Package 'SpecLatent'

January 9, 2025

Type Package

Title Spectral Methods for Latent Variable Models

Version 0.0.1.0

Description
This package contains spectral methods for latent variable models, including the generlized-GoM model and degree-heterogeneous latent class model.

Encoding UTF-8

RoxygenNote 7.3.1

BugReports https://github.com/lscientific/SpecLatent/issues

Depends R (>= 2.10)

Imports RSpectra, RcppHungarian

Suggests ggplot2

LazyData true

Contents

DhLCM *DhLCM*

Description

This function performs k-means clustering on the top K eigenvectors/left singular vectors, and estimates the DhLCM model parameters

Usage

```
DhLCM(
  R,
  K,
  spectral = "heteroPCA",
  norm = "L2",
  dist = "Bern",
  T0 = 20,
  nstart = 10,
  S0 = NULL,
  clustering_only = F
)
```

Arguments

R	Numeric matrix. Data matrix.
K	Positive integer. The number of top eigenvectors/left singular vectors to be extracted.
spectral	Numeric matrix or character. One of data matrix, "heteroPCA" and "SVD". If is a matrix, it is treated as U. Otherwise needs to be a string that specifies the method to be used to obtain the top K eigenvectors/left singular vectors. "heteroPCA" implements the heteroPCA method. "SVD" performs ordinary singular vector decomposition.
norm	Character or NULL. One of "L2", "L1", "SCORE", and NULL. Specifies the method to be used for normalization on the eigenvectors/left singular vectors. "L2" performs L2 normalization. "L1" performs L1 normalization. "SCORE" performs SCORE normalization. "NA" does not perform normalization.
dist	Character. One of "Bern", "Binom", and "Pois". Specifies the data distribution. "Bern" assumes the Bernoulli distribution. "Binom" assumes the Binomial distribution. "Pois" assumes the Poisson distribution.
T0	Positive integer. The number of iterations for heteroPCA. Only used when spectral is 'heteroPCA'
nstart	Positive integer. The number of initial starts in the kmeans function.
S0	Vector or NULL. If is not NULL, used to permute the labels.
clustering_only	Boolean. When true, only clustering is conducted.

Value

Named list. The list is made of:

- `U` — Numeric matrix. Estimation of the left singular matrix.
- `T_hat` — Numeric matrix. Estimation of the Θ matrix.
- `sigma2_hat` — Numeric vector (>=0). Asymptotic variance for each element of `T_hat`.
- `S_hat` — Numeric vector. Clustered membership for each subject.
- `Z_hat` — Numeric matrix. Clustered membership for each subject in binary matrix form.

References

Lyu, Zhongyuan, Ling Chen, and Yuqi Gu. "Degree-heterogeneous Latent Class Analysis for High-dimensional Discrete Data." arXiv preprint arXiv:2402.18745 (2024).

`diag_deletion` *diag_deletion*

Description

This function takes in a matrix, and returns the diagonal-deleted matrix

Usage

```
diag_deletion(X)
```

Arguments

`X`	Numeric matrix

Value

A matrix of with diagonals set to 0

`flatten` *flatten*

Description

Flatten the polytomous matrix to a fat binary matrix

Usage

```
flatten(R)
```

Arguments

`R`	integer matrix. The polytomous response data matrix.

Value

`R_flattened` **flattened binary matrix.**

gGoM *gGoM*

Description

Estimation algorithm for generalized-GoM model with potentially locally dependent data

Usage

```
gGoM(
  R,
  K,
  pol = T,
  dist = NULL,
  large = T,
  prune = T,
  r = 10,
  q = 0.4,
  e = 0.2
)
```

Arguments

R	data matrix.
K	integer. The number of extreme latent profiles. K should be at least 2.
pol	logical; if true, assume GoM model with polytomous response, and flattening is applied. Item parameter estimation T_hat is also flattened.
dist	character; One of "Bern", "Binom", and "Pois". Specifies the data distribution. "Bern" assumes the Bernoulli distribution. "Binom" assumes the Binomial distribution. "Pois" assumes the Poisson distribution.
large	logical; if true, K needs to be at least 3 and use the large-scale SVD function RSpectra::svds.
prune	logical; if true, the pruning step is performed.
r	the number of neighbors to consider in pruning. Used only when prune is TRUE. Default value is 10.
q	the cutoff for the upper quantile of row norms. Used only when prune is TRUE. Higher q leads to more points being pruned. Default value is 0.4.
e	the cutoff for the upper quantile of average distance. Used only when prune is TRUE. Higher e leads to more points being pruned. Default value is 0.2.
lower	the minimum value for item parameters. Default value is 0.
upper	the minimum value for item parameters. Default value is 1.

Value

The function returns a list with the following components:

- P_hat the estimated membership scores.
- T_hat the estimated item response parameters.

- R_hat the estimated response expectation.
- S_hat the estimated indices of pure subjects.
- t computation time.

References

Chen, Ling, and Yuqi Gu. "A spectral method for identifiable grade of membership analysis with binary responses." Psychometrika (2024): 1-32.

Chen, Ling, Chengzhu Huang, and Yuqi Gu. "Generalized Grade-of-Membership Estimation for High-dimensional Locally Dependent Data." arXiv preprint arXiv:2412.19796 (2024).

---

gomSVD *gomSVD*

Description

Estimation algorithm for gGoM model with the left singular matrix.

