Dirichlet Process Mixture Models

April 20, 2007

R topics documented:

`mdp-package` *Dirichlet process mixture (MDP) of normal model.*

Description

The package implements posterior inference for MDP models with multivariate normal kernel and conjugate DP base measure. The mixture is with respect to both, mean and covariance matrix of the kernel. The package includes posterior predictive inference for one- and two-dimensional subvectors, and posterior predictive mean functions and surfaces for two and three-dimensional subvectors.

Details

Package:	mdp
Type:	Package
Version:	1.0
Date:	2007-04-01
License:	GNU public domain software.

The function `mdp` implements posterior MCMC simulation. The function assumes a conditionally conjugate MDP model, with conjugate kernel and Dirichlet process base measure. The functions `plt.pxy`, `plt.exy` and `plt.ex` plot posterior predictive inference.

The package uses the model described in *MacEachern and Mueller (1998)*. Let $y_i \sim H(y)$ denote an i.i.d. sample from an unknown distribution $H(y)$. We model the random probability measure H as a location and scale mixture of multivariate normal kernels,

$$H(y) = \int N(\mu, V) dG(\mu, V)$$

with a Dirichlet process (DP) prior for the mixing measure, $G \sim DP(G_0, \alpha)$ and a conditionally

conjugate base measure $G_0(\mu, V^{-1}) = G_{\mu 0}(\mu) G_{V0}(V^{-1})$ where

$$G_{\mu 0}(\mu) = N(m, B) \text{ and } G_{V0}(V^{-1}) = Wishart(s, (sS)^{-1}).$$

The Wishart distribution is parametrized such that $E(V_{-1}) = S^{-1}$. We assume conjugate hyperpriors

$$m \sim N(a, A), B^{-1} \sim Wishart(c, (cC)^{-1}), S \sim Wishart(q, R/q).$$

The model is completed with a Gamma prior for the total mass paramter $\alpha \sim Ga(a_0, b_0)$.

Author(s)

Peter Mueller

Maintainer: Peter Mueller <pm@wotan.mdacc.tmc.edu>

References

The package uses the parametrization defined in:

MacEachern, S.N. and Mueller, P. (1998). "Estimating Mixture of Dirichlet Process Models," Journal of Computational and Graphical Statistics, 7, 223–239.

See Also

See also the DPpackage, at http://student.kuleuven.be/~s0166452/software.html.

`mdp` *MDP – Dirichlet process mixture of normals*

Description

Fits a semiparametric Dirichlet process mixture of normals.

Usage

```
mdp(n = NULL, p = NULL,
    n.iter = 10000, n.discard = 1000, n.reinit = 10000,
    n.batch = 100, n.predupdate = 100, n.printallpars = 1000,
    m.prior = F, B.prior = F, verbose = 3,
    pxy = F, exy = F, ex = F,
    s = 15, S.init = NULL, q = 5, R = NULL, B.init = NULL,
    cc = 5, C = NULL, m.init = NULL, a = NULL, A = NULL,
    alpha = 1, a0 = 1, b0 = 1,
    k0 = NULL, Y = NULL)
```

Arguments

n	number of data records
p	dimension of each data record
n.iter	number MCMC iterations
n.discard	initial transient
n.reinit	reinitialize every n.reinit iterations (not normally used)
n.batch	save imputed parameters every n.batch iterations
n.predupdate	save posterior predictive summaries every n.predupdate iterations
n.printallpars	print all parameters every n.printallpars iterations
m.prior	indicator for resampling (1) versus fixing (0) the mean of the base measure, m
B.prior	indicator for resampling (1) versus fixing (0) the variance-covariance matrix of the base measure, B
verbose	level of comments
pxy	indicator for evaluating posterior predictive $p(x, y \mid data)$ for a future observation. Indices of x, y are given in ix and iy
exy	indicator for evaluating posterior predictive expectation $E(z \mid x, y, data)$ for a future observation. Indices of x, y, z are given in ix, iy and iz
ex	indicator for evaluating posterior predictive $p(x, y \mid data)$ for a future observation. Indices of x, y are given in ix and iy
S.init	initial value for the kernel covariance matrix S
s	degrees of freedom for the inverse Wishart prior on S
q	degrees of freedom for the inverse Wishart base measure for $G_0(V_i)$
R	expectation of the inverse Wishart base measure $G_0(V_i)$
R	Describe R here
B.init	initial value for the covariance matrix B of the base measure $G_0(\mu) = N(m, B)$
cc	degrees of freedom of the inverse Wishart hyperprior for B
C	mean of the inverse Wishart hyperprior for B
m.init	initial value for the mean m of the base measure $G_0(\mu) = N(m, B)$
a	hyperprior mean for m
A	hyperprior covariance matrix for m
alpha	initial value for the total mass parameter $alpha$
a0	hyperprior parameters for prior on total mass paramter α
b0	hyperprior parameters for prior on total mass paramter α
k0	initial number of distinct clusters
Y	n by p data matrix

Details

See mdp-package for a statement of the probability model. The function mdp initializes and carries out MCMC posterior simulation. Simulation output is saved in the working directory. Change it by using setwd if desired.

