Package 'BNPTSclust'

April 19, 2015

Type Package

Title A Bayesian Nonparametric Algorithm for Time Series Clustering

Version 1.1

Date 2015-04-19

Author Martell-Juarez, D.A. & Nieto-Barajas, L.E.

Maintainer David Alejandro Martell Juarez <alex91599@gmail.com>

Depends R(>= 3.1.2), mvtnorm, MASS

Description Performs the algorithm for time series clustering described in Nieto-Barajas and Contreras-Cristan (2014).

NeedsCompilation no

Repository CRAN

License GPL (>= 2)

Date/Publication 2015-04-19 22:19:50

R topics documented:

BNPTSclust-package *A Bayesian Nonparametric Algorithm for Time Series Clustering*

Description

This package performs the algorithm for time series clustering described in Nieto-Barajas and Contreras-Cristan (2014). The package contains functions to work with annual, monthly and quarterly time series data.

The main functions to accomplish the above are:

1) tseriesca

2) tseriescm

3) tseriescq

Details

Package:	BNPTSclust
Type:	Package
Version:	1.1
Date:	2015-04-19
License:	GPL2, GPL3

For a comprehensive guide on how to use the package, refer to the vignette attached to the package.

Author(s)

Martell-Juarez, D.A. and Nieto-Barajas, L.E.

Maintainer: David Alejandro Martell Juarez <alex91599@gmail.com>

References

Nieto-Barajas, L.E. and Contreras-Cristan, A. (2014) A Bayesian Nonparametric Approach for Time Series Clustering. *Bayesian Analysis* **Vol. 9, No. 1** 147–170.

clusterplots *Cluster groups plotting function.*

Description

Function that plots the time series clusters generated by either of the functions: "tseriesca", "tseriescm" or "tseriescq".

Usage

```
clusterplots(L, data)
```

Arguments

L	output list from the functions: "tseriesca", "tseriescm" or "tseriescq".
data	Data frame with the time series information.

Details

See the examples in the documentation files of "tseriesca", "tseriescm" or "tseriescq" for an example of this function's usage.

Value

The function returns the plots of the time series clusters directly.

Author(s)

Martell-Juarez, D.A.

---

comp11 *Univariate ties function*

Description

Computes the distinct observations and frequencies in a numeric vector.

Usage

```
comp11(y)
```

Arguments

y	Numeric vector.

Details

The code of the function is the same as the "comp1" function from the "BNPdensity" package. The change is in the output of the function. This function is for internal use.

Value

jstar	variable that rearranges "y" into a vector with its unique values.
nstar	frequency of each distinct observation in "y".
rstar	number of distinct observations in "y".
gn	variable that indicates the group number to which every entry in "y" belongs.

Note

For internal use.

Author(s)

Martell-Juarez, D.A., Barrios, E., Nieto-Barajas, L. and Pruenster, I.

`designmatrices` *Function that creates the design matrices necessary for the clustering algorithm to work.*

Description

Function that generates the design matrices of the clustering algorithm based on the parameters that the user wants to consider, i.e. level, polinomial trend and/or seasonal components. It also returns the number of parameters that are considered and not considered for clustering.

Usage

```
designmatrices(level, trend, seasonality, deg, T, n, fun)
```

Arguments

`level`	Variable that indicates if the level of the time series will be considered for clustering. If level = 0, then it is omitted. If level = 1, then it is taken into account.
`trend`	Variable that indicates if the polinomial trend of the model will be considered for clustering. If trend = 0, then it is omitted. If trend = 1, then it is taken into account.
`seasonality`	Variable that indicates if the seasonal components of the model will be considered for clustering. If seasonality = 0, then they are omitted. If seasonality = 1, then they are taken into account.
`deg`	Degree of the polinomial trend of the model.
`T`	Number of periods of the time series.
`n`	Number of time series.
`fun`	Clustering function being used.

