Package 'dpmixsim'

July 11, 2018

Title Dirichlet Process Mixture Model Simulation for Clustering and Image Segmentation

Version 0.0-9

Date 2018-07-11

Author Adelino Ferreira da Silva <afs@fct.unl.pt>

Maintainer Adelino Ferreira da Silva <afs@fct.unl.pt>

Depends R (>= 2.10.0), oro.nifti, cluster

Description The 'dpmixsim' package implements a Dirichlet Process Mixture (DPM) model for clustering and image segmentation. The DPM model is a Bayesian nonparametric methodology that relies on MCMC simulations for exploring mixture models with an unknown number of components. The code implements conjugate models with normal structure (conjugate normal-normal DP mixture model). The package's applications are oriented towards the classification of magnetic resonance images according to tissue type or region of interest.

License GPL (>= 2)

Repository CRAN

Date/Publication 2018-07-11 17:00:02 UTC

NeedsCompilation yes

R topics documented:

`dpmixsim` *Dirichlet Process mixture model for clustering and image segmentation*

Description

`dpmixsim` implements a Dirichlet Process mixture (DPM) model. The DPM model is a Bayesian nonparametric methodology that relies on MCMC simulations for exploring mixture models with an unknown number of components. The function implements conjugate models with normal structure (conjugate normal-normal DP mixture model).

Usage

```
dpmixsim(x, M=1, a=1, b=1, upalpha=1, a0=2, b0=2, maxiter=4000, rec=3000,
  fsave=NA, kmax=30, nclinit=NA, minvar=0.001)
```

Arguments

`x`	scaled input data as vector in range {0,1}
`M`	DP precision hyperparameter
`a`	Gamma prior hyperparameter
`b`	Gamma prior hyperparameter
`upalpha`	is a logical variable for simulations with {automatic,fixed} calibration of the precision hyperparameter M (default = 'TRUE')
`a0`	Gamma prior hyperparameter for M (default 2)
`b0`	Gamma prior hyperparameter for M (default 2)
`maxiter`	maximum number of MCMC iteration steps
`rec`	record the last 'rec' iteration steps
`fsave`	filename for saving the MCMC simulation (def: 'NULL' do not save)
`kmax`	maximum number of clusters in the simulation, (default 30)
`nclinit`	number of initial clusters to use at the beginning of the simulation. If not specified (NA) the number of initial clusters is equal to the length of x (one element per cluster); (default: NA)
`minvar`	minimum value admissible for a cluster variance (default=0.001). Decreasing 'minval' may improve resolution (distribution fitness), but increases the maximum number of admissible clusters ('kmax'). In this case, you may have to increase ('kmax') as well.

Details

Consider n observations x_1,...,x_n which we regard as exchangeable. We model the distribution from which the x_i are drawn as a mixture of distributions. Dirichlet process mixture models are based on Dirichlet process priors for the primary parameters θ_i. DP mixture models assume that the prior distribution function G itself is uncertain, drawn from a Dirichlet process $G \sim DP(MG_0)$, with base prior G_0 and precision parameter M. This specification may be expressed by the hierarchical model:

$x_i \sim N(.|\theta_i, \sigma^2)$
$\theta_i \sim \mathrm{G}$
$G \sim DP(MN(0,1))$
$\sigma^{-2} \sim Gamma(a,b)$

Value

simulation output as a list of draws containing:

`krec`	cluster indicator variables
`wrec`	cluster weights
`phirec`	theta cluster parameters
`varrec`	sigma cluster parameters

Author(s)

Adelino Ferreira da Silva, Universidade Nova de Lisboa, Faculdade de Ciencias e Tecnologia, Portugal, <afs@fct.unl.pt>.

References

Adelino Ferreira da Silva, A Dirichlet process mixture model for brain MRI tissue classification, *Medical Image Analysis* 11 (2007) 169-182.

Adelino Ferreira da Silva, Bayesian mixture models of variable dimension for image segmentation, *Comput. Methods Programs Biomed.* 94 (2009) 1-14.

