Package 'RMLPCA'

January 20, 2025

Title Maximum Likelihood Principal Component Analysis

Version 0.0.1

Description R implementation of Maximum Likelihood Principal Component Analysis
The main idea of this package is to have an alternative way of PCA for
subspace modeling that considers measurement errors.
More details can be found in Peter D. Wentzell (2009)
<doi:10.1016/B978-0-444-64165-6.03029-9>.

URL https://github.com/renanestatcamp/RMLPCA

BugReports https://github.com/renanestatcamp/RMLPCA/issues

License MIT + file LICENSE

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

Suggests testthat, knitr, rmarkdown

Imports base, Matrix, pracma, RSpectra

Depends R (>= 2.10)

NeedsCompilation no

Author Renan Santos Barbosa [aut, cre]

Maintainer Renan Santos Barbosa <renansantosbarbosa@usp.br>

Repository CRAN

Date/Publication 2020-11-05 08:10:02 UTC

Contents

`cov_d` *Covariance matrix for mlpca_d model*

Description

A random covariance matrix to simulate data errors The main ideia is described in figure 3 on Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

Usage

```
cov_d
```

Format

A matrix with 20 rows and 20 columns

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

cov_e *Covariance matrices for mlpca_e model*

Description

A random array of covariance matrices to simulate data errors The main ideia is described in figure 3 on Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

Usage

```
cov_e
```

Format

An array of dimension 20,20,30

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

data_clean *Error free data for all examples.*

Description

A dataset generated by the rotation of a bivariate normal density, the method applied to get this dataset is described on Wentzell, P. D., and S. Hou. "Exploratory data analysis with noisy measurements." Journal of Chemometrics 26.6 (2012): 264-281.

Usage

```
data_clean
```

Format

A matrix with 300 rows and 20 columns

References

Wentzell, P. D., and S. Hou. "Exploratory data analysis with noisy measurements." Journal of Chemometrics 26.6 (2012): 264-281.

data_cleaned_mlpca_b *Cleaned dataset after applied MLPCA B used for tests only*

Description

A dataset where the values are estimated after mlpca_b is applied.

Usage

```
data_cleaned_mlpca_b
```

Format

A matrix with 300 rows and 20 columns

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

data_cleaned_mlpca_c *Cleaned dataset after applied MLPCA C used for tests only*

Description

A dataset where the values are estimated after mlpca_c is applied.

Usage

```
data_cleaned_mlpca_c
```

Format

A matrix with 300 rows and 20 columns

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

data_cleaned_mlpca_d *Cleaned dataset after applied MLPCA D used for tests only*

Description

A dataset where the values are estimated after mlpca_d is applied.

Usage

```
data_cleaned_mlpca_d
```

Format

A matrix with 300 rows and 20 columns

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

data_cleaned_mlpca_e *Cleaned dataset after applied MLPCA E used for tests only*

Description

A dataset where the values are estimated after mlpca_e is applied.

Usage

```
data_cleaned_mlpca_e
```

Format

A matrix with 30 rows and 20 columns

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

data_clean_e *Error free data for all examples.*

Description

A dataset generated by the rotation of a bivariate normal density, the method applied to get this dataset is described on Wentzell, P. D., and S. Hou. "Exploratory data analysis with noisy measurements." Journal of Chemometrics 26.6 (2012): 264-281.

Usage

```
data_clean_e
```

Format

A matrix with 30 rows and 20 columns

References

Wentzell, P. D., and S. Hou. "Exploratory data analysis with noisy measurements." Journal of Chemometrics 26.6 (2012): 264-281.

data_error_b *Errors generated for mlpca_b model*

Description

A dataset where each column contain values from a normal density with mean = 0 and standard deviation from 0.2 to 1, the standard deviations differs in the column. The main ideia is described in figure 3 on Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

Usage

```
data_error_b
```

Format

A matrix with 300 rows and 20 columns

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

data_error_c *Errors generated for mlpca_c model*

Description

A dataset where each column contain values from a normal density with mean = 0 and standard deviations simulated by a lognormal density with meanlog = -4.75 and sdlog = 2.5, all the standard deviations are different. The main ideia is described in figure 3 on Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

