Package ‘RMLPCA’

October 12, 2022

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)
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

R topics documented:

  cov_d ................................................................. 2
  cov_e ............................................................... 3
  data_clean ........................................................ 3
  data_cleaned_mlpca_b .......................................... 4
  data_cleaned_mlpca_c .......................................... 4
  data_cleaned_mlpca_d .......................................... 5
Description

A random covariance matrix to simulate data errors. The main idea 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

**cov_e**

**Description**
A random array of covariance matrices to simulate data errors The main idea 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**

data_clean

**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**
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**

---

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**
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

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
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


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 idea 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

**data_error_c**

<table>
<thead>
<tr>
<th>data_error_c</th>
<th>Errors generated for mlpca_c model</th>
</tr>
</thead>
</table>

**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 idea 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**


---

**data_error_d**

<table>
<thead>
<tr>
<th>data_error_d</th>
<th>Errors generated for mlpca_d model</th>
</tr>
</thead>
</table>

**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 idea 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**

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 idea 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


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 performing 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.
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.

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


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)
Usage

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

Arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>MxN matrix of measurements</td>
</tr>
<tr>
<td>Xsd</td>
<td>MxN matrix of measurements error standard deviations</td>
</tr>
<tr>
<td>p</td>
<td>Rank of the model’s subspace, p must be than the minimum of M and N</td>
</tr>
<tr>
<td>MaxIter</td>
<td>Maximum no. of iterations</td>
</tr>
</tbody>
</table>

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


Examples

```r
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)
```
Maximum likelihood principal component analysis for mode D error conditions

Description

Performs maximum likelihood principal components analysis for mode D error conditions (common 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 common 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


Examples

```r
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.
RMLPCA: A package for computing MLPCA algorithms b,c,d and e

Description

The RMLPCA package provides four algorithms that deal 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 idea 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**


---

**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 idea 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**

Index

* datasets
  cov_d, 2
  cov_e, 3
  data_clean, 3
  data_clean_e, 6
  data_cleaned_mlpca_b, 4
  data_cleaned_mlpca_c, 4
  data_cleaned_mlpca_d, 5
  data_cleaned_mlpca_e, 5
  data_error_b, 6
  data_error_c, 7
  data_error_d, 7
  data_error_e, 8
  sds_b, 14
  sds_c, 14

cov_d, 2
 cov_e, 3

 data_clean, 3
 data_clean_e, 6
 data_cleaned_mlpca_b, 4
 data_cleaned_mlpca_c, 4
 data_cleaned_mlpca_d, 5
 data_cleaned_mlpca_e, 5
 data_error_b, 6
 data_error_c, 7
 data_error_d, 7
 data_error_e, 8

 mlpca_b, 8
 mlpca_c, 9
 mlpca_d, 11
 mlpca_e, 12

 RMLPCA, 13

 sds_b, 14
 sds_c, 14