Package 'profdpm'

February 20, 2015

Type Package

Title Profile Dirichlet Process Mixtures

Version 3.3

Date 2013-05-24

Author Matt Shotwell

Maintainer Matt Shotwell <matt.shotwell@vanderbilt.edu>

Description This package facilitates profile inference (inference at the posterior mode) for a class of product partition models (PPM). The Dirichlet process mixture is currently the only available member of this class. These methods search for the maximum posterior (MAP) estimate for the data partition in a PPM.

License GPL (>= 2)

LazyLoad yes

NeedsCompilation yes

Repository CRAN

Date/Publication 2013-05-25 07:09:56

R topics documented:

pci *Partition Comparison Indices*

Description

This function computes several partition comparison indices.

Usage

```
pci(x1, x2)
```

Arguments

x1	a factor
x2	a factor

Details

This function computes indices of similarity between two factors representing the cluster partition of n items. The two vectors must be of the same length. Let n_{11} be the number of item pairs that occur in the same cluster in both partitions x1 and x2, n_{00} the number of item pairs that occur in differenct clusters in both partitions, n_{10} the number of item pairs that occur in the same cluster in partition x1 but in different clusters in parition x2, and n_{01} the number of item pairs that occur in different clusters in partition x1 but in the same cluster in partition x2. The Rand index is given by

$$\frac{n_{11} + n_{00}}{n_{11} + n_{00} + n_{01} + n_{10}}.$$

The Fowlkes and Mallows index is given by

$$\frac{n_{11}}{\sqrt{(n_{11} + n_{01})(n_{11} + n_{10})}}.$$

The Wallace indices are respectively given by

$$\frac{n_{11}}{n_{11} + n_{10}} \qquad \frac{n_{11}}{n_{11} + n_{01}}.$$

The Jaccard index is given by

$$\frac{n_{11}}{n_{11} + n_{01} + n_{10}}.$$

Value

A named vector with the following elements:

R	Rand index
FM	Fowlkes and Mallows index
W10	Wallace 10 index
W01	Wallace 01 index
J	Jaccard index

Author(s)

Matt Shotwell

References

Matthew S. Shotwell (2013). **profdpm**: An R Package for MAP Estimation in a Class of Conjugate Product Partition Models. *Journal of Statistical Software*, **53(8)**, 1-18. URL http://www.jstatsoft.org/v53/i08/.

Rand, W. (1971) Objective Criteria for the Evaluation of Clustering Methods. Journal of the American Statistical Association 66:846-850

Fowlkes, E. B. and Mallows, C. L. (1983) A Method for Comparing Two Hierarchical Clusterings. Journal of the American Statistical Association 78:553-569

Wallace, D. L. (1983) A Method for Comparing Two Hierarchical Clusterings: comment. Journal of the American Statistical Association 78:569-576

`profBinary` *Binary Product Partition Models*

Description

This function finds the most probable cluster partition in a binary product partition model (PPM). The Dirichlet process mixture of binary models is the default PPM.

Usage

```
profBinary(formula, data, clust, param, method="agglomerative",
           maxiter=1000, crit=1e-6, verbose=FALSE, sampler=FALSE)
```

Arguments

`formula`	a one-sided formula specifying a set of binary response variables.
`data`	a dataframe where `formula` is evaluated
`clust`	optional vector of factors (or coercible to factors) indicating initial clustering among observations.
`param`	optional list containing the any of the named elements `alpha`, `a0`, and `b0` corresponding to the prior parameters of the beta-binary Dirichlet process mixture. The prior parameters of the beta-binary Dirichlet process mixture should all be scalars.
`method`	character string indicating the optimization method to be used. Meaningful values for this string are `"stochastic"`, `"gibbs"`, `"agglomerative"`, and `"none"`.

- The `"stochastic"` method is an iterative stochastic search utilizing 'explode' and 'merge' operations on the clusters of a partition. At the explode step, a randomly selected subset of observations are redistributed uniformly at random to an existing or new cluster. Each of the exploded observations are then merged with an existing cluster in a sequentially optimal fashion.

Optimization involves computing a moving average of the relative change in the marginal posterior distribution over the possible clusters after each iteration. The optimization stopping criterion is the minumim value this quantity can take before stopping the optimization cycle. If the optimization cycle reaches the maximum allowable iterations before meeting the stopping criterion, a warning is issued.