Usage

```
gomSVD(U, V, d, prune = T, r = 10, q = 0.4, e = 0.2)
```

Arguments

U	the pruned left singular matrix from data SVD.
V	the right singular matrix from data SVD.
d	the vector containing the singular values.
prune	logical; if true, the pruning step is performed.
r	the number of neighbors to consider in pruning. Used only when prune is TRUE. Default value is 10.
q	the cutoff for the upper quantile of row norms. Used only when prune is TRUE. Higher q leads to more points being pruned. Default value is 0.4.
e	the cutoff for the upper quantile of average distance. Used only when prune is TRUE. Higher e leads to more points being pruned. Default value is 0.2.

Value

The function returns a list with the following components:

- P_hat the estimated membership scores.
- T_hat the estimated item response parameters (not truncated).
- R_hat the estimated response expectation (not truncated).
- S_hat the estimated indices of pure subjects.
- t computation time.

heteroPCA *heteroPCA*

Description

This function implements the HeteroPCA algorithm

Usage

```
heteroPCA(R, K, T0)
```

Arguments

R	Numeric matrix. The matrix to perform HeteroPCA.
K	Positive integer. The number of top eigenvectors to be extracted.
T0	Positive integer. The number of iterations.

Value

Numeric matrix U_hat

References

Zhang, Anru R., T. Tony Cai, and Yihong Wu. "Heteroskedastic PCA: Algorithm, optimality, and applications." The Annals of Statistics 50.1 (2022): 53-80.

perm *perm*

Description

This function performs permutation

Usage

```
perm(x, p)
```

Arguments

x	Numeric vector. Vector of labels with integer values 1, ..., K
p	Numeric vector. An integer permutation vector.

Value

Permuted vector x_perm

pruning *pruning*

Description

Locate noisy points in the data simplex.

Usage

```
pruning(mat, r = 10, q = 0.4, e = 0.2)
```

Arguments

mat	a numeric matrix to be pruned.
r	the number of neighbors to consider. Default value is 10.
q	the cutoff for the upper quantile of row norms. Higher 'q' leads to more points being pruned. Default value is 0.4.
e	the cutoff for the upper quantile of average distance. Higher 'e' leads to more points being pruned. Default value is 0.2.

Value

indices **the index vector of the rows to be pruned from the left singular matrix.**

References

Mao, X., Sarkar, P., & Chakrabarti, D. (2021). Estimating mixed memberships with sharp eigenvector deviations. Journal of the American Statistical Association, 116(536), 1928-1940.

rescale_T *rescale_T*

Description

Re-scale the item parameter estimation T_hat for polytomous GoM

Usage

```
rescale_T(T_mat, Cs)
```

Arguments

T_mat	Numeric matrix. Item parameter matrix.
Cs	Integer vector. The number of categories for each item

Value

T_mat **flattened item parameter matrix estimation.**

spa *spa*

Description

A sequential projection algorithm (SPA) to find the pure subjects

Usage

```
spa(mat)
```

Arguments

mat	the (pruned) left singular matrix to conduct SPA on.

Value

S_hat **a vector of the pure subject indices.**

References

Gillis, N. and Vavasis, S. A. (2013). Fast and robust recursive algorithms for separable nonnegative matrix factorization. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(4):698–714.

SpecLatent *Spectral Methods for Latent Variable Models*

Description

This package contains spectral methods for latent variable models, including the generalized-GoM model and degree-heterogeneous latent class model.

Details

Spectral Methods for Latent Variable Models

A package that contains spectral methods for latent variable models.

Authors

Ling Chen (<lc3521@columbia.edu>), Yuqi Gu, Zhongyuan Lyu, Chengzhu Huang, Seunghyun Lee

Data Example

The package includes a sample dataset, ‘anes‘, from the American National Election Studies (ANES) 2022 pilot study

Author(s)

Ling Chen (<lc3521@columbia.edu>), Yuqi Gu, Zhongyuan Lyu, Chengzhu Huang, Seunghyun Lee

See Also

Useful links:

- Report bugs at https://github.com/lscientific/SpecLatent/issues

Examples

```
# Load the dataset
data(anes)

# Inspect the first few rows and columns of the dataset
anes[1:6, 1:6]

# Estimation of the polytomous GoM model parameters
res <- gGoM(anes[, 2:146], K=3, pol=T, dist=NULL, large=T, prune=T, r=10, q=0.4, e=0.2)

# ternary plot
library(ggtern)
# ternary
data_tern <- as.data.frame(res$P_hat)
data_tern$party <- as.factor(anes$party)
ggtern(data=data_tern, aes(V2, V1, V3)) +
  geom_point(size=0.7, aes(color=party), alpha=1) +
  xlab("Conservative") + ylab("Indifferent") + zlab("Liberal") +
  theme(axis.title=element_text(size=8.5),
        legend.title= element_text(size=9),
        legend.text=element_text(size=9),
        legend.key.size=unit(0.6, 'cm'),
        legend.position = "bottom",
        tern.axis.title.L = element_text(hjust = 0.2, vjust = 0.5),
        tern.axis.title.R = element_text(hjust = 0.8, vjust = 0.5),
        legend.box.margin = margin(-35, 0, 0, 0),
        plot.margin = margin(-15, -15, 0, -30)) +
  scale_color_manual(values=c('#377EC2', "#7FBFBB", '#e26b57'))
```

Index