See Also

readsliceimg, postdataseg, postdpmixciz, postimgclgrp, postimgcomps, postkcluster, premask, readsliceimg

Examples

```
## Not run:
## Example 1: simple test using `galaxy' data
  data("galaxy")
  x0 <- galaxy$speed
  x  <- prescale(x0)
  maxiter <- 4000; rec <- 3000; ngrid <- 100
  res <- dpmixsim(x, M=1, a=1, b=0.1, upalpha=1, maxiter=maxiter, rec=rec,
    nclinit=4)
  z <-  postdpmixciz(x=x, res=res, rec=rec, ngrid=ngrid, plot=T)
```

```
  ##
  res <- dpmixsim(x, M=2, a=1, b=0.001, upalpha=0, maxiter=maxiter,
    rec=rec, nclinit=4)
  z <-  postdpmixciz(x, res=res, rec=rec, ngrid=ngrid, plot=T)
##-----------------
## Example 2:
  demo(testMarronWand)
##-----------------
## Example 3: MRI segmentation
## Testing note: this example should reproduce the equivalent segmented
## images used in the author's references
  slicedata <- readsliceimg(fbase="t1_pn3_rf0", swap=FALSE)
  image(slicedata$niislice, col=gray((0:255)/256), main="original image")
  x0 <- premask(slicedata, subsamp=TRUE)
  x  <- prescale(x0)
  rec <- 3000
  res <- dpmixsim(x, M=1, a=1, b=1, upalpha=1, maxiter=4000,
      rec=rec, nclinit=8, minvar=0.002)
  ## post-simulation
  ngrid <- 200
  z <- postdpmixciz(x, res=res, rec=rec, ngrid=ngrid, plot=TRUE)
  x0 <- premask(slicedata, subsamp=FALSE) # use full-sized image after estimation
  x  <- prescale(x0)
  cx   <- postdataseg(x, z, ngrid=ngrid)
  cat("*** view grouped segmentations:\n")
  postimgclgrp(slicedata$mask, cx, palcolor=FALSE)
  cat("*** display all clusters:\n")
  postimgcomps(slicedata$mask, cx)
  cat("*** re-cluster with 4 clusters:\n")
  postkcluster(slicedata$mask, cx, clk=4)

## End(Not run)
```

`galaxy` *Galaxy velocities*

Description

This data set considers physical information on velocities (km/second) for 82 galaxies reported by Roeder (1990). These are drawn from six well-separated conic sections of the Corona Borealis region.

Usage

```
data(galaxy)
```

Format

A data frame with 82 observations on the following variable.

`speed` a numeric vector giving the speed of galaxies ((km/second))

Source

Roeder, K. (1990) Density estimation with confidence sets exemplified by superclusters and voids in the galaxies, Journal of the American Statistical Association, 85: 617-624.

References

Escobar, M.D. and West, M. (1995) Bayesian Density Estimation and Inference Using Mixtures. Journal of the American Statistical Association, 90: 577-588.

Examples

```
data(galaxy)
## maybe str(galaxy) ; plot(galaxy) ...
```

---

postdataseg *Data segmentation*

Description

postdataseg performs data segmentation based on labelled cluster estimates.

Usage

```
postdataseg(x, z, ngrid, dbg=FALSE)
```

Arguments

x	full-sized scaled image data prepared by premask
z	cluster labels produced by postdpmixciz
ngrid	dimension of the grid used in estimation
dbg	logical variable to show debugging output (default = 'FALSE')

Details

Once the distributions of the indicator variables z_i are calculated we can separate the components of the mixture. Individual components are selected according to the most probable z_i value in a given region of the distributional space, leading to a partition of this space into regions. Intensity threshold values are associated with the partition of the distributional space to drive the image segmentation. In brief, the partition of the distributional space induced by the z values is used to segment the data space. From a computational point of view, the use of these two separate spaces enables us to optimize the MCMC implementation.

