Package ‘MetabolAnalyze’
February 19, 2015

Type Package
Title Probabilistic latent variable models for metabolomic data.
Version 1.3
Date 2010-05-12
Author Nyamundanda Gift, Isobel Claire Gormley and Lorraine Brennan.
Maintainer Claire Gormley <claire.gormley@ucd.ie>
Description Fits probabilistic principal components analysis, probabilistic principal components and covariates analysis and mixtures of probabilistic principal components models to metabolomic spectral data.
Depends mclust, mvtnorm, ellipse, gtools, gplots
License GPL-2
LazyLoad yes
Repository CRAN
Date/Publication 2012-08-29 14:22:54
NeedsCompilation no

R topics documented:

MetabolAnalyze-package ........................................... 2
BrainSpectra .......................................................... 2
loadings.jack.plot .................................................... 3
loadings.plot .......................................................... 4
mppca.loadings.plot .................................................. 5
mppca.metabol ......................................................... 6
mppca.scores.plot ..................................................... 7
ppca.metabol .......................................................... 8
ppca.metabol.jack ..................................................... 10
ppca.scores.plot ..................................................... 12
ppcca.metabol ........................................................ 13
ppcca.metabol.jack .................................................. 14
ppcca.scores.plot .................................................... 16
UrineSpectra .......................................................... 17
MetabolAnalyze-package

Probabilistic latent variable models for metabolomic data.

Description

Fits probabilistic principal components analysis (PPCA), probabilistic principal components and covariates analysis (PPCCA) and mixtures of probabilistic principal component analysis (MPPCA) models to metabolomic spectral data. Estimates of the uncertainty associated with the model parameter estimates are provided.

Details

Package: MetabolAnalyze
Type: Package
Version: 1.0
Date: 2010-05-12
License: GPL-2
LazyLoad: yes

Author(s)

Nyamundanda Gift, Isobel Claire Gormley and Lorraine Brennan.
Claire Gormley <claire.gormley@ucd.ie>

References


BrainSpectra

NMR spectral data from brain tissue samples.

Description

NMR spectral data from brain tissue samples of 33 rats, where each tissue sample originates in one of four known brain regions. Each spectrum has 164 spectral bins, measured in parts per million (ppm).
loadings.jack.plot

Usage

data(BrainSpectra)

Format

A list containing

1. a matrix with 33 rows and 164 columns
2. a vector indicating the brain region of origin of each sample where:
   • 1 = Brain stem
   • 2 = Cerebellum
   • 3 = Hippocampus
   • 4 = Pre-frontal cortex

Details

This is simulated data, based on parameter estimates from a mixture of PPCA models with 4 groups and 7 principal components fitted to a similar real data set.

Source


loadings.jack.plot  Plot loadings and their associated confidence intervals.

Description

A function to plot the loadings and confidence intervals resulting from fitting a PPCA model or a PPCCA model to metabolomic data.

Usage

loadings.jack.plot(output)

Arguments

output  An object resulting from fitting a PPCA model or a PPCCA model.

Details

The function produces a plot of those loadings on the first principal component which are significantly different from zero, and higher than a user specified cutoff point. Error bars associated with the estimates, derived using the jackknife, are also plotted.
Author(s)
Nyamundanda Gift, Isobel Claire Gormley and Lorraine Brennan

References

See Also
ppca.metabol, ppcca.metabol

loadings.plot

Plot loadings.

Description
A function to plot the loadings resulting from fitting a PPCA model or a PPCCA model to metabolomic data. A barplot or a scatterplot can be produced.

Usage
loadings.plot(output, barplot = FALSE, labelsize = 0.3)

Arguments
output An object resulting from fitting a PPCA model or a PPCCA model.
barplot Logical indicating whether a barplot of the loadings is required rather than a scatter plot. By default a scatter plot is produced.
labelsize Size of the text of the spectral bin labels on the resulting plot.

Details
A function to plot the loadings resulting from fitting a PPCA model or a PPCCA model to metabolomic data. A barplot or a scatterplot can be produced. The size of the text of the spectral bin labels on the bar plot can also be adjusted if the number of bins plotted is large.

