or parametric bootstrap resampling iteration; this is optional.

Details

See the details in `apval_Bai1996`.

Value

A list including the following elements:

`sam.info`	the basic information about the two groups of samples, including the samples sizes and dimension.
`cov.assumption`	this output reminds users that the two sample populations have a common covariance matrix.
`method`	this output reminds users that the p-values are obtained using permutation.
`pval`	the p-value of the test proposed by Bai and Saranadasa (1996).

Note

The permutation technique assumes that the distributions of the two sample populations are the same under the null hypothesis.

References

Bai ZD and Saranadasa H (1996). "Effect of high dimension: by an example of a two sample problem." *Statistica Sinica*, **6**(2), 311–329.

See Also

apval_Bai1996

Examples

```
#library(MASS)
#set.seed(1234)
#n1 <- n2 <- 50
#p <- 200
#mu1 <- rep(0, p)
#mu2 <- mu1
#mu2[1:10] <- 0.2
#true.cov <- 0.4^(abs(outer(1:p, 1:p, "-"))) # AR1 covariance
#sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov)
#sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov)
# increase n.iter to reduce Monte Carlo error.
#epval_Bai1996(sam1, sam2, n.iter = 10)
```

---

epval_Cai2014 *Empirical Permutation- or Resampling-Based p-value of the Test Proposed by Cai et al (2014)*

Description

Calculates p-value of the test for testing equality of two-sample high-dimensional mean vectors proposed by Cai et al (2014) based on permutation or parametric bootstrap resampling.

Usage

```
epval_Cai2014(sam1, sam2, eq.cov = TRUE, n.iter = 1000, cov1.est, cov2.est,
              bandwidth1, bandwidth2, cv.fold = 5, norm = "F", seeds)
```

Arguments

sam1	an n1 by p matrix from sample population 1. Each row represents a p-dimensional sample.
sam2	an n2 by p matrix from sample population 2. Each row represents a p-dimensional sample.
eq.cov	a logical value. The default is TRUE, indicating that the two sample populations have same covariance; otherwise, the covariances are assumed to be different. If eq.cov is TRUE, the permutation method is used to calculate p-values; otherwise, the parametric bootstrap resampling is used.

n.iter	a numeric integer indicating the number of permutation/resampling iterations. The default is 1,000.
cov1.est	This and the following arguments are only effective when eq.cov = FALSE and the parametric bootstrap resampling is used to calculate p-values. This argument specifies a consistent estimate of the covariance matrix of sample population 1 when eq.cov is FALSE. This can be obtained from various apporoaches (e.g., banding, tapering, and thresholding; see Pourahmadi 2013). If not specified, this function uses a banding approach proposed by Bickel and Levina (2008) to estimate the covariance matrix.
cov2.est	a consistent estimate of the covariance matrix of sample population 2 when eq.cov is FALSE. It is similar with the argument cov1.est.
bandwidth1	a vector of nonnegative integers indicating the candidate bandwidths to be used in the banding approach (Bickel and Levina, 2008) for estimating the covariance of sample population 1 when eq.cov is FALSE. This argument is effective when cov1.est is not provided. The default is a vector containing 50 candidate bandwidths chosen from {0, 1, 2, ..., p}.
bandwidth2	similar with the argument bandwidth1; it is used to specify candidate bandwidths for estimating the covariance of sample population 2 when eq.cov is FALSE.
cv.fold	an integer greater than or equal to 2 indicating the fold of cross-validation. The default is 5. See page 211 in Bickel and Levina (2008).
norm	a character string indicating the type of matrix norm for the calculation of risk function in cross-validation. This argument will be passed to the norm function. The default is the Frobenius norm ("F").
seeds	a vector of seeds for each permutation or parametric bootstrap resampling iteration; this is optional.

Details

See the details in apval_Cai2014.

Value

A list including the following elements:

sam.info	the basic information about the two groups of samples, including the samples sizes and dimension.
opt.bw1	the optimal bandwidth determined by the cross-validation when eq.cov was FALSE and cov1.est was not specified.
opt.bw2	the optimal bandwidth determined by the cross-validation when eq.cov was FALSE and cov2.est was not specified.
cov.assumption	the equality assumption on the covariances of the two sample populations; this was specified by the argument eq.cov.
method	this output reminds users that the p-values are obtained using permutation or parametric bootstrap resampling.
pval	the p-value of the test proposed by Cai et al (2014).

