Hierarchical DP Mixture

April 21, 2007

R topics documented:

`hdpmn-package` *Hierarchical mixture of Dirichlet process of normals (HDPMN)*

Description

Inference for a DP mixture of normal model for related random probability measures.

Details

Package:	hdpmn
Type:	Package
Version:	1.0
Date:	2007-04-15
License:	GNU

The function `hdpmn` intializes and carries out Markov chain Monte Carlo posterior simulation. Use the function `hdpmnPredict` to obtain posterior predictive draws which can be used to estimate desired summaries.

The model is a DP mixture of normals for related random probability measures H_j. Each random measure is assumed to arise as a mixture $H_j = \epsilon F_0 + (1-\epsilon)F_j$ of one common distribution F_0 and a distribution F_j that is specific to the j-th submodel.

See *Mueller, Quintana and Rosner (2004)* for details of the model. In summary, the implemented model is as follows. Without loss of generality we assume that each submodel corresponds to a different study in a set of related studies. Let θ_{ij} denote the i-th observation in the j-th study (we use θ, assuming that the model would typically be used for a random effects distribution). We assume that $\theta_{ji}, i = 1, \ldots, n_j$ are samples from a random probability measure for the j-th study, which in turn is a mixture of a measure F_0 that is common to all studies, and an idiosyncratic measure F_j that is specific to the j-th study.

$$\theta_{ji} \sim \epsilon F_0 + (1-\epsilon)F_j$$

The random probability measures F_j in turn are given a Dirichlet process mixture of normal prior. We assume

$$F_j(\theta) = \int N(\mu, S) dG_j(\mu),\ j = 0, 1, \ldots, J$$

with $G_j \sim DP(G^\star(\eta), \alpha)$. Here η are hyperparameters that index the base measure of the DP prior. We use a normal base measure and a conjugate hyperprior

$$G^\star(\mu) = N(m, B),\ \text{with}\ m \sim N(a, A),\ \text{and}\ B^{-1} \sim Wishart(c, (cC)^{-1})$$

The Wisharet prior is parametrized such that $E(B^{-1} = C^{-1})$. Let δ_x denote a point mass at x. We complete the model with the hyperpriors

$$S^{-1} \sim W(q, (qR)^{-1}),\ p(\epsilon) = \pi_0\delta_0 + \pi_1\delta_1 + (1 - \pi_0 - \pi_1)Be(a_\epsilon, b_\epsilon)$$

Regression on observation-specific covariates x_{ji} can be achieved by including x_{ji} with the outcome θ_{ji}, and proceeding as if (x_{ji}, θ_{ji}) were generated as θ_{ji} in the model described above. See *Mueller et al. (2004, section 3.3)* for details.

Author(s)

Peter Mueller

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

References

Mueller, P., Quintana, F. and Rosner, G. (2004). "Hierarchical Meta-Analysis over Related Nonparametric Bayesian Models." *Journal of the Royal Statistical Society, Series B*, 66, 735–749.

hdpmn *MCMC for the hierarchical DP mixture*

Description

Initializes and runs posterior MCMC for the hierarchical DP mixture of normals for dependent random probability measures.

Usage

```
hdpmn(Z = NULL, study = NULL,
      npa = NULL,  nstudies = NULL,
      n.iter = 1000, n.discard = 100, n.batch = 50,
      verbose = 3, seed1 = 981963, seed2 = 6869504,
      mcmc.eps = 0, eps = 0.1, ae = 1, be = 1, pe1 = 0.1, pe0 = 0.1,
      pz = NULL, px = NULL,
      m.prior = 0, B.prior = 0, S.prior = 0,
      alpha.prior = 0, n.predupdate = 100,
      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, header = T)
```

