Package 'ARHT'

March 27, 2018

Type Package

Title Adaptable Regularized Hotelling's T^2 Test for High-Dimensional Data

Version 0.1.0

Description Perform the Adaptable Regularized Hotelling's T^2 test (ARHT) proposed by Li et al., (2016) <arXiv:1609.08725>. Both one-sample and two-sample mean test are available with
various probabilistic alternative prior models. It contains a function to consistently estimate higher order moments of the population covariance spectral distribution using the spectral of the sample covariance matrix (Bai et al. (2010) <doi:10.1111/j.1467-842X.2010.00590.x>).
In addition, it contains a function to sample from 3-variate chi-squared random vectors approximately
with a given correlation matrix when the degrees of freedom are large.

License GPL (>= 2)

Encoding UTF-8

LazyData true

Depends R (>= 2.10)

Imports stats

RoxygenNote 6.0.1

Suggests testthat

NeedsCompilation no

Author Haoran Li [aut, cre]

Maintainer Haoran Li <hrli@ucdavis.edu>

Repository CRAN

Date/Publication 2018-03-27 15:47:55 UTC

R topics documented:

ARHT *An adaptable generalized Hotelling's T^2 test for high dimensional data*

Description

This function performs the adaptable regularized Hotelling's T^2 test (ARHT) (Li et al., (2016) <arXiv:1609.08725>) for the one-sample and two-sample test problem, where we're interested in detecting the mean vector in the one-sample problem or the difference between mean vectors in the two-sample problem in a high dimensional regime.

Usage

```
ARHT(X, Y = NULL, mu_0 = NULL, prob_alt_prior = list(c(1, 0, 0), c(0, 1,
  0), c(0, 0, 1)), Type1error_calib = c("cube_root", "sqrt", "chi_sq",
  "none"), lambda_range = NULL, nlambda = 2000, bs_size = 1e+05)
```

Arguments

`X`	the n1-by-p observation matrix with numeric column variables.
`Y`	an optional n2-by-p observation matrix; if `NULL`, a one-sample test is conducted on `X`; otherwise, a two-sample test is conducted on `X` and `Y`.
`mu_0`	the null hypothesis vector to be tested; if `NULL`, the default value is the 0 vector of length p.
`prob_alt_prior`	a non-empty list; Each field is a numeric vector with sum 1. The default value is the "canonical weights" `list(c(1,0,0), c(0,1,0), c(0,0,1))`; Each field represents a probabilistic prior model specified by weights of I_p, Σ, Σ^2, etc, where Σ is the population covariance matrix of the observations.
`Type1error_calib`	the method to calibrate Type 1 error rate of ARHT. Choose its first element when more than one are specified. Four values are allowed: • `cube_root` The default value; cube-root transformation; • `sqrt` Square-root transformation; • `chi_sq` Chi-square approximation, not available when more than three models are specified in `prob_alt_prior`; • `none` No calibration.
`lambda_range`	optional user-supplied lambda range; If `NULL`, ARHT chooses its own range.
`nlambda`	optional user-supplied number of lambda's in grid search; default to be `2000`; the grid is progressively coarser.
`bs_size`	positive numeric with default value `1e5`; only effective when more than one prior models are specified in `prob_alt_prior`; control the size of the bootstrap sample used to approximate the ARHT p-value.

Details

The method incorporates ridge-regularization in the classic Hotelling's T^2 test with the regularization parameter chosen such that the asymptotic power under a class of probabilistic alternative prior models is maximized. ARHT combines different prior models by taking the maximum of statistics under all models. ARHT is distributed as the maximum of a correlated multivariate normal random vector. We estimate its covariance matrix and bootstrap its distribution. The returned p-value is a Monte Carlo approximation to its true value using the bootstrap sample, therefore not deterministic. Various methods are available to calibrate the slightly inflated Type 1 error rate of ARHT, including Cube-root transformation, square-root transformation and chi-square approximation.

Value

- `ARHT_pvalue`: The p-value of ARHT test.
  - If `length(prob_alt_prior)==1`, it is identical to `RHT_pvalue`.
  - If `length(prob_alt_prior)>1`, it is the p-value after combining results from all prior models. The value is bootstrapped, therefore not deterministic.
- `RHT_opt_lambda`: The optimal lambda's chosen under each of the prior models in `prob_alt_prior`. It has the same length and order as `prob_alt_prior`.
- `RHT_pvalue`: The p-value of RHT tests with the lambda's in `RHT_opt_lambda`.
- `RHT_std`: The standardized RHT statistics with the lambda's in `RHT_opt_lambda`. Take its maximum to get the statistic of ARHT test.
- `Theta1`: As defined in Li et al. (2016) <arXiv:1609.08725>, the estimated asymptotic means of RHT statistics with the lambda's in `RHT_opt_lambda`.
- `Theta2`: As defined in Li et al. (2016) <arXiv:1609.08725>, `2*Theta2` are the estimated asymptotic variances of RHT statistics the lambda's in `RHT_opt_lambda`.
- `Corr_RHT`: The estimated correlation matrix of the statistics in `RHT_std`.