Note

This function does not transform the data with their precision matrix (see Cai et al, 2014). To calculate the p-value of the test statisic with transformation, users can input transformed samples to `sam1` and `sam2`.

References

Bickel PJ and Levina E (2008). "Regularized estimation of large covariance matrices." *The Annals of Statistics*, **36**(1), 199–227.

Cai TT, Liu W, and Xia Y (2014). "Two-sample test of high dimensional means under dependence." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(2), 349–372.

Pourahmadi M (2013). *High-Dimensional Covariance Estimation*. John Wiley & Sons, Hoboken, NJ.

See Also

`apval_Cai2014`

Examples

```
#library(MASS)
#set.seed(1234)
#n1 <- n2 <- 50
#p <- 200
#mu1 <- rep(0, p)
#mu2 <- mu1
#mu2[1:10] <- 0.2
#true.cov <- 0.4^(abs(outer(1:p, 1:p, "-"))) # AR1 covariance
#sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov)
#sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov)
# increase n.iter to reduce Monte Carlo error
#epval_Cai2014(sam1, sam2, n.iter = 10)

# the two sample populations have different covariances
#true.cov1 <- 0.2^(abs(outer(1:p, 1:p, "-")))
#true.cov2 <- 0.6^(abs(outer(1:p, 1:p, "-")))
#sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov1)
#sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov2)
# increase n.iter to reduce Monte Carlo error
#epval_Cai2014(sam1, sam2, eq.cov = FALSE, n.iter = 10,
# bandwidth1 = 10, bandwidth2 = 10)
```

`epval_Chen2010` *Empirical Permutation- or Resampling-Based p-value of the Test Proposed by Chen and Qin (2010)*

Description

Calculates p-value of the test for testing equality of two-sample high-dimensional mean vectors proposed by Chen and Qin (2010) based on permutation or parametric bootstrap resampling.

Usage

```
epval_Chen2010(sam1, sam2, eq.cov = TRUE, n.iter = 1000, cov1.est, cov2.est,
               bandwidth1, bandwidth2, cv.fold = 5, norm = "F", seeds)
```

Arguments

sam1	an n1 by p matrix from sample population 1. Each row represents a p-dimensional sample.
sam2	an n2 by p matrix from sample population 2. Each row represents a p-dimensional sample.
eq.cov	a logical value. The default is TRUE, indicating that the two sample populations have same covariance; otherwise, the covariances are assumed to be different. If eq.cov is TRUE, the permutation method is used to calculate p-values; otherwise, the parametric bootstrap resampling is used.
n.iter	a numeric integer indicating the number of permutation/resampling iterations. The default is 1,000.
cov1.est	This and the following arguments are only effective when eq.cov = FALSE and the parametric bootstrap resampling is used to calculate p-values. This argument specifies a consistent estimate of the covariance matrix of sample population 1 when eq.cov is FALSE. This can be obtained from various apporoaches (e.g., banding, tapering, and thresholding; see Pourahmadi 2013). If not specified, this function uses a banding approach proposed by Bickel and Levina (2008) to estimate the covariance matrix.
cov2.est	a consistent estimate of the covariance matrix of sample population 2 when eq.cov is FALSE. It is similar with the argument cov1.est.
bandwidth1	a vector of nonnegative integers indicating the candidate bandwidths to be used in the banding approach (Bickel and Levina, 2008) for estimating the covariance of sample population 1 when eq.cov is FALSE. This argument is effective when cov1.est is not provided. The default is a vector containing 50 candidate bandwidths chosen from {0, 1, 2, ..., p}.
bandwidth2	similar with the argument bandwidth1; it is used to specify candidate bandwidths for estimating the covariance of sample population 2 when eq.cov is FALSE.
cv.fold	an integer greater than or equal to 2 indicating the fold of cross-validation. The default is 5. See page 211 in Bickel and Levina (2008).
norm	a character string indicating the type of matrix norm for the calculation of risk function in cross-validation. This argument will be passed to the norm function. The default is the Frobenius norm ("F").
seeds	a vector of seeds for each permutation or parametric bootstrap resampling iteration; this is optional.

Details

See the details in apval_Chen2010.