Usage

```
data_error_c
```

Format

A matrix with 300 rows and 20 columns

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

data_error_d *Errors generated for mlpca_d model*

Description

A dataset where the values come from a 20 -multivariate normal density where all the means are 0 and the covariance matrix from cov_d. The main ideia is described in figure 3 on Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

Usage

```
data_error_d
```

Format

A matrix with 300 rows and 20 columns

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

data_error_e *Errors generated for mlpca_e model*

Description

A dataset where the values come from a 20 -multivariate normal density where all the means are 0 and the covariance matrix from cov_e. The main ideia is described in figure 3 on Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

Usage

```
data_error_e
```

Format

A matrix with 30 rows and 20 columns

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

mlpca_b *Maximum likelihood principal component analysis for mode B error conditions*

Description

Performs maximum likelihood principal components analysis for mode B error conditions (independent errors, homoscedastic within a column). Equivalent to perfoming PCA on data scaled by the error SD, but results are rescaled to the original space.

Usage

```
mlpca_b(X, Xsd, p)
```

Arguments

X	MxN matrix of measurements.
Xsd	MxN matrix of measurements error standard deviations.
p	Rank of the model's subspace, p must be than the minimum of M and N.

Details

The returned parameters, U, S and V, are analogs to the truncated SVD solution, but have somewhat different properties since they represent the MLPCA solution. In particular, the solutions for different values of p are not necessarily nested (the rank 1 solution may not be in the space of the rank 2 solution) and the eigenvectors do not necessarily account for decreasing amounts of variance, since MLPCA is a subspace modeling technique and not a variance modeling technique.

Value

The parameters returned are the results of SVD on the estimated subspace. The quantity Ssq represents the sum of squares of weighted residuals. All the results are nested in a list format.

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

Examples

```
library(RMLPCA)
data(data_clean)
data(data_error_b)
data(sds_b)

# data that you will usually have on hands
data_noisy <- data_clean + data_error_b

# run mlpca_b with rank p = 2
results <- RMLPCA::mlpca_b(
  X = data_noisy,
  Xsd = sds_b,
  p = 2
)

# estimated clean dataset
data_cleaned_mlpca <- results$U %*% results$S %*% t(results$V)
```

---

mlpca_c *Maximum likelihood principal component analysis for mode C error conditions*

Description

Performs maximum likelihood principal components analysis for mode C error conditions (independent errors, general heteroscedastic case). Employs ALS algorithm.

Usage

```
mlpca_c(X, Xsd, p, MaxIter = 20000)
```

Arguments

X	MxN matrix of measurements
Xsd	MxN matrix of measurements error standard deviations
p	Rank of the model's subspace, p must be than the minimum of M and N
MaxIter	Maximum no. of iterations

Details

The returned parameters, U, S and V, are analogs to the truncated SVD solution, but have somewhat different properties since they represent the MLPCA solution. In particular, the solutions for different values of p are not necessarily nested (the rank 1 solution may not be in the space of the rank 2 solution) and the eigenvectors do not necessarily account for decreasing amounts of variance, since MLPCA is a subspace modeling technique and not a variance modeling technique.

Value

The parameters returned are the results of SVD on the estimated subspace. The quantity Ssq represents the sum of squares of weighted residuals. ErrFlag indicates the convergence condition, with 0 indicating normal termination and 1 indicating the maximum number of iterations have been exceeded.

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

Examples

```
library(RMLPCA)
data(data_clean)
data(data_error_c)
data(sds_c)

# data that you will usually have on hands
data_noisy <- data_clean + data_error_c

# run mlpca_c with rank p = 5
results <- RMLPCA::mlpca_c(
  X = data_noisy,
  Xsd = sds_c,
  p = 2
)

# estimated clean dataset
data_cleaned_mlpca <- results$U %*% results$S %*% t(results$V)
```

`mlpca_d` *Maximum likelihood principal component analysis for mode D error conditions*

Description

Performs maximum likelihood principal components analysis for mode D error conditions (commom row covariance matrices). Employs rotation and scaling of the original data.