Author(s)
Nyamundanda Gift, Isobel Claire Gormley and Lorraine Brennan

References

See Also
ppca.metabol, ppcca.metabol
Plot loadings resulting from fitting a MPPCA model.

Description

A function to plot the loadings resulting from fitting a MPPCA model to metabolomic data. A barplot or a scatterplot can be produced.

Usage

mppca.loadings.plot(output, Y, barplot = FALSE, labelsize = 0.3)

Arguments

- **output**: An object resulting from fitting a MPPCA model.
- **Y**: The N x p matrix of observations to which the MPPCA model is fitted.
- **barplot**: Logical indicating whether a barplot of the loadings is required rather than a scatter plot. By default a scatter plot is produced.
- **labelsize**: Size of the text of the spectral bin labels on the resulting plot.

Details

A function which produces a series of plots illustrating the loadings resulting from fitting a MPPCA model to metabolomic data. A barplot or a scatterplot can be produced. The size of the text of the spectral bin labels on the bar plot can also be adjusted if the number of bins plotted is large.

Author(s)

Nyamundanda Gift, Isobel Claire Gormley and Lorraine Brennan

References


See Also

mppca.metabol
Fit a mixture of probabilistic principal components analysis (MPPCA) model to a metabolomic data set via the EM algorithm to perform simultaneous dimension reduction and clustering.

Description

This function fits a mixture of probabilistic principal components analysis model to metabolomic spectral data via the EM algorithm.

Usage

mppca.metabol(Y, minq=1, maxq=2, ming, maxg, scale = "none", epsilon = 0.1, plot.BIC = FALSE)

Arguments

Y
An N x p data matrix where each row is a spectrum.

minq
The minimum number of principal components to be fit. By default minq is 1.

maxq
The maximum number of principal components to be fit. By default maxq is 2.

ming
The minimum number of groups to be fit.

maxg
The maximum number of groups to be fit.

scale
Type of scaling of the data which is required. The default is "none". Options include "pareto" and "unit" scaling. See scaling for further details.

epsilon
Value on which the convergence assessment criterion is based. Set by default to 0.1.

plot.BIC
Logical indicating whether or not a plot of the BIC values for the different models fitted should be provided. By default, the plot is not produced.

Details

This function fits a mixture of probabilistic principal components analysis models to metabolomic spectral data via the EM algorithm. A range of models with different numbers of groups and different numbers of principal components can be fitted. The model performs simultaneous clustering of observations into unknown groups and dimension reduction simultaneously.

Value

A list containing:

q
The number of principal components in the optimal MPPCA model, selected by the BIC.

g
The number of groups in the optimal MPPCA model, selected by the BIC.

sig
The posterior mode estimate of the variance of the error terms.
scores A list of length g, each entry of which is a n_g x q matrix of estimates of the latent locations of each observation in group g in the principal subspace.

loadings An array of dimension p x q x g, each sheet of which contains the maximum likelihood estimate of the p x q loadings matrix for a group.

Pi The vector indicating the probability of belonging to each group.

mean A p x g matrix, each column of which contains a group mean.

tau An N x g matrix, each row of which contains the posterior group membership probabilities for an observation.

clustering A vector of length N indicating the group to which each observation belongs.

BIC A matrix containing the BIC values for the fitted models.

AIC A matrix containing the AIC values for the fitted models.

Author(s)
Nyamundanda Gift, Isobel Claire Gormley and Lorraine Brennan.

References

See Also
mppca.scores.plot, mppca.loadings.plot

Examples

data(BrainSpectra)
## Not run:
mdlfit<-mppca.metabol(BrainSpectra[[1]], minq=7, maxq=7, ming=4, maxg=4, plot.BIC = TRUE)
mppca.scores.plot(mdlfit)
mppca.loadings.plot(mdlfit, BrainSpectra[[1]])

## End(Not run)
Arguments

output An object resulting from fitting a MPPCA model.
group Should it be relevant, a vector indicating the known treatment group membership of each observation prior to clustering.
gplegend Logical indicating whether a legend should be plotted.