Value

`Z`	Design matrix of the parameters not considered for clustering.
`X`	Design matrix of the parameters considered for clustering.
`p`	Number of parameters not considered for clustering.
`d`	Number of parameters considered for clustering.

Note

For internal use.

Author(s)

Martell-Juarez, D.A.

`diagplots` *Diagnostic plots function.*

Description

Function that produces the diagnostic plots to assess the convergence of the Markov Chains generated by either of the functions: "tseriesca", "tseriescm" or "tseriescq".

Usage

```
diagplots(L)
```

Arguments

`L`	output list from the functions: "tseriesca", "tseriescm" or "tseriescq".

Details

See the examples in the documentation files of "tseriesca", "tseriescm" or "tseriescq" for an example of this function's usage.

Value

The function returns three different kinds of plots to assess convergence of the generated Markov Chain: trace plots, histograms and ergodic mean plots.

Author(s)

Martell-Juarez, D.A.

gdp *GDP per person employed from 1990 to 2012*

Description

This data set contains the yearly GDP per person employed from 1990 to 2012 for 121 countries.

Usage

```
data(gdp)
```

Format

Data frame with 20 rows and 121 columns.

Source

http://data.worldbank.org/indicator/SL.GDP.PCAP.EM.KD

houses *House price statistics in Scotland from 2004 to 2014.*

Description

This data set contains the average price of houses from the 1st quarter of 2004 to the 4th quarter of 2014 by the local authority areas of Scotland

Usage

```
data(houses)
```

Format

Data frame with 44 rows and 33 columns.

Source

http://www.ros.gov.uk/public/news/quarterly_statistics.html

References

http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

`scaleandperiods` *Scaling data function.*

Description

This function scales the time series data in the interval [0,1] as deemed necessary in Nieto-Barajas and Contreras-Cristan (2014) for the time series clustering algorithm. It also obtains the time periods of the data set provided.

Usage

```
scaleandperiods(data)
```

Arguments

`data`	Data frame with the time series information.

Details

The function considers that the time periods of the data appear as row names.

Value

`periods`	array with the time periods of the data.
`mydata`	data frame with the time series data scaled in [0,1].
`cts`	variable that indicates if some time series were removed because they were constant in time. If no time series were removed, cts = 0. If there were time series removed, cts indicates the column of such time series.

Note

For internal use.

Author(s)

Martell-Juarez, D.A.

stocks *Mexican stock exchange market prices*

Description

This data set contains the monthly adjusted closing prices of 58 shares of the mexican stock exchange market from September 2006 to August 2011.

Usage

```
data(stocks)
```

Format

Data frame with 60 rows and 58 columns.

Source

http://www.dowjones.com/factiva/

References

This is the data set used by Nieto-Barajas, L.E. & Contreras-Cristan, A. (2014) as application for their paper.

tseriesca *Function for annual time series clustering.*

Description

Function that performs the time series clustering algorithm described in Nieto-Barajas and Contreras-Cristan (2014) for annual time series data.

Usage

```
tseriesca(data, maxiter = 1000, burnin = floor(0.1 * maxiter),
          thinning = 5, level = FALSE, trend = TRUE, deg = 2, c0eps = 2,
          c1eps = 1, c0beta = 2, c1beta = 1, c0alpha = 2, c1alpha = 1,
          priora = FALSE, pia = 0.5, q0a = 1, q1a = 1, priorb = FALSE,
          q0b = 1, q1b = 1, a = 0.25, b = 0, indlpml = FALSE)
```