- The `"gibbs"` method implements the Polya urn Gibbs sampler. This method draws samples from the posterior distribution over the cluster partition in a sequential Gibbs fashion. The sample value with greatest posterior mass is returned. See MacEachern (1994) for details.
- The `"agglomerative"` method initially places each observation into seperate clusters. At each iteration, two of the remaining clusters are merged, where the merged clusters are chosen such that the resulting increase in the posterior mass function is maximized. This is repeated until only one cluster remains. The MAP estimate is the cluster partition, among those considered, which maximizes the posterior mass function over the possible cluster partitions. See Ward (1963) for additional details.
- The `"fast"` method is a modified version of Sequential Update and Greedy Search (SUGS). This SUGS algorithm assigns observations to clusters sequentially. Initally, the first observation forms a singleton cluster. Subsequent observations are assigned to an existing cluster, or form a new singleton cluster by optimizing the associated posterior probabilities, conditional on previous cluster assignments. See Wang and Dunson (2010) for additional details.
- The `"none"` method is typically used in conjunction with the `clust` option to specify an initial cluster partition. If `"none"` is specified without `clust`, a simple algorithm is used to initialize the cluster partition. Otherwise, the cluster partition is initialized using the `clust` argument. The posterior statistics are then computed for initialized clusters.

`maxiter` integer value specifying the maximum number of iterations for the optimization algorithm.

`crit` numeric scalar constituting a stopping criterion for the `"stochastic"` and `"gibbs"` optimization methods.

`verbose` logical value indicating whether the routine should be verbose in printing.

`sampler` for the "gibbs" method, return the last sampled value instead of the MAP estimate

Details

This function fits a Dirichlet process mixture of binary models (DPMBM) using the profile method. This method will cluster binary observations vectors (rows of `y`) into clusters. The cluster partition is estimated by maximizing the marginal posterior distribution over all possible cluster partitions. Each cluster has an associated binary model. The binary model assigns Bernoulli probabilities independently to each binary valued outcome, corresponding to the columns of `y`. The prior parameters `a0` and `b0` assign a beta prior distribution to each outcome probability. Conditional on the estimated cluster partition, each outcome probability is beta distributed *a posteriori*. The function

profBinary returns the associated posterior parameters of the beta destribution for each cluster and outcome probability.

Missing observations (NA) are removed automatically and a warning is issued. The return value contains the reduced observation matrix.

Value

An instance of the class profBinary containing the following objects

y	the numeric matrix of observations, where rows with missing observations (NA) are removed
param	the list of prior parameters
clust	a numeric vector of integers indicating cluster membership for each non-missing observation
a	a list of numeric vectors containing the posterior vector a for each cluster
b	a list of numeric vectors containing the posterior vector b for each cluster
logp	the unnormalized log value of the marginal posterior mass function for the cluster partition evaluated at clust
model	a model frame, resulting from a call to model.frame

Author(s)

Matt Shotwell

References

Matthew S. Shotwell (2013). **profdpm**: An R Package for MAP Estimation in a Class of Conjugate Product Partition Models. *Journal of Statistical Software*, **53(8)**, 1-18. URL http://www.jstatsoft.org/v53/i08/.

Ward, J. H. (1963) Heirarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58:236-244 MacEachern, S. N. (1994) Estimating Normal Means with Conjugate Style Dirichlet Process Prior. Communications in Statistics B 23:727-741

See Also

pci

Examples

```
library(profdpm)
set.seed(42)

# simulate two clusters of multivariate binary data
p <- seq(0.9,0.1,length.out=3)
y1 <- matrix(rbinom(333, 1, p), 111, 3, TRUE)
y2 <- matrix(rbinom(333, 1, rev(p)), 111, 3, TRUE)
dat <- as.data.frame(rbind(y1, y2))

# fit the PPM
```

```
fitb <- profBinary(~0+., data=dat)

# plot the data ordered by cluster
image(t(as.matrix(fitb$model)[order(fitb$clust),]),
      xaxt="n", yaxt="n", col=0:1)
axis(3, labels=paste("V", 1:3, sep=""), at=0:2/2)

# plot the data ordered and colored by cluster
image(t(as.matrix(fitb$model) * fitb$clust)[, order(fitb$clust)],
      xaxt="n", yaxt="n", col=0:length(unique(fitb$clust)))
axis(3, labels=paste("V", 1:3, sep=""), at=0:2/2)
```

`profdpm` *Profile Inference for Dirichlet Process Mixtures and Other Product Partition Models*

Description

This package facilitates profile inference (inference at the posterior mode) for a class of product partition models (PPM). The Dirichlet process mixture is a specific case in this class and is selected by default. These methods search for the maximum posterior (MAP) estimate for the cluster partition in the PPM.

Details

Package:	profdpm
Type:	Package
Version:	3.3
Date:	2013-05-24
License:	GPL (>=2)
LazyLoad:	yes

Contents

- `profLinear` Linear Product Partition Models
- `profBinary` Binary Product Partition Models
- `pci` Partition Comparison Indices

Author(s)

Matt Shotwell Maintainer: Matt Shotwell <matt.shotwell@vanderbilt.edu>

References

Matthew S. Shotwell (2013). **profdpm**: An R Package for MAP Estimation in a Class of Conjugate Product Partition Models. *Journal of Statistical Software*, **53(8)**, 1-18. URL http://www.jstatsoft.org/v53/i08/.

profLinear *Linear Product Partition Models*

Description

This function finds the most probable cluster partition in a linear product partition model (PPM). The Dirichlet process mixture of linear models is the default PPM.