Details

This function produces a series of scatterplots, for each group uncovered. For group g, each scatterplot illustrates the estimated score for each observation allocated to that group within the reduced q dimensional space. The uncertainty associated with the score estimate is also illustrated through its 95

It is often the case that observations are known to belong to treatment groups, for example, and the MPPCA model is employed to uncover any underlying subgroups, possibly related to disease subtypes. The treatment group membership of each observation can be illustrated on the plots produced by utilizing the ‘group’ argument.

Author(s)

Nyamundanda Gift, Isobel Claire Gormley and Lorraine Brennan

References


See Also

mppca.metabol

ppca.metabol
Fit a probabilistic principal components analysis (PPCA) model to a metabolomic data set via the EM algorithm.

Description

This function fits a probabilistic principal components analysis model to metabolomic spectral data via the EM algorithm.

Usage

ppca.metabol(Y, minq=1, maxq=2, scale = "none", epsilon = 0.1, plot.BIC = FALSE, printout=TRUE)
Arguments

y
  An N x p data matrix where each row is a spectrum.
minq
  The minimum number of principal components to be fit. By default minq is 1.
maxq
  The maximum number of principal components to be fit. By default maxq is 2.
scale
  Type of scaling of the data which is required. The default is "none". Options include "pareto" and "unit" scaling. See scaling for further details.
epsilon
  Value on which the convergence assessment criterion is based. Set by default to 0.1.
plot.BIC
  Logical indicating whether or not a plot of the BIC values for the different models fitted should be provided. By default, the plot is not produced.
printout
  Logical indicating whether or not a statement is printed on screen detailing the progress of the algorithm.

Details

This function fits a probabilistic principal components analysis model to metabolomic spectral data via the EM algorithm. A range of models with different numbers of principal components can be fitted.

Value

A list containing:

q
  The number of principal components in the optimal PPCA model, selected by the BIC.
sig
  The posterior mode estimate of the variance of the error terms.
scores
  An N x q matrix of estimates of the latent locations of each observation in the principal subspace.
loadings
  The maximum likelihood estimate of the p x q loadings matrix.
BIC
  A vector containing the BIC values for the fitted models.
AIC
  A vector containing the AIC values for the fitted models.

Author(s)

Nyamundanda Gift, Isobel Claire Gormley and Lorraine Brennan.

References


See Also

ppca.metabol.jack, loadings.plot, pPCA.scores.plot
Fit a probabilistic principal components analysis model to a metabolomic data set, and assess uncertainty via the jackknife.

**Description**

Fit a probabilistic principal components analysis (PPCA) model to a metabolomic data set via the EM algorithm, and assess uncertainty in the obtained loadings estimates via the jackknife.

**Usage**

```r
call = pPCA.metabol.jack(Y, minq = 1, maxq = 2, scale = "none", epsilon = 0.1, conflevel = 0.95)
```

**Arguments**

- `Y`: An N x p data matrix where each row is a spectrum.
- `minq`: The minimum number of principal components to be fit. By default minq is 1.
- `maxq`: The maximum number of principal components to be fit. By default maxq is 2.
- `scale`: Type of scaling of the data which is required. The default is "none". Options include 'pareto' and "unit" scaling. See `scaling` for further details.
- `epsilon`: Value on which the convergence assessment criterion is based. Set by default to 0.1.
- `conflevel`: Level of confidence required for the loadings confidence intervals. By default 95\% confidence intervals are computed.

**Details**

A (range of) PPCA model(s) are fitted and an optimal model (i.e. number of principal components, q) is selected. Confidence intervals for the obtained loadings are then obtained via the jackknife i.e. a model with q principal components is fitted to the dataset N times, where an observation is removed from the dataset each time.