Arguments

data	Data frame with the time series information.
maxiter	Maximum number of iterations for Gibbs sampling.
burnin	Burn-in period of the Markov Chain generated by Gibbs sampling.
thinning	Number that indicates how many Gibbs sampling simulations should be skipped to form the Markov Chain.
level	Flag that indicates if the level of the time series will be considered for clustering. If TRUE, then it is taken into account.
trend	Flag that indicates if the polinomial trend of the model will be considered for clustering. If TRUE, then it is taken into account.
deg	Degree of the polinomial trend of the model.
c0eps	Shape parameter of the hyper-prior distribution on sig2eps.
c1eps	Rate parameter of the hyper-prior distribution on sig2eps.
c0beta	Shape parameter of the hyper-prior distribution on sig2beta.
c1beta	Rate parameter of the hyper-prior distribution on sig2beta.
c0alpha	Shape parameter of the hyper-prior distribution on sig2alpha.
c1alpha	Rate parameter of the hyper-prior distribution on sig2alpha.
priora	Flag that indicates if a prior on parameter "a" is to be assigned. If TRUE, a prior on "a" is assigned.
pia	Mixing proportion of the prior distribution on parameter "a".
q0a	Shape parameter of the continuous part of the prior distribution on parameter "a".
q1a	Shape parameter of the continuous part of the prior distribution on parameter "a".
priorb	Flag that indicates if a prior on parameter "b" is to be assigned. If TRUE, a prior on "b" is assigned.
q0b	Shape parameter of the prior distribution on parameter "b".
q1b	Shape parameter of the prior distribution on parameter "b".
a	Initial/fixed value of parameter "a".
b	Initial/fixed value of parameter "b".
indlpml	Flag that indicates if the LPML is to be calculated. If TRUE, LPML is calculated.

Details

It is assumed that the time series data is organized into a data frame with the time periods included as its row names.

Value

mstar	Number of groups of the chosen cluster configuration.
gnstar	Array that contains the group number to which each time series belongs.
HM	Heterogeneity Measure of the chosen cluster configuration.
arrho	Acceptance rate of the parameter "rho".
ara	Acceptance rate of the parameter "a".
arb	Acceptance rate of the parameter "b".
sig2epssample	Matrix that in its columns contains the sample of each sig2eps_i's posterior distribution after Gibbs sampling.
sig2alphasample	Matrix that in its columns contains the sample of each sig2alpha_i's posterior distribution after Gibbs sampling.
sig2betasample	Matrix that in its columns contains the sample of each sig2beta_i's posterior distribution after Gibbs sampling.
sig2thesample	Vector that contains the sample of sig2the's posterior distribution after Gibbs sampling.
rhosample	Vector that contains the sample of rho's posterior distribution after Gibbs sampling.
asample	Vector that contains the sample of a's posterior distribution after Gibbs sampling.
bsample	Vector that contains the sample of b's posterior distribution after Gibbs sampling.
msample	Vector that contains the sample of the number of groups at each Gibbs sampling iteration.
lpml	If indlpml = TRUE, lpml contains the value of the LPML of the chosen model.

Author(s)

Martell-Juarez, D.A. and Nieto-Barajas, L.E.

Examples

```
## Do not run
#
# data(gdp)
# tseriesca.out <- tseriesca(gdp,maxiter = 4000,level=FALSE,trend=TRUE,
#                            c0eps = 0.001,c1eps = 0.001,c0beta = 0.001,
#                            c1beta = 0.001,c0alpha = 0.001,
#                            c1alpha= 0.001,priorb = TRUE,a = 0,b = 0.1)
#
# The console output of the above example is:
#
# Number of groups of the chosen cluster configuration : 13
# Time series in group 1 : 1 111
# Time series in group 2 : 2 8
```

```
# Time series in group 3 : 3 4 5 6 7 10 11 12 13 14 15 16 17 18 19 20 21
# 22 24 25 26 28 29 30 31 32 33 34 35 36 37 38 40 41 42 43 44 45 46 47 49
# 50 51 52 55 56 57 58 59 61 62 63 65 67 68 69 70 71 74 75 76 77 78 79 80
# 81 82 83 84 85 86 89 92 92 93 94 95 96 97 100 101 102 103 104 105 106
# 107 108 109 110 113 114 117 118 120
# Time series in group 4 : 9 23 48 54 60 87
# Time series in group 5 : 27
# Time series in group 6 : 39
# Time series in group 7 : 53 73 88
# Time series in group 8 : 64
# Time series in group 9 : 66 98 112
# Time series in group 10 : 72
# Time series in group 11 : 90 116 119 121
# Time series in group 12 : 99
# Time series in group 13 : 115
# HM Measure : 99.50627
#
# Make sure that chain convergence is always assessed. Run the following
# code to show the cluster and diagnostic plots:

data(gdp)
data(tseriesca.out)
attach(tseriesca.out)

clusterplots(tseriesca.out,gdp)
diagplots(tseriesca.out)
```