Arguments

`Z`	Data and covariates, as a (`n` by `p`) matrix or file name, where $p = \texttt{pz} + \texttt{px}$. The i-th row of `Z` reports the i-th observation, as a combined vector of the `pz`-dimensional response and a `px`-dimensional covariate vector.
`study`	(1 by `n`) vector of study indicators. The i-th index is the study j that response i belongs to.
`npa`	Total number of observations (patients), counting across all studies.
`nstudies`	Number of studies, $= J$ in $j = 1, \ldots, J$.
`n.iter`	Number of Markov chain Monte Carlo iterations.
`n.discard`	Initial burn-in to be discarded.
`n.batch`	Save imputed paramter values every `n.batch` iterations.
`verbose`	0 is silent, 3 is verbose.
`seed1`	Random variate seed.
`seed2`	Random variate seed.
`mcmc.eps`	Indicator for resampling ϵ. If zero, ϵ will be fixed.
`eps`	Initial value for ϵ.
`ae`	Prior paramters for a Beta prior on ϵ
`be`	
`pe1`	Point mass at $\epsilon = 1$
`pe0`	Point mass at $\epsilon = 0$
`pz`	Dimension of the data vector θ_{ji}, without the covariate vector.
`px`	Dimension of a covariate vector.
`m.prior`	Indicator for resampling m, the mean of the normal base measure.
`B.prior`	Indicator for resampling B, the covariance matrix of the normal base measure.
`S.prior`	Indicator for resampling S, the covariance matrix of the normal kernel in the DP mixture.
`alpha.prior`	Indicator for resampling α, the total mass parameter of the DP prior.
`n.predupdate`	Batch size to update posterior predictive inference in the MCMC simulation.
`S.init`	initial value for S. `NULL` is allowed.
`q`	degrees of freedom in the inverse Wishart prior for S
`R`	matrix-variate parameter for the inverse Wishart prior for S
`B.init`	initial value for B, the covariance matrix in the DP base measure.
`cc`	degrees of freedom in the inverse Wishart prior for B
`C`	matrix-variate parameter for the inverse Wishart prior for B
`m.init`	initial value for m.
`a`	mean of the normal hyperprior for m.
`A`	covariance matrix of the normal hyperprior for m.
`alpha`	initial value of the total mass parameter in the DP prior.
`a0`	hyperparameters in the Gamma prior for $alpha$.
`b0`	
`k0`	initial number of clusters.
`work.dir`	directory where working files will be saved. `NULL` indicates to use the current working directory.
`header`	indicator whether the data file `Z`) includes a header.

Details

The function sets up and carries out posterior Markov chain Monte Carlo (MCMC) simulation for a hierarchical DP mixture model.

See hdpmn-package for a statement of the probability model.

Value

The function returns no value. MCMC simulations are saved in files in the designated working directory.

Use hdpmnPredict to plot summaries.

Author(s)

Peter Mueller ⟨pm@wotan.mdacc.tmc.edu⟩

References

Mueller, P., Quintana, F. and Rosner, G. (2004). “Hierarchical Meta-Analysis over Related Nonparametric Bayesian Models.” *Journal of the Royal Statistical Society, Series B*, 66, 735–749.