Value

cx	vector of image cluster values

Author(s)

A. Ferreira da Silva, Universidade Nova de Lisboa, Faculdade de Ciencias e Tecnologia, <afs@fct.unl.pt>.

See Also

dpmixsim, readsliceimg, premask, postdpmixciz

Examples

```
## Not run:
## see Example 2 in dpmixsim.

## End(Not run)
```

---

postdpmixciz *Summary statistics and cluster estimation*

Description

postdpmixciz computes post-simulation summary statistics, and estimates cluster partition.

Usage

```
postdpmixciz(x, res, kmax=30, rec=300, ngrid=200, plot=TRUE)
```

Arguments

x	data used in the simulation
kmax	maximum number of clusters
res	output of the MCMC simulation
rec	number of recorded iteration steps
ngrid	dimension of the grid used in density estimation
plot	logical variable to omit plots (default = 'TRUE'

Value

z	cluster partition estimation

Author(s)

A. Ferreira da Silva, Universidade Nova de Lisboa, Faculdade de Ciencias e Tecnologia, <afs@fct.unl.pt>.

References

Adelino Ferreira da Silva, A Dirichlet process mixture model for brain MRI tissue classification, *Medical Image Analysis* 11 (2007) 169-182.

Adelino Ferreira da Silva, Bayesian mixture models of variable dimension for image segmentation, *Comput. Methods Programs Biomed.* 94 (2009) 1-14.

See Also

dpmixsim

Examples

```
## Not run:
## Example: MRI brain image segmentation
  slicedata <- readsliceimg(fbase="t1_pn3_rf0", swap=FALSE)
  image(slicedata$niislice, col=gray((0:255)/256), main="original image")
  x0 <- premask(slicedata, subsamp=TRUE)
  x  <- prescale(x0)
  rec <- 3000
  res <- dpmixsim(x, M=1, a=1, b=2, upalpha=1, maxiter=4000,
      rec=rec, nclinit=8)
  ## post-simulation
  ngrid <- 200
  z <- postdpmixciz(x, res=res, rec=rec, ngrid=ngrid, plot=TRUE)

## End(Not run)
```

---

postimgclgrp *Segment image with the estimated number of components*

Description

postimgclgrp displays the segmented image with the estimated number of components

Usage

```
postimgclgrp(mask, cx, palcolor=TRUE)
```

Arguments

mask	full-sized scaled image data prepared by premask
cx	data segmentation prepared by postdataseg
palcolor	logical variable for selecting colored/grey image visualization (default = 'TRUE')

Details

Display image segmentation with the estimated number of components.

Author(s)

A. Ferreira da Silva, Universidade Nova de Lisboa, Faculdade de Ciencias e Tecnologia, <afs@fct.unl.pt>.

References

Adelino Ferreira da Silva, A Dirichlet process mixture model for brain MRI tissue classification, *Medical Image Analysis* 11 (2007) 169-182.

Adelino Ferreira da Silva, Bayesian mixture models of variable dimension for image segmentation, *Comput. Methods Programs Biomed.* 94 (2009) 1-14.

See Also

dpmixsim, readsliceimg, premask, postdpmixciz, postdataseg

Examples

```
## Not run:
## see Examples in `dpmixsim'.

## End(Not run)
```

`postimgcomps` *Display cluster components*

Description

`postimgcomps` displays the components of the segmented image with the estimated number of components

Usage

```
postimgcomps(mask, cx)
```

Arguments

`mask`	scaled masked full-sized image data prepared by `premask`
`cx`	data segmentation prepared by `postdataseg`

Details

Display components based on the estimated number of clusters.

Author(s)

A. Ferreira da Silva, Universidade Nova de Lisboa, Faculdade de Ciencias e Tecnologia, <afs@fct.unl.pt>.

References

Adelino Ferreira da Silva, A Dirichlet process mixture model for brain MRI tissue classification, *Medical Image Analysis* 11 (2007) 169-182.

Adelino Ferreira da Silva, Bayesian mixture models of variable dimension for image segmentation, *Comput. Methods Programs Biomed.* 94 (2009) 1-14.

