Package ‘BNPTSclust’

August 19, 2019

Type Package
Title A Bayesian Nonparametric Algorithm for Time Series Clustering
Version 2.0
Date 2019-08-19
Author Martell-Juarez, D.A. & Nieto-Barajas, L.E.
Maintainer David Alejandro Martell Juarez <alex91599@gmail.com>
Depends R(>= 3.6.0), 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 2019-08-19 21:40:03 UTC

R topics documented:

BNPTSclust-package .................................................. 2
clusterplots ............................................................. 2
compII ................................................................. 3
designmatrices ......................................................... 4
diagplots ............................................................... 5
gdp ................................................................. 6
houses ................................................................. 6
scaleandperiods ......................................................... 7
stocks ................................................................. 8
tseriesca ............................................................ 8
tseriesca.out ......................................................... 11
tseriescm ............................................................ 11
tseriescm.out ......................................................... 14
tseriescq ............................................................ 14
tseriescq.out ........................................................ 17

Index 18
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: 2.0
Date: 2019-08-19
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


Description

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

clusterplots

Cluster groups plotting function.
**Usage**

```r
counterplots(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**

```r
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.
designmatrices

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

**Note**

For internal use.

**Author(s)**

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

**Description**

Function that generates the design matrices of the clustering algorithm based on the parameters that the user wants to consider, i.e. level, polynomial 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 polynomial 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 polynomial 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.
diagplots

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**


**References**

**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**

```r
scaleandperiods(data, scale)
```

**Arguments**

- **data**: Data frame with the time series information.
- **scale**: Flag that indicates if the time series data should be scaled to the [0,1] interval with a linear transformation as proposed by Nieto-Barajas and Contreras-Cristan (2014). If TRUE, then the time series are scaled to the [0,1] interval. Its value comes directly from the “scale” argument of the clustering functions.

**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**

```r
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**

```r
tseriesca(data, maxiter = 500, burnin = floor(0.1 * maxiter),
    thinning = 5, scale = TRUE, level = FALSE, trend = TRUE, deg = 2,
    c0eps = 2, c1eps = 1, c0beta = 2, c1beta = 1, c0alpha = 2, c1alpha = 1,
    priora = TRUE, pia = 0.5, q0a = 1, q1a = 1, priorb = TRUE, 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.
scale: Flag that indicates if the time series data should be scaled to the [0,1] interval with a linear transformation as proposed by Nieto-Barajas and Contreras-Cristan (2014). If TRUE, then the time series are scaled to the [0,1] interval.
level: Flag that indicates if the level of the time series will be considered for clustering. If TRUE, then it is considered.
trend: Flag that indicates if the polynomial trend of the model will be considered for clustering. If TRUE, then it is taken into account.
deg: Degree of the polynomial 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".
piorb: 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".
indlpm1: 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

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>mstar</td>
<td>Number of groups of the chosen cluster configuration.</td>
</tr>
<tr>
<td>gnstar</td>
<td>Array that contains the group number to which each time series belongs.</td>
</tr>
<tr>
<td>HM</td>
<td>Heterogeneity Measure of the chosen cluster configuration.</td>
</tr>
<tr>
<td>arrho</td>
<td>Acceptance rate of the parameter &quot;rho&quot;.</td>
</tr>
<tr>
<td>ara</td>
<td>Acceptance rate of the parameter &quot;a&quot;.</td>
</tr>
<tr>
<td>arb</td>
<td>Acceptance rate of the parameter &quot;b&quot;.</td>
</tr>
<tr>
<td>sig2epssample</td>
<td>Matrix that in its columns contains the sample of each sig2eps_i's posterior distribution after Gibbs sampling.</td>
</tr>
<tr>
<td>sig2alphasample</td>
<td>Matrix that in its columns contains the sample of each sig2alpha_i's posterior distribution after Gibbs sampling.</td>
</tr>
<tr>
<td>sig2betasample</td>
<td>Matrix that in its columns contains the sample of each sig2beta_i's posterior distribution after Gibbs sampling.</td>
</tr>
<tr>
<td>sig2thesample</td>
<td>Vector that contains the sample of sig2the's posterior distribution after Gibbs sampling.</td>
</tr>
<tr>
<td>rhosample</td>
<td>Vector that contains the sample of rho's posterior distribution after Gibbs sampling.</td>
</tr>
<tr>
<td>asample</td>
<td>Vector that contains the sample of a's posterior distribution after Gibbs sampling.</td>
</tr>
<tr>
<td>bsample</td>
<td>Vector that contains the sample of b's posterior distribution after Gibbs sampling.</td>
</tr>
<tr>
<td>msample</td>
<td>Vector that contains the sample of the number of groups at each Gibbs sampling iteration.</td>
</tr>
<tr>
<td>lpml</td>
<td>If indlpml = TRUE, lpml contains the value of the LPML of the chosen model.</td>
</tr>
<tr>
<td>scale</td>
<td>Flag that indicates if the time series data were scaled to the [0,1] interval with a linear transformation. This will be taken as an input for the plotting functions.</td>
</tr>
</tbody>
</table>