Value

The function returns no value. Simulation output is written to files.

Note

Careful, `mdp` writes temporary files into the current working directory. The same files are used by `plt.pxy`, `plt.ex` and `plt.exy` to plot posterior predictive distributions and expectations.

References

the package uses the parametrization defined in:

MacEachern, S.N. and Mueller, P. (1998). "Estimating Mixture of Dirichlet Process Models," Journal of Computational and Graphical Statistics, 7, 223–239.

See Also

See also the DPpackage, at http://student.kuleuven.be/~s0166452/software.html.

Examples

```
## Not run:
## Data from Lubischew, A. (1962), "On the use of discriminant functions
##           in taxonomy," Biometrics, 18, 455-477.
data.dir <- system.file("demo",package="mdp")
beetles <- file.path(data.dir,"beetle.data")
Y <- read.table(beetles)   # beetle data. The last column reports species
p <- ncol(Y)
Y <- as.matrix(Y[,-p]) # don't use the species indicator

## run MCMC
##
mdp(Y=Y,
    pxy=1,exy=1,ex=1,ix=1,iy=2,iz=3,
    q=10,cc=10)

## plot results
##
plt.pxy(img=T,lg=T)          ## biv density estimate p(x,y),
                             ## x=1st col (ix), y=2nd column (iy)
points(Y[,1],Y[,2],pch=19)   # add data points

plt.pxy(img=F,lg=F)          # same as a contour plot
points(Y[,1],Y[,2],pch=19)

plt.ex(sd=T)                 ## E(z | x, data), for z=3rd col (iz)

plt.exy()                    ## E(z | x,y, data)

## note: for marginals on different coordinates need to
## re-run mdp with new values for ix,iy,iz
## End(Not run)
```

plt.ex *Conditional mean curve f(x)*

Description

Plots the estimated conditional mean curve $E[f(x) \mid data]$, for the $f_H(x) = E_H(z \mid x)$. The first expectation is under the posterior distribution on H given the data. The second expectation is with respect to H.

Usage

```
plt.ex <- function(xlab="X",zlab="Z",
                   xlim=NULL,ylim=NULL,
                   bty="l",
                   sd=FALSE,
                   sim=FALSE)
```

Arguments

xlab	label on the x-axis
ylab	label on the vertical axis (plotting z).
xlim	domain of the x-axis
ylim	domain of the vertical axis
sd	
sim	

indicator for adding 10 random draws for f_H.

Details

Need to call mdp first to carry out the posterior Markov chain Monte Carlo simulation. The function plt.ex uses the simulation output to produce the desired posterior estimated conditional mean function. The function assumes the simulation output is saved in the current working directory. Change it by using setwd if necessary.

The random draws (under sim=TRUE) are generated from $p(f_H \mid data)$. See mdp-package for a statement of the probability model for the random probability measure H.

Value

The function returns no value.

Note

Careful, mdp writes temporary files into the current working directory.

References

the package uses the parametrization defined in:

MacEachern, S.N. and Mueller, P. (1998). "Estimating Mixture of Dirichlet Process Models," Journal of Computational and Graphical Statistics, 7, 223–239.

See Also

See also the R library DPpackage, at http://student.kuleuven.be/~s0166452/software.html.

Examples

```
## Not run:
## Data from Lubischew, A. (1962), "On the use of discriminant functions
##           in taxonomy," Biometrics, 18, 455-477.
data.dir <- system.file("demo",package="mdp")
beetles <- file.path(data.dir,"beetle.data")
Y <- read.table(beetles)   # beetle data. The last column reports species
p <- ncol(Y)
Y <- as.matrix(Y[,-p]) # don't use the species indicator

## run MCMC
##
mdp(Y=Y,
    pxy=1,exy=1,ex=1,ix=1,iy=2,iz=3,
    q=10,cc=10)

## plot results
##
plt.pxy(img=T,lg=T)          ## biv density estimate p(x,y),
                             ## x=1st col (ix), y=2nd column (iy)
points(Y[,1],Y[,2],pch=19)   # add data points

plt.pxy(img=F,lg=F)          # same as a contour plot
points(Y[,1],Y[,2],pch=19)

plt.ex(sd=T)                 ## E(z | x, data), for z=3rd col (iz)

plt.exy()                    ## E(z | x,y, data)

## note: for marginals on different coordinates need to
## re-run mdp with new values for ix,iy,iz
## End(Not run)
```

`plt.exy` *Conditional mean surface f(x,y).*

Description

Plots the estimated conditional mean curve $E[f(x,y) \mid data]$, for the $f_H(x,y) = E_H(z \mid x,y)$. The first expectation is under the posterior distribution on H given the data. The second expectation is with respect to H.