Value

A list including the following elements:

sam.info	the basic information about the two groups of samples, including the samples sizes and dimension.
opt.bw1	the optimal bandwidth determined by the cross-validation when eq.cov was FALSE and cov1.est was not specified.
opt.bw2	the optimal bandwidth determined by the cross-validation when eq.cov was FALSE and cov2.est was not specified.
cov.assumption	the equality assumption on the covariances of the two sample populations; this was specified by the argument eq.cov.
method	this output reminds users that the p-values are obtained using permutation or parametric bootstrap resampling.
pval	the p-value of the test proposed by Chen and Qin (2010).

References

Bickel PJ and Levina E (2008). "Regularized estimation of large covariance matrices." *The Annals of Statistics*, **36**(1), 199–227.

Chen SX and Qin YL (2010). "A two-sample test for high-dimensional data with applications to gene-set testing." *The Annals of Statistics*, **38**(2), 808–835.

Pourahmadi M (2013). *High-Dimensional Covariance Estimation*. John Wiley & Sons, Hoboken, NJ.

See Also

apval_Chen2010

Examples

```
#library(MASS)
#set.seed(1234)
#n1 <- n2 <- 50
#p <- 200
#mu1 <- rep(0, p)
#mu2 <- mu1
#mu2[1:10] <- 0.2
#true.cov <- 0.4^(abs(outer(1:p, 1:p, "-"))) # AR1 covariance
#sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov)
#sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov)
# increase n.iter to reduce Monte Carlo error.
#epval_Chen2010(sam1, sam2, n.iter = 10)

# the two sample populations have different covariances
```

```
#true.cov1 <- 0.2^(abs(outer(1:p, 1:p, "-")))
#true.cov2 <- 0.6^(abs(outer(1:p, 1:p, "-")))
#sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov1)
#sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov2)
# increase n.iter to reduce Monte Carlo error
#epval_Chen2010(sam1, sam2, eq.cov = FALSE, n.iter = 10,
# bandwidth1 = 10, bandwidth2 = 10)
```

`epval_Chen2014` *Empirical Permutation- or Resampling-Based p-value of the Test Proposed by Chen et al (2014)*

Description

Calculates p-value of the test for testing equality of two-sample high-dimensional mean vectors proposed by Chen et al (2014) based on permutation or parametric bootstrap resampling.

Usage

```
epval_Chen2014(sam1, sam2, eq.cov = TRUE, n.iter = 1000, cov1.est, cov2.est,
               bandwidth1, bandwidth2, cv.fold = 5, norm = "F", seeds)
```

Arguments

`sam1`	an n1 by p matrix from sample population 1. Each row represents a p-dimensional sample.
`sam2`	an n2 by p matrix from sample population 2. Each row represents a p-dimensional sample.
`eq.cov`	a logical value. The default is `TRUE`, indicating that the two sample populations have same covariance; otherwise, the covariances are assumed to be different. If `eq.cov` is `TRUE`, the permutation method is used to calculate p-values; otherwise, the parametric bootstrap resampling is used.
`n.iter`	a numeric integer indicating the number of permutation/resampling iterations. The default is 1,000.
`cov1.est`	This and the following arguments are only effective when `eq.cov = FALSE` and the parametric bootstrap resampling is used to calculate p-values. This argument specifies a consistent estimate of the covariance matrix of sample population 1 when `eq.cov` is `FALSE`. This can be obtained from various apporoaches (e.g., banding, tapering, and thresholding; see Pourahmadi 2013). If not specified, this function uses a banding approach proposed by Bickel and Levina (2008) to estimate the covariance matrix.
`cov2.est`	a consistent estimate of the covariance matrix of sample population 2 when `eq.cov` is `FALSE`. It is similar with the argument `cov1.est`.

bandwidth1	a vector of nonnegative integers indicating the candidate bandwidths to be used in the banding approach (Bickel and Levina, 2008) for estimating the covariance of sample population 1 when eq.cov is FALSE. This argument is effective when cov1.est is not provided. The default is a vector containing 50 candidate bandwidths chosen from {0, 1, 2, ..., p}.
bandwidth2	similar with the argument bandwidth1; it is used to specify candidate bandwidths for estimating the covariance of sample population 2 when eq.cov is FALSE.
cv.fold	an integer greater than or equal to 2 indicating the fold of cross-validation. The default is 5. See page 211 in Bickel and Levina (2008).
norm	a character string indicating the type of matrix norm for the calculation of risk function in cross-validation. This argument will be passed to the norm function. The default is the Frobenius norm ("F").
seeds	a vector of seeds for each permutation or parametric bootstrap resampling iteration; this is optional.