References

Li, H. Aue, A., Paul, D. Peng, J., & Wang, P. (2016). *An adaptable generalization of Hotelling's T^2 test in high dimension.* <arXiv:1609:08725>.

Chen, L., Paul, D., Prentice, R., & Wang, P. (2011). *A regularized Hotelling's T^2 test for pathway analysis in proteomic studies.* Journal of the American Statistical Association, 106(496), 1345-1360.

Examples

```
set.seed(10086)
# One-sample test
n1 = 300; p =500
dataX = matrix(rnorm(n1 * p), nrow = n1, ncol = p)
res1 = ARHT(dataX)

# Two-sample test
n2= 400
dataY = matrix(rnorm(n2 * p), nrow = n2, ncol = p )
res2 = ARHT(dataX, dataY, mu_0 = rep(0.01,p))
```

```
# Specify probabilistic alternative priors model
res3 = ARHT(dataX, dataY, mu_0 = rep(0.01,p),
     prob_alt_prior = list(c(1/3, 1/3, 1/3), c(0,1,0)))

# Change Type 1 error calibration method
res4 = ARHT(dataX, dataY, mu_0 = rep(0.01,p),
     Type1error_calib = "sqrt")

RejectOrNot = res4$ARHT_pvalue < 0.05
```

`moments_PSD` *Consistent estimators of high-order moments of the population spectral distribution for high-dimensional data*

Description

The function calculates consistent estimators of moments of the spectral distribution of the population covariance matrix given the spectral of the sample covariance matrix.

Usage

```
moments_PSD(eigenvalues, n, mom_degree)
```

Arguments

`eigenvalues`	all eigenvalues of the sample covariance matrix including 0's.
`n`	degree of freedom of the sample covariance matrix.
`mom_degree`	the maximum order of moments.

Value

Estimators of moments from the first to the `mom_degree` -th order.

References

Bai, Z., Chen, J., & Yao, J. (2010). *On estimation of the population spectral distribution from a high-dimensional sample covariance matrix.* Australian & New Zealand Journal of Statistics, 52(4), 423-437.

Examples

```
set.seed(10086)
n = 400; p= 500
pop_eig = seq(10,1,length = p)
# Data with covariance matrix diag(pop_eig)
Z = matrix(rnorm(n*p),n,p)
```

```
X = Z %*% diag(sqrt(pop_eig))
raw_eig = svd(cov(X))$d
emp_eig = raw_eig[raw_eig>=0]
# Moments of population spectral distribution
colMeans(outer(pop_eig, 1:4, "^"))
# Estimators
moments_PSD(emp_eig, n-1, 4)
```

`r3chisq` *3-variate positively correlated chi-squared sample generation when degrees of freedom are large*

Description

Generate samples approximately from three positively correlated chi-squared random variables $(\chi^2(d_1), \chi^2(d_2), \chi^2(d_3))$ when the degrees of freedom (d_1, d_2, d_3) are large.

Usage

```
r3chisq(size, df, corr_mat)
```

Arguments

`size`	sample size.
`df`	the degree of freedoms of the marginal distributions. Must be non-negative, but can be non-integer. The function uses `ceiling(df)` if non-integer.
`corr_mat`	the target correlation matrix; negative elements will be set to 0.

Details

It is generally hard to sample from $(\chi^2(d_1), \chi^2(d_2), \chi^2(d_3))$ with a designed correlation matrix. In the algorithm, we approximate the random vector by $(z^TQ_1z, z^TQ_2z, z^TQ_3z)$ where z is a standard norm random vector and Q_1, Q_2, Q_3 are diagonal matrices with diagonal elements 1's and 0's. The designed positive correlations is approximated by carefully selecting common locations of 1's on the diagonals. The generated sample may have slightly larger marginal degrees of freedom than the inputted `df`, also slightly different covariances.

Value

- `sample`: a `size`-by-3 matrix contains the generated sample.
- `approx_cov`: the true covariance matrix of `sample`.

References

Li, H., Aue, A., Paul, D., Peng, J., & Wang, P. (2016). *An adaptable generalization of Hotelling's T^2 test in high dimension.* arXiv preprint <arXiv:1609.08725>.

Examples

```
set.seed(10086)
cor_examp = matrix(c(1,1/6,2/3,1/6,1,2/3,2/3,2/3,1),3,3)
a_sam = r3chisq(size = 10000,
                df = c(80,90,100),
                corr_mat = cor_examp)
cov(a_sam$sample) - a_sam$approx_cov
cov2cor(a_sam$approx_cov) - cor_examp
```

Index