See Also

dpmixsim, readsliceimg, premask, postdpmixciz, postdataseg, postimgclgrp

Examples

```
## Not run:
## see Examples in `dpmixsim'.

## End(Not run)
```

`postkcluster` *Segmentation with a fixed number of clusters*

Description

`postkcluster` re-clusters the data with a user-specified number of components, and displays the segmented image.

Usage

```
postkcluster(mask, cx, clk=4, plot=TRUE)
```

Arguments

`mask`	masked full-sized image data prepared by `premask`
`cx`	data segmentation prepared by `postdataseg`
`clk`	desired fixed number of components, including the background component, to use in the data segmentation; default '`clk=4`': gray matter (GM), white matter (WM), CSF, and background
`plot`	logical variable; enables suspension of output images (default = '`TRUE`')

Details

Partitioning clustering around medoids (PAM) is applied to the classes simulated from `dpmixsim` as a post-processing step. This procedure may be applied to merge clusters, and reduce the number of clusters to the specified value '`clk`'. `postkcluster` computes a `clara` object using **cluster** (see Struyf et.al.), a list representing a clustering of the data into '`clk`' clusters.

Author(s)

Adelino Ferreira da Silva, Universidade Nova de Lisboa, Faculdade de Ciencias e Tecnologia, Portugal, <afs@fct.unl.pt>.

References

Adelino Ferreira da Silva, A Dirichlet process mixture model for brain MRI tissue classification, *Medical Image Analysis* 11 (2007) 169-182.

Adelino Ferreira da Silva, Bayesian mixture models of variable dimension for image segmentation, *Comput. Methods Programs Biomed.* 94 (2009) 1-14.

Anja Struyf, Mia Hubert & Peter J. Rousseeuw (1996): Clustering in an Object-Oriented Environment. Journal of Statistical Software, 1. http://www.stat.ucla.edu/journals/jss/

See Also

dpmixsim, readsliceimg, premask, postdpmixciz, postdataseg, postimgcomps

Examples

```
## Not run:
## see Examples in `dpmixsim'.

## End(Not run)
```

`premask` *Data preparation*

Description

`premask` applies a pre-defined mask to a MRI slice in order to select regions of interest (ROIs) for processing

Usage

```
premask(slicedata, subsamp=TRUE)
```

Arguments

`slicedata`	list as output by `read.sliceimg`
`subsamp`	logical variable; if 'TRUE' a downsampled image by a factor of 2 is used in the MCMC simulation, otherwise the full-sized image is taken. After parameter estimation, the full-sized image should be used for clustering and image segmentation. The use of downsampled images can substantially reduce runtime, with little quality degradation.

Value

xv	processed data vector

Author(s)

A. Ferreira da Silva, Universidade Nova de Lisboa, Faculdade de Ciencias e Tecnologia, <afs@fct.unl.pt>.

See Also

dpmixsim, readsliceimg

Examples

```
## Not run:
  slicedata <- readsliceimg(fbase="t1_pn3_rf0", swap=FALSE)
  x0 <- premask(slicedata, subsamp=TRUE)
  x  <- prescale(x0)
  print(str(x))

## End(Not run)
```

---

prescale *Data preparation*

Description

prescale scales data to be in the range {0,1}, as a preparation for simulation.

Usage

```
prescale(xv)
```

Arguments

xv	unscaled data vector

Value

x	scaled data vector

Author(s)

A. Ferreira da Silva, Universidade Nova de Lisboa, Faculdade de Ciencias e Tecnologia, <afs@fct.unl.pt>.

See Also

dpmixsim, readsliceimg

Examples

```
## Not run:
  data("galaxy")
  x0 <- galaxy$speed
  x  <- prescale(x0)
  print(range(x))

## End(Not run)
```

`readsliceimg` *Read MRI slice data*

Description

`readsliceimg` reads MRI and mask data.