Details

See the details in apval_Chen2014.

Value

A list including the following elements:

sam.info	the basic information about the two groups of samples, including the samples sizes and dimension.
opt.bw1	the optimal bandwidth determined by the cross-validation when eq.cov was FALSE and cov1.est was not specified.
opt.bw2	the optimal bandwidth determined by the cross-validation when eq.cov was FALSE and cov2.est was not specified.
cov.assumption	the equality assumption on the covariances of the two sample populations; this was specified by the argument eq.cov.
method	this output reminds users that the p-values are obtained using permutation or parametric bootstrap resampling.
pval	the p-value of the test proposed by Chen et al (2014).

Note

This function does not transform the data with their precision matrix (see Chen et al, 2014). To calculate the p-value of the test statisic with transformation, users can input transformed samples to sam1 and sam2.

References

Bickel PJ and Levina E (2008). "Regularized estimation of large covariance matrices." *The Annals of Statistics*, **36**(1), 199–227.

Chen SX, Li J, and Zhong PS (2014). "Two-Sample Tests for High Dimensional Means with Thresholding and Data Transformation." arXiv preprint arXiv:1410.2848.

Pourahmadi M (2013). *High-Dimensional Covariance Estimation*. John Wiley & Sons, Hoboken, NJ.

See Also

apval_Chen2014

Examples

```
#library(MASS)
#set.seed(1234)
#n1 <- n2 <- 50
#p <- 200
#mu1 <- rep(0, p)
#mu2 <- mu1
#mu2[1:10] <- 0.2
#true.cov <- 0.4^(abs(outer(1:p, 1:p, "-"))) # AR1 covariance
#sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov)
#sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov)
# increase n.iter to reduce Monte Carlo error
#epval_Chen2014(sam1, sam2, n.iter = 10)

# the two sample populations have different covariances
#true.cov1 <- 0.2^(abs(outer(1:p, 1:p, "-")))
#true.cov2 <- 0.6^(abs(outer(1:p, 1:p, "-")))
#sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov1)
#sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov2)
# increase n.iter to reduce Monte Carlo error
#epval_Chen2014(sam1, sam2, eq.cov = FALSE, n.iter = 10,
# bandwidth1 = 10, bandwidth2 = 10)
```

---

epval_Sri2008 — *Empirical Permutation-Based p-value of the Test Proposed by Srivastava and Du (2008)*

Description

Calculates p-value of the test for testing equality of two-sample high-dimensional mean vectors proposed by Srivastava and Du (1996) based on permutation.

Usage

```
epval_Sri2008(sam1, sam2, n.iter = 1000, seeds)
```

Arguments

sam1	an n1 by p matrix from sample population 1. Each row represents a p-dimensional sample.
sam2	an n2 by p matrix from sample population 2. Each row represents a p-dimensional sample.
n.iter	a numeric integer indicating the number of permutation iterations. The default is 1,000.
seeds	a vector of seeds for each permutation or parametric bootstrap resampling iteration; this is optional.

Details

See the details in apval_Sri2008.

Value

A list including the following elements:

sam.info	the basic information about the two groups of samples, including the samples sizes and dimension.
cov.assumption	this output reminds users that the two sample populations have a common covariance matrix.
method	this output reminds users that the p-values are obtained using permutation.
pval	the p-value of the test proposed by Srivastava and Du (2008).

Note

The permutation technique assumes that the distributions of the two sample populations are the same under the null hypothesis.

References

Srivastava MS and Du M (2008). "A test for the mean vector with fewer observations than the dimension." *Journal of Multivariate Analysis*, **99**(3), 386–402.

See Also

apval_Sri2008

Examples

```
#library(MASS)
#set.seed(1234)
#n1 <- n2 <- 50
#p <- 200
#mu1 <- rep(0, p)
#mu2 <- mu1
#mu2[1:10] <- 0.2
```

```
#true.cov <- 0.4^(abs(outer(1:p, 1:p, "-"))) # AR1 covariance
#sam1 <- mvrnorm(n = n1, mu = mu1, Sigma = true.cov)
#sam2 <- mvrnorm(n = n2, mu = mu2, Sigma = true.cov)
# increase n.iter to reduce Monte Carlo error.
#epval_Sri2008(sam1, sam2, n.iter = 10)
```

Index