On convergence of the algorithm, the number of loadings significantly different from zero is printed on screen. The user may then further examine the significant loadings when prompted by selecting a cutoff value from the table printed on screen. Bar plots detailing the resulting significantly high loadings are provided.
Value

A list containing:

- **q**
  - The number of principal components in the optimal PPCA model, selected by the BIC.
- **sig**
  - The posterior mode estimate of the variance of the error terms.
- **scores**
  - An N x q matrix of estimates of the latent locations of each observation in the principal subspace.
- **loadings**
  - The maximum likelihood estimate of the p x q loadings matrix.
- **SignifW**
  - The maximum likelihood estimate of the loadings matrix for those loadings significantly different from zero.
- **SignifHighW**
  - The maximum likelihood estimate of the loadings matrix for those loadings significantly different from zero and higher than a user selected cutoff point.
- **Lower**
  - The lower limit of the confidence interval for those loadings significantly different from zero.
- **Upper**
  - The upper limit of the confidence interval for those loadings significantly different from zero.
- **Cutoffs**
  - A table detailing a range of cutoff points and the associated number of selected spectral bins.
- **number**
  - The number of spectral bins selected by the user.
- **cutoff**
  - The cutoff value selected by the user.
- **BIC**
  - A vector containing the BIC values for the fitted models.
- **AIC**
  - A vector containing the AIC values for the fitted models.

Author(s)

Nyamundanda Gift, Isobel Claire Gormley and Lorraine Brennan.

References


See Also

- `ppca.metabol`, `loadings.jack.plot`, `ppca.scores.plot`

Examples

data(UrineSpectra)

```r
# Not run:
md1fit<-ppca.metabol.jack(UrineSpectra[[1]], minq=2, maxq=2, scale="none")
loadings.jack.plot(md1fit)
ppca.scores.plot(md1fit, group=UrineSpectra[[2]][[1]])
```

## End(Not run)
ppca.scores.plot  
Plot scores from a fitted PPCA model

Description
A function to plot the scores resulting from fitting a PPCA model to metabolomic data.

Usage
ppca.scores.plot(output, group = FALSE)

Arguments
- output: An object resulting from fitting a PPCA model.
- group: Should it be relevant, a vector indicating the known treatment group membership of each observation.

Details
This function produces a series of scatterplots each illustrating the estimated score for each observation within the reduced q dimensional space. The uncertainty associated with the score estimate is also illustrated through its 95%

It is often the case that observations are known to belong to treatment groups; the treatment group membership of each observation can be illustrated on the plots produced by utilizing the ‘group’ argument.

Author(s)
Nyamundanda Gift, Isobel Claire Gormley and Lorraine Brennan

References

See Also
ppca.metabol, ppcas.metabol.jack
ppcca.metabol

Fit a probabilistic principal components and covariates analysis (PPCCA) model to a metabolomic data set via the EM algorithm.

Description
This function fits a probabilistic principal components and covariates analysis model to metabolomic spectral data via the EM algorithm.

Usage
ppcca.metabol(Y, Covars, minq=1, maxq=2, scale = "none", epsilon = 0.1, plot.BIC = FALSE, printout=TRUE)

Arguments
Y  An N x p data matrix in which each row is a spectrum.
Covars  An N x L covariate data matrix in which each row is a set of covariates.
minq  The minimum number of principal components to be fit.
maxq  The maximum number of principal components to be fit.
scale  Type of scaling of the data which is required. The default is "none". Options include "pareto" and "unit" scaling. See scaling for further details.
epsilon  Value on which the convergence assessment criterion is based. Set by default to 0.1.
plot.BIC  Logical indicating whether or not a plot of the BIC values for the different models fitted should be provided. By default, the plot is not produced.
printout  Logical indicating whether or not a statement is printed on screen detailing the progress of the algorithm.

Details
This function fits a probabilistic principal components and covariates analysis model to metabolomic spectral data via the EM algorithm. A range of models with different numbers of principal components can be fitted.

Care should be taken with the form of covariates supplied. All covariates are standardized (to lie in [0,1]) within the ppcca.metabol function for stability reasons. Hence continuous covariates and binary valued categorical covariates are easily handled. For a categorical covariate with V levels, the equivalent V-1 dummy variables representation should be passed as an argument to ppcca.metabol.