Usage

```
readsliceimg(fbase="t1_pn3_rf0", swap=FALSE)
```

Arguments

`fbase`	Indicates the dataset prefix of the MRI dataset to use. The prefix applies to data files: '`{fbase}_slice_0092.nii.gz`', and '`{fbase}_slice_0092_mask.nii.gz`'. These data files were obtained from the BrainWeb repository of the McConnell Brain Imaging Center at the Montreal Neurological Institute. BrainWeb anatomical models uses MRI slices of dimension 181x217 pixels. The datasets included in the package for demonstration correspond to a T1 BrainWeb image for slice number 92, with 3% noise and 0% intensity non-uniformity.
`swap`	logical variable (default = '`FALSE`') for choosing the right/left data display convention consistent with FSLVIEW

Details

The FSL tools may be used to prepare the MRI data and the mask required as data input. The package **oro.nifti** is used for reading gzipped NIFTI files.

Value

a list containing

`fbase`	dataset prefix of the dataset used in the analysis
`niislice`	slice data at all timepoints
`mask`	slice mask
`nrow`	number of rows
`ncol`	number of columns
`swap`	relative orientation used in the data setup

Author(s)

A. Ferreira da Silva, Universidade Nova de Lisboa, Faculdade de Ciencias e Tecnologia, <afs@fct.unl.pt>.

References

Brandon Whitcher, Volker Schmid and Andrew Thornton, Package **oro.nifti**: Rigorous - NIfTI Input / Output, 2010.

FSL/FEAT Analysis tool, FMRIB Software Library (FSL) (www.fmrib.ox.ac.uk/fsl)

See Also

dpmixsim

Examples

```
## Not run:
slicedata <- readsliceimg(fbase="t1_pn3_rf0", swap=FALSE)
print(str(slicedata))

## End(Not run)
```

---

t1_pn3_rf0_slice_0092.Rd

Example of a pre-processed MRI slice from the BrainWeb database

---

Description

The file 't1_pn3_rf0_slice_0092.nii.gz' is a pre-processed image of slice '92' with '3%' noise extracted from the Brainweb database file 't1_icbm_normal_1mm_pn3_rf0\[1\].mnc.gz'. BrainWeb simulations are based on an anatomical model of normal brain, which can serve as the ground truth for any analysis procedure. BrainWeb datasets and are provided by the McConnell Brain Imaging Center at the Montreal Neurological Institute, http://www.bic.mni.mcgill.ca/, (see Collins et. al. 1998).

Format

The file 't1_pn3_rf0_slice_0092.ni.gz' is in gzipped NIFTI format. The R-package **oro.nifti** is required to read gzipped NIFTI files.

References

D.L. Collins, et.al., Design and construction of a realistic digital brain phantom, *IEEE Trans. on Medical Imaging* 17~(3) (1998) 463-468.

S.M. Smith, et. al., Advances in Functional and Structural MR Image Analysis and Implementation as FSL, *NeuroImage*, 23(S1):208-219, 2004.

Brandon Whitcher, Volker Schmid and Andrew Thornton, Package **oro.nifti**: Rigorous - NIfTI Input / Output, 2010.

t1_pn3_rf0_slice_0092_mask.Rd

Mask file for MRI slice

Description

The 't1_pn3_rf0_slice_0092_mask.nii.gz' defines the mask for 't1_pn3_rf0_slice_0092.nii.gz', as used in the examples. The mask used here is an all-brain mask; it just removes non-brain regions, as the result of applying a brain extraction tool to the specified dataset. Other masks may be defined to select regions of interest (ROIs).

Format

The file 't1_pn3_rf0_slice_0092_mask.nii.gz' is in gzipped NIFTI format. The R-package **oro.nifti** is required to read gzipped NIFTI files.

References

D.~L. Collins, A.~P. Zijdenbos, V.~Kollokian, J.~G. Sled, N.~J. Kabani, C.~J. Holmes, A.~C. Evans, Design and construction of a realistic digital brain phantom, *IEEE Trans. on Medical Imaging* 17~(3) (1998) 463–468.

S.M. Smith, et. al., Advances in Functional and Structural MR Image Analysis and Implementation as FSL, *NeuroImage*, 23(S1):208-219, 2004.

Brandon Whitcher, Volker Schmid and Andrew Thornton, Package **oro.nifti**: Rigorous - NIfTI Input / Output, 2010.

Index