Author(s)

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

Examples

```r
## Do not run
#
# data(gdp)
# tseriesca.out <- tseriesca(gdp,maxiter = 4000,level=FALSE,trend=TRUE,
# c0eps = 0.1,c1eps = 0.1,c0beta = 0.1,
# c1beta = 0.1,c0alpha = 0.1,
# c1alpha= 0.1)
# 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 = 500, burnin = floor(0.1 * maxiter),
  thinning = 5, scale = TRUE, level = FALSE, trend = TRUE, seasonality = TRUE,
  deg = 2, c0eps = 2, c1eps = 1, c0beta = 2, c1beta = 1, c0alpha = 2,
  c1alpha = 1, priora = TRUE, pia = 0.5, q0a = 1, q1a = 1, priorb = TRUE,
  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.
scale  Flag that indicates if the time series data should be scaled to the [0,1] interval with a linear transformation as proposed by Nieto-Barajas and Contreras-Cristan (2014). If TRUE, then the time series are scaled to the [0,1] interval.
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.
- **scale**: Flag that indicates if the time series data were scaled to the [0,1] interval with a linear transformation. This will be taken as an input for the plotting functions.

### Author(s)

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

### Examples

```r
## Do not run
#
# data(stocks)
# tseriescm.out <- tseriescm(stocks,maxiter=4000,level=FALSE,trend=TRUE,
# seasonality=TRUE,priorb=FALSE,b=0)
#
# 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  
\textit{Output of tseriescm function for the Mexican stock exchange market prices dataset}

\section*{Description}

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

\section*{Usage}

\begin{verbatim}
data(tseriescm.out)
\end{verbatim}

\section*{Details}

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

\section*{Examples}

\begin{verbatim}
data(tseriescm.out)
\end{verbatim}

tseriescq  
\textit{Function for quarterly time series clustering.}

\section*{Description}

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

\section*{Usage}

\begin{verbatim}
tseriescq(data, maxiter = 500, burnin = floor(0.1 * maxiter),
            thinning = 5, scale = TRUE, level = FALSE, trend = TRUE, seasonality = TRUE,
            deg = 2, c0eps = 2, c1eps = 1, c0beta = 2, c1beta = 1, c0alpha = 2,
            c1alpha = 1, priora = TRUE, pia = 0.5, q0a = 1, q1a = 1, priorb = TRUE,
            q0b = 1, q1b = 1, a = 0.25, b = 0, indlpml = FALSE)
\end{verbatim}
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.
scale | Flag that indicates if the time series data should be scaled to the \([0,1]\) interval with a linear transformation as proposed by Nieto-Barajas and Contreras-Cristan (2014). If TRUE, then the time series are scaled to the \([0,1]\) interval.
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 polynomial 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 polynomial trend of the model.
c0eps | Shape parameter of the hyper-prior distribution on \(\text{sig2eps}\).
c1eps | Rate parameter of the hyper-prior distribution on \(\text{sig2eps}\).
c0beta | Shape parameter of the hyper-prior distribution on \(\text{sig2beta}\).
c1beta | Rate parameter of the hyper-prior distribution on \(\text{sig2beta}\).
c0alpha | Shape parameter of the hyper-prior distribution on \(\text{sig2alpha}\).
c1alpha | Rate parameter of the hyper-prior distribution on \(\text{sig2alpha}\).
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".
piorb | 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.

**scale**
Flag that indicates if the time series data were scaled to the [0,1] interval with a linear transformation. This will be taken as an input for the plotting functions.

Author(s)

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

Examples

```r
## Do not run
#
# data(houses)
# tseriescq.out <- tseriescq(houses,maxiter=4000,level=FALSE,trend=TRUE,
#   seasonality=TRUE,priora=TRUE)
#
# 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

*Topic datasets
  gdp, 6
  houses, 6
  stocks, 8
  tseriesca.out, 11
  tseriescm.out, 14
  tseriescq.out, 17

*Topic package
  BNPTSclust-package, 2

BNPTSclust (BNPTSclust-package), 2
BNPTSclust-package, 2

clusterplots, 2
comp11, 3

designmatrices, 4
diagplots, 5

gdp, 6
houses, 6
scaleandperiods, 7
stocks, 8

tseriesca, 8
tseriesca.out, 11
tseriescm, 11
tseriescm.out, 14
tseriescq, 14
tseriescq.out, 17