Package 'distanceHD'

January 31, 2025

Type Package

Title Distance Metrics for High-Dimensional Clustering

Version 1.2

Date/Publication 2025-01-31 16:50:11 UTC

Maintainer Jung Ae Lee <jungaeleeb@gmail.com>

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.

License GPL (>= 2)

Depends R (>= 3.5.0), MASS

Encoding UTF-8

NeedsCompilation no

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

Repository CRAN

Contents

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.

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
```

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)
```

Index