Usage

```
plt.exy <- function(xlab="X",ylab="Y",zlab="Z",
                    plt.contour=TRUE)
```

Arguments

xlab	label on the x-axis
ylab	label on the y-axis
zlab	label on the z-axis

plt.contour

indicator for using a contour plot vs. an image plot

Details

Need to call mdp first to carry out the posterior Markov chain Monte Carlo simulation. The function plt.exy uses the simulation output to produce the desired posterior estimated conditional mean function. The function assumes the simulation output is saved in the current working directory. Change it by using setwd if necessary.

See mdp-package for a statement of the probability model for the random probability measure H.

Value

The function returns no value.

Note

Careful, mdp writes temporary files into the current working directory.

References

the package uses the parametrization defined in:

MacEachern, S.N. and Mueller, P. (1998). "Estimating Mixture of Dirichlet Process Models," Journal of Computational and Graphical Statistics, 7, 223–239.

See Also

See also the R library DPpackage, at http://student.kuleuven.be/~s0166452/software.html.

Examples

```
## Not run:
## Data from Lubischew, A. (1962), "On the use of discriminant functions
##           in taxonomy," Biometrics, 18, 455-477.
data.dir <- system.file("demo",package="mdp")
beetles <- file.path(data.dir,"beetle.data")
Y <- read.table(beetles)   # beetle data. The last column reports species
p <- ncol(Y)
Y <- as.matrix(Y[,-p]) # don't use the species indicator

## run MCMC
##
mdp(Y=Y,
    pxy=1,exy=1,ex=1,ix=1,iy=2,iz=3,
    q=10,cc=10)

## plot results
```

```
##
plt.pxy(img=T,lg=T)          ## biv density estimate p(x,y),
                             ## x=1st col (ix), y=2nd column (iy)
points(Y[,1],Y[,2],pch=19)   # add data points

plt.pxy(img=F,lg=F)          # same as a contour plot
points(Y[,1],Y[,2],pch=19)

plt.ex(sd=T)                 ## E(z | x, data), for z=3rd col (iz)

plt.exy()                    ## E(z | x,y, data)

## note: for marginals on different coordinates need to
## re-run mdp with new values for ix,iy,iz
## End(Not run)
```

`plt.pxy` *Bivariate posterior predictive*

Description

Plots the bivariate density estimate $p(x, y) = E[H(x, y) \mid data]$.

Usage

```
plt.pxy <- function(xlab="X",ylab="Y",
                    xlim=NULL,ylim=NULL,
                    lg=F,
                    img=T)
```

Arguments

`xlab`	label on the x-axis
`ylab`	label on the y-axis
`xlim`	domain of the x-axis
`ylim`	domain of the y-axis
`lg`	
`img`	

indicator for image plot vs. contours

Details

Need to call `mdp` first to carry out the posterior Markov chain Monte Carlo simulation. The function `plt.pxy` uses the simulation output to produce the desired posterior predictive distribution. The function assumes the simulation output is saved in the current working directory. Change it by using `setwd` if necessary.

See `mdp-package` for a statement of the probability model for the random probability measure H.

Value

The function returns no value.

Note

Careful, `mdp` writes temporary files into the current working directory. The same files are used by `plt.pxy`, `plt.ex` and `plt.exy` to plot posterior predictive distributions and expectations.

References

the package uses the parametrization defined in:

MacEachern, S.N. and Mueller, P. (1998). "Estimating Mixture of Dirichlet Process Models," Journal of Computational and Graphical Statistics, 7, 223–239.

See Also

See also the R library DPpackage, at http://student.kuleuven.be/~s0166452/software.html.

Examples

```
## Not run:
## Data from Lubischew, A. (1962), "On the use of discriminant functions
##           in taxonomy," Biometrics, 18, 455-477.
data.dir <- system.file("demo",package="mdp")
beetles <- file.path(data.dir,"beetle.data")
Y <- read.table(beetles)   # beetle data. The last column reports species
p <- ncol(Y)
Y <- as.matrix(Y[,-p]) # don't use the species indicator

## run MCMC
##
mdp(Y=Y,
    pxy=1,exy=1,ex=1,ix=1,iy=2,iz=3,
    q=10,cc=10)

## plot results
##
plt.pxy(img=T,lg=T)          ## biv density estimate p(x,y),
                             ## x=1st col (ix), y=2nd column (iy)
points(Y[,1],Y[,2],pch=19)   # add data points

plt.pxy(img=F,lg=F)          # same as a contour plot
points(Y[,1],Y[,2],pch=19)

plt.ex(sd=T)                 ## E(z | x, data), for z=3rd col (iz)

plt.exy()                    ## E(z | x,y, data)

## note: for marginals on different coordinates need to
## re-run mdp with new values for ix,iy,iz
## End(Not run)
```