Usage

```
mlpca_d(X, Cov, p)
```

Arguments

`X`	IxJ matrix of measurements
`Cov`	JxJ matrix of measurement error covariance, which is commom to all rows
`p`	Rank of the model's subspace

Details

The returned parameters, U, S and V, are analogs to the truncated SVD solution, but have somewhat different properties since they represent the MLPCA solution. In particular, the solutions for different values of p are not necessarily nested (the rank 1 solution may not be in the space of the rank 2 solution) and the eigenvectors do not necessarily account for decreasing amounts of variance, since MLPCA is a subspace modeling technique and not a variance modeling technique.

Value

The parameters returned are the results of SVD on the estimated subspace. The quantity Ssq represents the sum of squares of weighted residuals.

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

Examples

```
library(RMLPCA)
data(data_clean)
data(data_error_d)
# covariance matrix
data(cov_d)
data(data_cleaned_mlpca_d)
# data that you will usually have on hands
data_noisy <- data_clean + data_error_d
```

```
# run mlpca_c with rank p = 5
results <- RMLPCA::mlpca_d(
  X = data_noisy,
  Cov = cov_d,
  p = 2
)

# estimated clean dataset
data_cleaned_mlpca <- results$U %*% results$S %*% t(results$V)
```

`mlpca_e` *Maximum likelihood principal component analysis for mode E error conditions*

Description

Performs maximum likelihood principal components analysis for mode E error conditions (correlated errors, with a different covariance matrix for each row, but no error correlation between the rows). Employs an ALS algorithm.

Usage

```
mlpca_e(X, Cov, p, MaxIter = 20000)
```

Arguments

`X`	IxJ matrix of measurements
`Cov`	JXJXI matrices of measurement error covariance
`p`	Rank of the model's subspace, p must be than the minimum of I and J
`MaxIter`	Maximum no. of iterations

Details

The returned parameters, U, S and V, are analogs to the truncated SVD solution, but have somewhat different properties since they represent the MLPCA solution. In particular, the solutions for different values of p are not necessarily nested (the rank 1 solution may not be in the space of the rank 2 solution) and the eigenvectors do not necessarily account for decreasing amounts of variance, since MLPCA is a subspace modeling technique and not a variance modeling technique.

Value

The parameters returned are the results of SVD on the estimated subspace. The quantity Ssq represents the sum of squares of weighted residuals. ErrFlag indicates the convergence condition, with 0 indicating normal termination and 1 indicating the maximum number of iterations have been exceeded.

Author(s)

Renan Santos Barbosa

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

Examples

```
library(RMLPCA)
data(data_clean_e)
data(data_error_e)
# covariance matrix
data(cov_e)
data(data_cleaned_mlpca_e)
# data that you will usually have on hands
data_noisy <- data_clean_e + data_error_e

# run mlpca_e with rank p = 1
results <- RMLPCA::mlpca_e(
  X = data_noisy,
  Cov = cov_e,
  p = 1
)

# estimated clean dataset
data_cleaned_mlpca <- results$U %*% results$S %*% t(results$V)
```

RMLPCA *RMLPCA: A package for computating MLPCA algorithms b,c,d and e*

Description

The RMLPCA package provides four algorithms that to deals with measurement errors

---

sds_b *Standard deviations for mlpca_b model*

Description

A dataset where each column contain the standard deviations from 0.2 to 1 that is necessary to run mlpca_b. The main ideia is described in figure 3 on Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

Usage

```
sds_b
```

Format

A matrix with 300 rows and 20 columns

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

---

sds_c *Standard deviations for mlpca_c model*

Description

A dataset where each value come from a lognormal density with meanlog = -4.75 and sdlog = 2.5. The main ideia is described in figure 3 on Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

Usage

```
sds_c
```

Format

A matrix with 300 rows and 20 columns

References

Wentzell, P. D. "Other topics in soft-modeling: maximum likelihood-based soft-modeling methods." (2009): 507-558.

Index