Value
A list containing:
q  The number of principal components in the optimal PPCCA model, selected by the BIC.
sig  The posterior mode estimate of the variance of the error terms.
scores  An N x q matrix of estimates of the latent locations of each observation in the principal subspace.
loadings  The maximum likelihood estimate of the p x q loadings matrix.
coefficients  The maximum likelihood estimates of the regression coefficients associated with the covariates in the PPCCA model.
BIC  A vector containing the BIC values for the fitted models.
AIC  A vector containing the AIC values for the fitted models.

Author(s)
Nyamundanda Gift, Isobel Claire Gormley and Lorraine Brennan.

References

See Also
ppcca.metabol.jack, ppcca.scores.plotloadings.plot

Examples

data(UrineSpectra)
### Not run:
mdlfit<-ppcca.metabol(UrineSpectra[[1]], UrineSpectra[[2]][,2], minq=2, maxq=2)
loadings.plot(mdlfit)
ppcca.scores.plot(mdlfit, UrineSpectra[[2]][,2], group=UrineSpectra[[2]][,1], covarnames="Weight")

### End(Not run)

---

**ppcca.metabol.jack**  *Fit a probabilistic principal components and covariates analysis model to a metabolomic data set, and assess uncertainty via the jackknife.*

Description

Fit a probabilistic principal components and covariates analysis (PPCCA) model to a metabolomic data set via the EM algorithm, and assess uncertainty in the obtained loadings estimates and the regression coefficients via the jackknife.

Usage

ppcca.metabol.jack(Y, Covars, minq=1, maxq=2, scale="none", epsilon=0.1, conflevel=0.95)
Arguments

\textbf{\texttt{Y}} \hspace{1cm} An N x p data matrix in which each row is a spectrum.

\textbf{\texttt{Covars}} \hspace{1cm} An N x L covariate data matrix where each row is a set of covariates.

\texttt{minq} \hspace{1cm} The minimum number of principal components to be fit. By default minq is 1.

\texttt{maxq} \hspace{1cm} The maximum number of principal components to be fit. By default maxq is 2.

\texttt{scale} \hspace{1cm} Type of scaling of the data which is required. The default is "none". Options include "pareto' and "unit" scaling. See \texttt{scaling} for further details.

\texttt{epsilon} \hspace{1cm} Value on which the convergence assessment criterion is based. Set by default to 0.1.

\texttt{conflevel} \hspace{1cm} Level of confidence required for the loadings and regression coefficients confidence intervals. By default 95\% confidence intervals are computed.

Details

A (range of) PPCCA model(s) are fitted and an optimal model (i.e. number of principal components, q) is selected. Confidence intervals for the obtained loadings and regression coefficients are then obtained via the jackknife i.e. a model with q principal components is fitted to the data N times, where an observation is removed from the dataset each time.

Care should be taken with the form of covariates supplied. All covariates are standardized (to lie in [0,1]) within the ppcca.metabol.jack function for stability reasons. Hence continuous covariates and binary valued categorical covariates are easily handled. For a categorical covariate with V levels, the equivalent V-1 dummy variables representation should be passed as an argument to ppcca.metabol.jack.

On convergence of the algorithm, the number of loadings significantly different from zero is printed on screen. The user may then further examine the significant loadings when prompted by selecting a cutoff value from the table printed on screen. Bar plots detailing the resulting significantly high loadings are provided.

Value

A list containing:

\texttt{q} \hspace{1cm} The number of principal components in the optimal PPCCA model, selected by the BIC.

\texttt{sig} \hspace{1cm} The posterior mode estimate of the variance of the error terms.

\texttt{scores} \hspace{1cm} An N x q matrix of estimates of the latent locations of each observation in the principal subspace.

\texttt{loadings} \hspace{1cm} The maximum likelihood estimate of the p x q loadings matrix.

\texttt{SignifW} \hspace{1cm} The maximum likelihood estimate of the loadings matrix for those loadings significantly different from zero.

\texttt{SignifHighW} \hspace{1cm} The maximum likelihood estimate of the loadings matrix for those loadings significantly different from zero and above the user selected cutoff point.