`tseriesca.out` *Output of tseriesca function for the GDP per person employed dataset*

Description

This object contains the output of the function tseriesca for the example described in its documentation file.

Usage

```
data(tseriesca.out)
```

Details

See function tseriesca for an explanation of how the output was obtained.

Examples

```
data(tseriesca.out)
```

`tseriescm` *Function for monthly time series clustering.*

Description

Function that performs the time series clustering algorithm described in Nieto-Barajas and Contreras-Cristan (2014) for monthly time series data.

Usage

```
tseriescm(data, maxiter = 1000, burnin = floor(0.1 * maxiter),
          thinning = 5, level = FALSE, trend = TRUE, seasonality = TRUE,
          deg = 2,c0eps = 2, c1eps = 1, c0beta = 2, c1beta = 1,
          c0alpha = 2,c1alpha = 1, priora = FALSE, pia = 0.5, q0a = 1,
          q1a = 1, priorb = FALSE, q0b = 1, q1b = 1, a = 0.25, b = 0,
          indlpml = FALSE)
```

Arguments

`data`	Data frame with the time series information.
`maxiter`	Maximum number of iterations for Gibbs sampling.
`burnin`	Burn-in period of the Markov Chain generated by Gibbs sampling.
`thinning`	Number that indicates how many Gibbs sampling simulations should be skipped to form the Markov Chain.
`level`	Flag that indicates if the level of the time series will be considered for clustering. If TRUE, then it is taken into account.
`trend`	Flag that indicates if the polinomial trend of the model will be considered for clustering. If TRUE, then it is taken into account.
`seasonality`	Flag that indicates if the seasonal components of the model will be considered for clustering. If TRUE, then they are taken into account.
`deg`	Degree of the polinomial trend of the model.
`c0eps`	Shape parameter of the hyper-prior distribution on sig2eps.
`c1eps`	Rate parameter of the hyper-prior distribution on sig2eps.
`c0beta`	Shape parameter of the hyper-prior distribution on sig2beta.
`c1beta`	Rate parameter of the hyper-prior distribution on sig2beta.
`c0alpha`	Shape parameter of the hyper-prior distribution on sig2alpha.
`c1alpha`	Rate parameter of the hyper-prior distribution on sig2alpha.
`priora`	Flag that indicates if a prior on parameter "a" is to be assigned. If TRUE, a prior on "a" is assigned.
`pia`	Mixing proportion of the prior distribution on parameter "a".
`q0a`	Shape parameter of the continuous part of the prior distribution on parameter "a".

q1a	Shape parameter of the continuous part of the prior distribution on parameter "a".
priorb	Flag that indicates if a prior on parameter "b" is to be assigned. If TRUE, a prior on "b" is assigned.
q0b	Shape parameter of the prior distribution on parameter "b".
q1b	Shape parameter of the prior distribution on parameter "b".
a	Initial/fixed value of parameter "a".
b	Initial/fixed value of parameter "b".
indlpml	Flag that indicates if the LPML is to be calculated. If TRUE, LPML is calculated.

Details

It is assumed that the time series data is organized into a data frame with the time periods included as its row names.

