Package 'distanceHD'

October 18, 2026

Type Package

Title Distance Metrics for High-Dimensional Clustering

Version 1.4

Date 2026-10-18

Description We provide three distance metrics for measuring the separation between two clusters in high-dimensional spaces. The first metric is the centroid distance, which calculates the Euclidean distance between the centers of the two groups. The second is a ridge Mahalanobis distance, which incorporates a ridge correction constant, alpha, to ensure that the covariance matrix is invertible. The third metric is the maximal data piling distance, which computes the orthogonal distance between the affine spaces spanned by each class. These three distances are asymptotically interconnected and are applicable in tasks such as discrimination, clustering, and outlier detection in high-dimensional settings. The package also provides a distance-based outlier detection procedure for high-dimensional, low-sample-size data, as described in Ahn, Lee and Lee (2018).

License GPL (>= 2)

Depends R (>= 3.5.0), MASS

Imports graphics, stats

Encoding UTF-8

NeedsCompilation no

Author Jung Ae Lee [aut, cre],
Jeongyoun Ahn [aut]

Maintainer Jung Ae Lee <jungaeleeb@gmail.com>

Repository CRAN

Date/Publication 2026-10-18 15:10:08 UTC

Contents

distanceHD-package 2
dist_cen 2
dist_mah 3
dist_mdp 4

HDout 5
leukemia 7
simdata 8

Index **9**

`distanceHD-package` *Distance Metrics for High-Dimensional Clustering*

Description

We provide three distance metrics for measuring the separation between two clusters in high-dimensional spaces. The first metric is the centroid distance, which calculates the Euclidean distance between the centers of the two groups. The second is a ridge Mahalanobis distance, which incorporates a ridge correction constant, alpha, to ensure that the covariance matrix is invertible. The third metric is the maximal data piling distance, which computes the orthogonal distance between the affine spaces spanned by each class. These three distances are asymptotically interconnected and are applicable in tasks such as discrimination, clustering, and outlier detection in high-dimensional settings. The package also provides a distance-based outlier detection procedure for high-dimensional, low-sample-size data, as described in Ahn, Lee and Lee (2018).

Author(s)

Jung Ae Lee <jungae.lee@umassmed.edu>; Jeongyoun Ahn <jyahn@kaist.ac.kr>

Maintainer: Jung Ae Lee <jungaeleeb@gmail.com>

References

1. Ahn J, Marron JS (2010). The maximal data piling direction for discrimination. Biometrika, 97(1):254-259.

2. Ahn J, Lee MH, Yoon YJ (2012). Clustering high dimension, low sample size data using the maximal data piling distance. Statistica Sinica, 22(2):443-464.

3. Ahn J, Lee MH, Lee JA (2019). Distance-based outlier detection for high dimension, low sample size data. Journal of Applied Statistics.46(1):13-29.

`dist_cen` *Centroid distance between two groups*

Description

Calculate the Centroid (Euclidean) distance between two groups in a high-dimensional space (d > n), with support for a single-member cluster. It also works in low-dimensional settings.

Usage

```
dist_cen(x, group)
```

Arguments

x	x is n by d matrix
group	group is a binary group label with the length of n1 and n2

Value

A numeric value of distance

Author(s)

Jung Ae Lee <jungae.lee@umassmed.edu>

Examples

```
data(leukemia)
group = leukemia$Y      # 38 patients status with a value of 1 or 2
x = leukemia$X          # 38 by 3051 genes

# apply the function
dist_cen(x, group)  # 25.4
```

---

dist_mah *ridge Mahalanobis distance between two groups*

Description

Calculate the Mahalanobis distance between two groups in a high-dimensional space (d > n), using a ridge correction on the covariance matrix to ensure invertibility. The method also supports a single-member cluster and works in low-dimensional settings without a ridge correction.

Usage

```
dist_mah(x, group, alpha)
```

Arguments

x	x is n by d matrix
group	group is a binary group label with the length of n1 and n2
alpha	alpha is a positive numeric value representing the ridge correction constant. If not specified, the default value is set to $\sqrt{\log(d)/n}$.

Value

A numeric value of distance

Author(s)

Jung Ae Lee <jungae.lee@umassmed.edu>

Examples

```
data(leukemia)
group = leukemia$Y      # 38 patients status with a value of 1 or 2
x = leukemia$X          # 38 x 3051 genes

# apply the function
dist_mah(x, group)   # 26.8

# default alpha
d = 3051; n = 38
alpha = sqrt(log(d)/n)
dist_mah(x, group, alpha)   # 26.8
```

`dist_mdp` *Maximal data piling (MDP) distance between two groups*

Description

Calculate the MDP (maximal data piling) distance between two groups in a high-dimensional space (d > n), with support for a single-member cluster. It also works in low-dimensional settings.

Usage

```
dist_mdp(x, group)
```

Arguments

x	x is n by d matrix
group	group is a binary group label with the length of n1 and n2

Value

A numeric value of distance

Author(s)

Jeongyoun Ahn <jyahn@kaist.ac.kr>

Examples

```
data(leukemia)
group = leukemia$Y      # 38 patients status with a value of 1 or 2
x = leukemia$X          # 38 x 3051 genes

# apply the function
dist_mah(x, group)   # 26.8
```

HDout *Outlier detection in high-dimensional data*

Description

Distance-based outlier detection for high dimension, low sample size data (Ahn, Lee and Lee, 2018). The observation farthest from the rest is flagged as a potential outlier and removed, one at a time, up to `Mout` observations. Each step is displayed by plotting the observed distances against the distances expected without outliers, estimated by a parametric bootstrap.