Examples

```
## Not run:
###
### hdpm.R - Demo
###

require(hdpmn)

## data files
data.dir <- system.file("data",package="hdpmn")

## data files
Z <- file.path(data.dir,"CALGBz.txt")
     ## data (first 7 columns) and covariates (last 3 columns)
X <- file.path(data.dir,"CALGBz0.txt")
     ## same as Z, for future patients (for prediction)
S <- file.path(data.dir,"CALGBstudy.txt")
     ## table of patient number and study index
pa.st <- as.matrix(read.table(S,header=T))
study <- pa.st[,2]  # get study index

## run MCMC -- save working files in current working directory
hdpmn(Z=Z,nstudies=3,n.iter=500,study=study,px=3,q=15,cc=15,
      work.dir=work.dir)

## post-process MCMC output for predictive inference
## save posterior predictive simulations in z00 ... z30
z10 <- hdpmnPredict(X=X,px=3,j=1,r=0) # post prediction for study 1
z20 <- hdpmnPredict(X=X,px=3,j=2,r=0) # .. study 2
z30 <- hdpmnPredict(X=X,px=3,j=3,r=0) # .. population at large (= study 3)

z11 <- hdpmnPredict(X=X,px=3,j=1,r=1) # idiosyncratic measures study 1
z21 <- hdpmnPredict(X=X,px=3,j=2,r=1) # .. study 2
z00 <- hdpmnPredict(X=X,px=3,j=0,r=0) # common measure
```

```
## covariates (and dummy random effects) of future patients
X <- as.matrix(read.table(X,header=T))
colnames(z00) <- c("PATIENT",colnames(X))

## plot estimated density for future patients in study 1, 2 and
## in population at large
idx <- which(z10[,1]==1)   ## PATIENT 1
options(digits=2)

par(mfrow=c(2,1))

## plot prediction fo study 1,2,population
plot  (density(z10[idx,8]),
       ylim=c(0,1.5),xlim=c(-0.5,2.5),
       xlab="SLOPE OF RECOVERY",bty="l",main="FUTURE PAT 1")
lines (density(z20[idx,8]),type="l",col=2)
lines (density(z30[idx,8]),type="l",col=3)
legend(-0.5,1.5,col=1:3,legend=c("STUDY 1","STUDY 2","POPULATION"),
       lty=c(1,1,1),bty="n")

## common and idiosyncratic measures
plot (density(z00[idx,8]),type="l",col=4,lty=1,
       ylim=c(0,1.5),xlim=c(-0.5,2.5),
       xlab="SLOPE OF RECOVERY",bty="l",main="COMMON & IDIOSYNC PARTS")
lines (density(z11[idx,8]),type="l",col=1,lty=2)
lines (density(z21[idx,8]),type="l",col=2,lty=2)
legend(1.5,1.5,col=c(1,2,4),lty=c(2,2,1),
       legend=c("STUDY 1 (idiosyn.)",
                "STUDY 2 (idiosyn.)",
                "COMMON"),bty="n")

## plot estimated density for future patients in study 1, 2 and
## in population at large
idx <- which(z10[,1]==2)   ## PATIENT 2
options(digits=2)

par(mfrow=c(2,1))

plot  (density(z10[idx,8]),
       ylim=c(0,1.5),xlim=c(-0.5,2.5),
       xlab="SLOPE OF RECOVERY",bty="l",main="FUTURE PAT 2")
lines (density(z20[idx,8]),type="l",col=2)
lines (density(z30[idx,8]),type="l",col=3)
legend(-0.5,1.5,col=1:3,legend=c("STUDY 1","STUDY 2","POPULATION"),
       lty=c(1,1,1),bty="n")

plot (density(z00[idx,8]),type="l",col=4,lty=1,
       ylim=c(0,1.5),xlim=c(-0.5,2.5),
       xlab="SLOPE OF RECOVERY",bty="l",main="COMMON & IDIOSYNC PARTS")
lines (density(z11[idx,8]),type="l",col=1,lty=2)
lines (density(z21[idx,8]),type="l",col=2,lty=2)
legend(1.5,1.5,col=c(1,2,4),lty=c(2,2,1),
       legend=c("STUDY 1 (idiosyn.)",
                "STUDY 2 (idiosyn.)",
                "COMMON"),bty="n")
```

```
## plot nadir count by covariate, for population
z2 <- z30[,3]; ctx <- z30[,9]; gm <- z30[,10]; amf <- z30[,11]
## fix covariates gm (GM-CSF) and amf (aminofostine)
idx <- which( (gm==-1.78) & (amf== -0.36) )
boxplot(split(z2,ctx),
        xlab="CYCLOPHOSPHAMIDE",bty="n",ylab="NADIR COUNT")
## End(Not run)
```

`hdpmnPredict` *Posterior inference for the dependent random probability measures.*

Description

Generates posterior predictive draws for future patients (observations) from the random probability measures.

Usage

```
hdpmnPredict(
  j = 1, r = 0,
  nsim = 100,
  npa = NULL, p = NULL, px = NULL,
  idx.x = NULL,
  X = NULL,
  work.dir = NULL, header = T)
```

Arguments

`j` study

`r` indicator for including (0) or not (1) the common measure.

`nsim` Number of imputed posterior simulations to use.

`p` number of columns for `X`, i.e., dimension of the random effects vector including the covariates.

`px` Dimension of covariate vector.

`idx.x` vector of size `px`, columns (starting to count at 1 for the 1st column) that contain the `px` covariates. The remaining columns are dummies corresponding to the (p-`px`)-dimensional response vector.

`npa` Number of rows in `X`, i.e., number of **future patients for posterior predictive inference.**

`X` Random effects (dummy values) and covariates for future patients, (`npa` by `p`) matrix. `X` can be a matrix of a file name.

`work.dir` directory to save working files.

`header` indicator for a header line in `X`

Details

Must run hdpmn first to generate posterior simulations.

The function carries out post-processing of the MCMC posterior simulation to generate posterior predictive simulation for future observations from the random probability measures defined in the model. See hdpmn-package for a statement of the probability model.

For npa assumed future patients with given covariates (specified in X) the function computes posterior predictive inference of future responses. The subvector of responses is a dummy to match the dimension.

Value

The function returns a matrix zout with p+1 columns of posterior predictive simulations for the npa future patients. The first column is a patient index. An index i refers to the i-th row in the matrix X of given patient covariates. Columns 2 through p+1 are posterior predictive simulations including the (unchanged) covariate vector in the locations indicated by idx.

Scatterplots, density estimates etc. of the posterior predictive simulations can be used to evaluate posterior means for the RPMs, and to evaluate posterior predictive probabilities for events of interest for future subjects (patients).

See the examples below for examples on how to summarize the posterior predictive simulations.