\texttt{LowerCI_W} \hspace{1cm} The lower limit of the confidence interval for those loadings significantly different from zero.
UpperCI_W  The upper limit of the confidence interval for those loadings significantly different from zero.

coefficients  The maximum likelihood estimates of the regression coefficients.

coeffCI  A matrix detailing the upper and lower limits of the confidence intervals for the regression parameters.

Cutoffs  A table detailing a range of cutoff points and the associated number of selected spectral bins.

number  The number of spectral bins selected by the user.

cutoff  The cutoff value selected by the user.

BIC  A vector containing the BIC values for the fitted models.

AIC  A vector containing the AIC values for the fitted models.

Author(s)
Nyamundanda Gift, Isobel Claire Gormley and Lorraine Brennan.

References

See Also
ppcca.metabol, ppcca.scores.plot, loadings.jack.plot

Examples
data(UrineSpectra)
## Not run:
mdlfit<-ppcca.metabol.jack(UrineSpectra[[1]], UrineSpectra[[2]][,2], minq=2, maxq=2)
loadings.jack.plot(mdlfit)
ppcca.scores.plot(mdlfit, UrineSpectra[[2]][,2], group=UrineSpectra[[2]][,1], covarnames="Weight")
## End(Not run)

ppcca.scores.plot  Plot scores from a fitted PPCCA model.

Description
A function to plot the scores resulting from fitting a PPCCA model to metabolomic data.

Usage
ppcca.scores.plot(output, Covars, group = FALSE, covarnames=NULL)
Arguments

- **output**: An object resulting from fitting a PPCCA model.
- **Covars**: An N x L covariate data matrix where each row is a set of covariates.
- **group**: Should it be relevant, a vector indicating the known treatment group membership of each observation.
- **covarnames**: Should it be relevant, a vector string indicating the names of the covariates.

Details

This function produces a series of scatterplots each illustrating the estimated score for each observation within the reduced q dimensional space. The uncertainty associated with the score estimate is also illustrated through its 95

It is often the case that observations are known to belong to treatment groups; the treatment group membership of each observation can be illustrated on the plots produced by utilizing the ‘group’ argument.

Author(s)

Nyamundanda Gift, Isobel Claire Gormley and Lorraine Brennan

References


See Also

- `ppcca.metabol`, `ppcca.metabol.jack`

---

**UrineSpectra**

NMR metabolomic spectra from urine samples of 18 mice.

Description

NMR metabolomic spectra from urine samples of 18 mice, each belonging to one of two treatment groups. Each spectrum has 189 spectral bins, measured in parts per million (ppm).

Covariates associated with the mice were also recorded: the weight of each mouse is provided.

Usage

`data(UrineSpectra)`
Format

A list containing

1. a matrix with 18 rows and 189 columns
2. a data frame with 18 observations on 2 variables:
   • Treatment group membership of each animal.
   • Weight (in grammes) of each animal.

Details

This is simulated data, based on parameter estimates from a PPCA model with two principal components fitted to a similar real data set.

Source

Index

*Topic **datasets**

BrainSpectra, 2
UrineSpectra, 17

*Topic **methods**

loadings.jack.plot, 3
loadings.plot, 4
mppca.loadings.plot, 5
mppca.metabol, 6
mppca.scores.plot, 7
ppca.metabol, 8
ppca.metabol.jack, 10
ppca.scores.plot, 12
ppcca.metabol, 13
ppcca.metabol.jack, 14
ppcca.scores.plot, 16

BrainSpectra, 2

loadings.jack.plot, 3, 11, 16
loadings.plot, 4, 9, 14

MetabolAnalyse

(MetabolAnalyse-package), 2
MetabolAnalyse-package, 2
mppca.loadings.plot, 5, 7
mppca.metabol, 5, 6, 8
mppca.scores.plot, 7, 7

ppca.metabol, 4, 8, 11, 12
ppca.metabol.jack, 4, 9, 10, 12
ppca.scores.plot, 9, 11, 12
ppca.metabol, 4, 13, 16, 17
ppcca.metabol.jack, 4, 14, 14, 17
ppcca.scores.plot, 14, 16, 16

scaling, 6, 9, 10, 13, 15

UrineSpectra, 17