Usage

```
HDout(data, method = 1:3, Mout = 4, Nsim = 30, plotY = TRUE)
```

Arguments

`data`	`data` is a p by n data matrix (p > n).
`method`	choice of distance: `1` = centroid distance (`dist_cen`), `2` = ridge Mahalanobis (rM) distance (`dist_mah`), `3` = maximal data piling (MDP) distance (`dist_mdp`). The default `1:3` runs all three.
`Mout`	maximum number of outliers to test. The default is 4.
`Nsim`	number of bootstrap simulations. The default is 30; 10 is faster and 100 is more accurate.
`plotY`	if `TRUE` (default), the results are displayed in a plot.

Details

At each step (`Nout` = 1, ..., `Mout`), the leave-one-out distance of each remaining observation j is the distance between the single-member group {j} and the group of all other observations, computed with `dist_cen`, `dist_mah` or `dist_mdp`. For the rM distance, the ridge correction constant is $\sqrt{\log(p)/(m-1)}$, where m is the number of remaining observations. The observation with the largest distance is flagged as a potential outlier and removed.

The bootstrap samples are drawn from a normal distribution with the sample mean and covariance of the data without the flagged observation(s); when p >= n, a ridge term $\sqrt{\log(p)/m}\, I_p$ is added to the covariance. For each of the `Nsim` samples, the sorted leave-one-out distances are computed and then averaged.

The plot has one row per method and one column per step. Each panel plots the observed sorted distances (y-axis) against the average simulated sorted distances (x-axis); the flagged observation is shown as a filled point labelled with its column index. A flagged point lying far above the pattern of the others suggests an outlier.

Value

A list (returned invisibly) with the following elements:

- `dj_c`: sorted centroid distances at the last iteration (`Mout`).
- `dj_rM`: sorted rM distances at the last iteration.
- `dj_MDP`: sorted MDP distances at the last iteration.
- `outliers`: a list of the column indices flagged at steps 1, ..., `Mout`, one vector per method.

Distances are 0 for methods that were not run.

Author(s)

Jung Ae Lee <jungae.lee@umassmed.edu>

References

Ahn, J., Lee, M. H. and Lee, J. A. (2018). Distance-based outlier detection for high dimension, low sample size data, Journal of Applied Statistics. doi:10.1080/02664763.2018.1452901

See Also

`dist_cen`, `dist_mah`, `dist_mdp`

Examples

```
# simulated data example: p = 1000 variables, n = 19 observations
data(simdata)
dim(simdata)   # 1000 by 19

y = simdata    # p by n matrix (p >= n)

# first 200 variables, small Mout and Nsim keep the example fast;
# HDout(y) with the defaults (Mout = 4, Nsim = 30) takes longer
out = HDout(y[1:200, ], Mout = 2, Nsim = 5)
out$outliers   # observations flagged at each step, per method


# leukemia data: first 200 genes, transposed to genes (p) by patients (n);
# centroid distance only, without the plot
data(leukemia)
HDout(t(leukemia$X[, 1:200]), method = 1, Mout = 2, Nsim = 5, plotY = FALSE)$outliers
```

leukemia *Gene expression data from Golub et al. (1999)*

Description

Gene expression data (3051 genes and 38 tumor mRNA samples) from the leukemia microarray study of Golub et al. (1999).)

Usage

```
data(leukemia)
```

Format

A list with the following elements:

- X: a (38 by 3051) matrix giving the expression levels of 3051 genes for 38 leukemia patients. Each row corresponds to a patient, each column to a gene.
- Y: a numeric vector of length 38 giving the cancer class of each patient.
- gene.names: a matrix containing the names of the 3051 genes for the gene expression matrix X. The three columns correspond to the gene index, ID, and Name, respectively.

Source

The data are described in Golub et al. (1999) and can be freely downloaded from http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?paper_id=43.

References

S. Dudoit, J. Fridlyand and T. P. Speed (2002). Comparison of discrimination methods for the classification of tumors using gene expression data, Journal of the American Statistical Association 97, 77-87.

Golub et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286, 531-537.

See Also

plsgenomics::leukemia

Examples

```
data(leukemia)

# how many samples and genes?
dim(leukemia$X)

# how many samples of class 1 and 2, respectively?
table(leukemia$Y)
```

simdata *Simulated high-dimensional data for outlier detection*

Description

A simulated high dimension, low sample size data set with 1000 variables and 19 observations, used to illustrate the outlier detection procedure HDout.

Usage

```
data(simdata)
```

Format

A numeric matrix of dimension 1000 by 19 (p by n, p >= n). Each row corresponds to a variable and each column to an observation, which is the input format of HDout.

See Also

HDout

Examples

```
data(simdata)

# how many variables (p) and observations (n)?
dim(simdata)    # 1000 x 19


y = simdata     # p by n matrix (p >= n)

out = HDout(y[1:200, ], Mout = 2, Nsim = 5)   # first 200 variables, for speed
out$outliers   # observations flagged at each step, per method
```

Index

∗ **centroid distance**
distanceHD-package, 2
∗ **datasets**
leukemia, 7
simdata, 8
∗ **maximal data piling distance**
distanceHD-package, 2
∗ **ridge Mahalanobis distance**
distanceHD-package, 2

dist_cen, 2, *5, 6*
dist_centroid *(*dist_cen*)*, 2
dist_ma *(*dist_mah*)*, 3
dist_mah, 3, *5, 6*
dist_md *(*dist_mdp*)*, 4
dist_mdp, 4, *5, 6*
distanceHD *(*distanceHD-package*)*, 2
distanceHD-package, 2

HDout, 5, *8*

leukemia, 7

simdata, 8