Value

mstar	Number of groups of the chosen cluster configuration.
gnstar	Array that contains the group number to which each time series belongs.
HM	Heterogeneity Measure of the chosen cluster configuration.
arrho	Acceptance rate of the parameter "rho".
ara	Acceptance rate of the parameter "a".
arb	Acceptance rate of the parameter "b".
sig2epssample	Matrix that in its columns contains the sample of each sig2eps_i's posterior distribution after Gibbs sampling.
sig2alphasample	Matrix that in its columns contains the sample of each sig2alpha_i's posterior distribution after Gibbs sampling.
sig2betasample	Matrix that in its columns contains the sample of each sig2beta_i's posterior distribution after Gibbs sampling.
sig2thesample	Vector that contains the sample of sig2the's posterior distribution after Gibbs sampling.
rhosample	Vector that contains the sample of rho's posterior distribution after Gibbs sampling.
asample	Vector that contains the sample of a's posterior distribution after Gibbs sampling.
bsample	Vector that contains the sample of b's posterior distribution after Gibbs sampling.
msample	Vector that contains the sample of the number of groups at each Gibbs sampling iteration.
lpml	If indlpml = TRUE, lpml contains the value of the LPML of the chosen model.

Author(s)

Martell-Juarez, D.A. and Nieto-Barajas, L.E.

Examples

```
## Do not run
#
# data(stocks)
# tseriescm.out <- tseriescm(stocks,maxiter=4000,level=FALSE,trend=TRUE,
#                            seasonality=TRUE,a=0,b=1)
#
# The console output of the above example is:
#
# Number of groups of the chosen cluster configuration:  9
# Time series in group 1 : 1 2 4 5 7 10 12 13 19 21 22 25 29 30 31 33 34
# 40 41 42 43 44 46 47 48 49 52 57 58
# Time series in group 2 : 3 6 8 9 11 14 15 17 18 26 27 28 32 35 36 37 38
# 45 50 51 53 56
# Time series in group 3 : 16
# Time series in group 4 : 20
# Time series in group 5 : 23
# Time series in group 6 : 24
# Time series in group 7 : 39
# Time series in group 8 : 54
# Time series in group 9 : 55
# HM Measure:  199.2226
#
# Make sure that chain convergence is always assessed. Run the following
# code to show the cluster and diagnostic plots:

data(stocks)
data(tseriescm.out)
attach(tseriescm.out)

clusterplots(tseriescm.out,stocks)
diagplots(tseriescm.out)
```

---

tseriescm.out *Output of tseriescm function for the Mexican stock exchange market prices dataset*

Description

This object contains the output of the function tseriescm for the example described in its documentation file.

Usage

```
data(tseriescm.out)
```

Details

See function tseriescm for an explanation of how the output was obtained.

Examples

```
data(tseriescm.out)
```

`tseriescq` *Function for quarterly time series clustering.*

Description

Function that performs the time series clustering algorithm described in Nieto-Barajas and Contreras-Cristan (2014) for quarterly time series data.

Usage

```
tseriescq(data, maxiter = 1000, burnin = floor(0.1 * maxiter),
          thinning = 5, level = FALSE, trend = TRUE, seasonality = TRUE,
          deg = 2,c0eps = 2, c1eps = 1, c0beta = 2, c1beta = 1,
          c0alpha = 2,c1alpha = 1, priora = FALSE, pia = 0.5, q0a = 1,
          q1a = 1,priorb = FALSE, q0b = 1, q1b = 1, a = 0.25, b = 0,
          indlpml = FALSE)
```

Arguments

`data`	Data frame with the time series information.
`maxiter`	Maximum number of iterations for Gibbs sampling.
`burnin`	Burn-in period of the Markov Chain generated by Gibbs sampling.
`thinning`	Number that indicates how many Gibbs sampling simulations should be skipped to form the Markov Chain.
`level`	Flag that indicates if the level of the time series will be considered for clustering. If TRUE, then it is taken into account.
`trend`	Flag that indicates if the polinomial trend of the model will be considered for clustering. If TRUE, then it is taken into account.
`seasonality`	Flag that indicates if the seasonal components of the model will be considered for clustering. If TRUE, then they are taken into account.
`deg`	Degree of the polinomial trend of the model.
`c0eps`	Shape parameter of the hyper-prior distribution on sig2eps.
`c1eps`	Rate parameter of the hyper-prior distribution on sig2eps.
`c0beta`	Shape parameter of the hyper-prior distribution on sig2beta.
`c1beta`	Rate parameter of the hyper-prior distribution on sig2beta.
`c0alpha`	Shape parameter of the hyper-prior distribution on sig2alpha.