Author(s)

Peter Mueller

Examples

```
## Not run:
###
### hdpm.R - Demo
###

require(hdpmn)

## data files
data.dir <- system.file("data",package="hdpmn")

## data files
Z <- file.path(data.dir,"CALGBz.txt")
     ## data (first 7 columns) and covariates (last 3 columns)
X <- file.path(data.dir,"CALGBz0.txt")
     ## same as Z, for future patients (for prediction)
S <- file.path(data.dir,"CALGBstudy.txt")
     ## table of patient number and study index
pa.st <- as.matrix(read.table(S,header=T))
study <- pa.st[,2]  # get study index

## run MCMC -- save working files in current working directory
hdpmn(Z=Z,nstudies=3,n.iter=500,study=study,px=3,q=15,cc=15,
      work.dir=work.dir)

## post-process MCMC output for predictive inference
## save posterior predictive simulations in z00 ... z30
z10 <- hdpmnPredict(X=X,px=3,j=1,r=0) # post prediction for study 1
z20 <- hdpmnPredict(X=X,px=3,j=2,r=0) # .. study 2
```

```
z30 <- hdpmnPredict(X=X,px=3,j=3,r=0) # .. population at large (= study 3)

z11 <- hdpmnPredict(X=X,px=3,j=1,r=1) # idiosyncratic measures study 1
z21 <- hdpmnPredict(X=X,px=3,j=2,r=1) # .. study 2
z00 <- hdpmnPredict(X=X,px=3,j=0,r=0) # common measure

## covariates (and dummy random effects) of future patients
X <- as.matrix(read.table(X,header=T))
colnames(z00) <- c("PATIENT",colnames(X))

## plot estimated density for future patients in study 1, 2 and
## in population at large
idx <- which(z10[,1]==1)   ## PATIENT 1
options(digits=2)

par(mfrow=c(2,1))

## plot prediction fo study 1,2,population
plot  (density(z10[idx,8]),
       ylim=c(0,1.5),xlim=c(-0.5,2.5),
       xlab="SLOPE OF RECOVERY",bty="l",main="FUTURE PAT 1")
lines (density(z20[idx,8]),type="l",col=2)
lines (density(z30[idx,8]),type="l",col=3)
legend(-0.5,1.5,col=1:3,legend=c("STUDY 1","STUDY 2","POPULATION"),
       lty=c(1,1,1),bty="n")

## common and idiosyncratic measures
plot (density(z00[idx,8]),type="l",col=4,lty=1,
       ylim=c(0,1.5),xlim=c(-0.5,2.5),
       xlab="SLOPE OF RECOVERY",bty="l",main="COMMON & IDIOSYNC PARTS")
lines (density(z11[idx,8]),type="l",col=1,lty=2)
lines (density(z21[idx,8]),type="l",col=2,lty=2)
legend(1.5,1.5,col=c(1,2,4),lty=c(2,2,1),
       legend=c("STUDY 1 (idiosyn.)",
                "STUDY 2 (idiosyn.)",
                "COMMON"),bty="n")

## plot estimated density for future patients in study 1, 2 and
## in population at large
idx <- which(z10[,1]==2)   ## PATIENT 2
options(digits=2)

par(mfrow=c(2,1))

plot  (density(z10[idx,8]),
       ylim=c(0,1.5),xlim=c(-0.5,2.5),
       xlab="SLOPE OF RECOVERY",bty="l",main="FUTURE PAT 2")
lines (density(z20[idx,8]),type="l",col=2)
lines (density(z30[idx,8]),type="l",col=3)
legend(-0.5,1.5,col=1:3,legend=c("STUDY 1","STUDY 2","POPULATION"),
       lty=c(1,1,1),bty="n")

plot (density(z00[idx,8]),type="l",col=4,lty=1,
       ylim=c(0,1.5),xlim=c(-0.5,2.5),
       xlab="SLOPE OF RECOVERY",bty="l",main="COMMON & IDIOSYNC PARTS")
lines (density(z11[idx,8]),type="l",col=1,lty=2)
lines (density(z21[idx,8]),type="l",col=2,lty=2)
```

```
legend(1.5,1.5,col=c(1,2,4),lty=c(2,2,1),
       legend=c("STUDY 1 (idiosyn.)",
                "STUDY 2 (idiosyn.)",
                "COMMON"),bty="n")

## plot nadir count by covariate, for population
z2 <- z30[,3]; ctx <- z30[,9]; gm <- z30[,10]; amf <- z30[,11]
## fix covariates gm (GM-CSF) and amf (aminofostine)
idx <- which( (gm==-1.78) & (amf== -0.36) )
boxplot(split(z2,ctx),
        xlab="CYCLOPHOSPHAMIDE",bty="n",ylab="NADIR COUNT")
## End(Not run)
```

Index