c1alpha	Rate parameter of the hyper-prior distribution on sig2alpha.
priora	Flag that indicates if a prior on parameter "a" is to be assigned. If TRUE, a prior on "a" is assigned.
pia	Mixing proportion of the prior distribution on parameter "a".
q0a	Shape parameter of the continuous part of the prior distribution on parameter "a".
q1a	Shape parameter of the continuous part of the prior distribution on parameter "a".
priorb	Flag that indicates if a prior on parameter "b" is to be assigned. If TRUE, a prior on "b" is assigned.
q0b	Shape parameter of the prior distribution on parameter "b".
q1b	Shape parameter of the prior distribution on parameter "b".
a	Initial/fixed value of parameter "a".
b	Initial/fixed value of parameter "b".
indlpml	Flag that indicates if the LPML is to be calculated. If TRUE, LPML is calculated.

Details

It is assumed that the time series data is organized into a data frame with the time periods included as its row names.

Value

mstar	Number of groups of the chosen cluster configuration.
gnstar	Array that contains the group number to which each time series belongs.
HM	Heterogeneity Measure of the chosen cluster configuration.
arrho	Acceptance rate of the parameter "rho".
ara	Acceptance rate of the parameter "a".
arb	Acceptance rate of the parameter "b".
sig2epssample	Matrix that in its columns contains the sample of each sig2eps_i's posterior distribution after Gibbs sampling.
sig2alphasample	Matrix that in its columns contains the sample of each sig2alpha_i's posterior distribution after Gibbs sampling.
sig2betasample	Matrix that in its columns contains the sample of each sig2beta_i's posterior distribution after Gibbs sampling.
sig2thesample	Vector that contains the sample of sig2the's posterior distribution after Gibbs sampling.
rhosample	Vector that contains the sample of rho's posterior distribution after Gibbs sampling.
asample	Vector that contains the sample of a's posterior distribution after Gibbs sampling.

bsample	Vector that contains the sample of b's posterior distribution after Gibbs sampling.
msample	Vector that contains the sample of the number of groups at each Gibbs sampling iteration.
lpml	If indlpml = TRUE, lpml contains the value of the LPML of the chosen model.

Author(s)

Martell-Juarez, D.A. and Nieto-Barajas, L.E.

Examples

```
## Do not run
#
# data(houses)
# tseriescq.out <- tseriescq(houses,maxiter=4000,level=FALSE,trend=TRUE,
#                            seasonality=TRUE,priora=TRUE)
#
# The console output of the above example is:
#
# Number of groups of the chosen cluster configuration : 9
# Time series in group 1 : 1
# Time series in group 2 : 2 3 4 5 6 7 9 10 11 12 13 15 16 17 18 19 20 21 # 25 26 27 29 30 31 33
# Time series in group 3 : 8 23
# Time series in group 4 : 14
# Time series in group 5 : 22
# Time series in group 6 : 24
# Time series in group 7 : 28
# Time series in group 8 : 32
# Time series in group 9 : 34
# HM Measure : 126.9543
#
# Make sure that chain convergence is always assessed. Run the following
# code to show the cluster and diagnostic plots:

data(houses)
data(tseriescq.out)
attach(tseriescq.out)

clusterplots(tseriescq.out,houses)
diagplots(tseriescq.out)
```

---

tseriescq.out *Output of tseriescq function for the House price statistics in Scotland dataset*

Description

This object contains the output of the function tseriescq for the example described in its documentation file.

Usage

```
data(tseriescq.out)
```

Details

See function tseriescq for an explanation of how the output was obtained.

Examples

```
data(tseriescq.out)
```

Index