Usage

```
profLinear(formula, data, group, clust, param, method="agglomerative",
           maxiter=1000, crit=1e-6, verbose=FALSE, sampler=FALSE)
```

Arguments

formula	a formula.
data	a dataframe where formula is evaluated
group	optional vector of factors (or coercible to factors) indicating grouping among observations. Observations that are grouped will be clustered together. This is useful if several values form a single longitudinal observation.
clust	optional vector of factors (or coercible to factors) indicating initial clustering among observations. Grouped observations (see group) must have the same clust value.
param	optional list containing the any of the named elements alpha, a0, b0, m0, and s0 corresponding to the prior parameters of the normal-gamma Dirichlet process mixture. The prior parameters of the normal-gamma Dirichlet process mixture should all be scalars except m0 which should be a vector of length equal to the number of columns in x.
method	character string indicating the optimization method to be used. Meaningful values for this string are "stochastic", "gibbs", "agglomerative", and "none".

- The "stochastic" method is an iterative stochastic search utilizing 'explode' and 'merge' operations on the clusters of a partition. At the explode step, a randomly selected subset of observations are redistributed uniformly at random to an existing or new cluster. Each of the exploded observations are then merged with an existing cluster in a sequentially optimal fashion. Optimization involves computing a moving average of the relative change in the marginal posterior distribution over the possible clusters after each iteration. The optimization stopping criterion is the minumim value this quantity can take before stopping the optimization cycle. If the optimization cycle reaches the maximum allowable iterations before meeting the stopping criterion, a warning is issued.
- The "gibbs" method implements the Polya urn Gibbs sampler. This method draws samples from the posterior distribution over the cluster partition in a sequential Gibbs fashion. The sample value with greatest posterior mass is returned. See MacEachern(1994) for details.

- The "agglomerative" method initially places each observation into seperate clusters. At each iteration, two of the remaining clusters are merged, where the merged clusters are chosen such that the resulting increase in the posterior mass function is maximized. This is repeated until only one cluster remains. The MAP estimate is the cluster partition, among those considered, which maximizes the posterior mass function over the possible cluster partitions. See Ward (1963) for additional details.
- The "fast" method is a modified version of Sequential Update and Greedy Search (SUGS). This SUGS algorithm assigns observations to clusters sequentially. Initally, the first observation forms a singleton cluster. Subsequent observations are assigned to an existing cluster, or form a new singleton cluster by optimizing the associated posterior probabilities, conditional on previous cluster assignments. See Wang and Dunson (2010) for additional details.
- The "none" method is typically used in conjunction with the clust option to specify an initial cluster partition. If "none" is specified without clust, a simple algorithm is used to initialize the cluster partition. Otherwise, the cluster partition is initialized using the clust argument. The posterior statistics are then computed for initialized clusters.

maxiter	integer value specifying the maximum number of iterations for the optimization algorithm.
crit	numeric scalar constituting a stopping criterion for the "stochastic" and "gibbs" optimization methods.
verbose	logical value indicating whether the routine should be verbose in printing.
sampler	for the "gibbs" method, return the last sampled value instead of the MAP estimate

Details

This function fits a Dirichlet process mixture of linear models (DPMLM) using the profile method. This method will group the observations into clusters. The clusters are determined by maximizing the marginal posterior distribution over the space of possible clusters. Each cluster has an associated linear model. Notationally, the linear model for cluster k has the form

$$y = \gamma' x + \epsilon,$$

where y and x are the observation vector and covariate matrix for a particular cluster, ϵ has a multivariate normal distribution with mean zero and precision matrix τI, and γ is the vector of linear coefficients. In the DPMLM, conditional on the clustering, γ and τ have a joint normal-gamma posterior distribution of the form

$$p(m, \tau | y, x) = N(\gamma | m, \tau S) G(\tau | a, b),$$

where $N(\cdot)$ is the multivariate normal density function with mean vector m and precision matrix τS and $G(\cdot)$ is the gamma density function with shape and scale parameters a and b. In addition to the cluster indicators, the posterior quantities S, m, a, and b are provided for each cluster in the return value.

Missing observations (NA) are removed automatically and a warning is issued.

Value

An instance of the class `profLinear` containing the following objects

`y`	the numeric outcome vector, where missing observations (`NA`) are removed
`x`	the numeric design matrix, where missing covariates (`NA`) are removed
`group`	the grouping vector, where missing group values (`NA`) are removed
`param`	the list of prior parameters
`clust`	a numeric vector of integers indicating cluster membership for each non-missing observation
`a`	a numeric vector containing the posterior parameter a for each cluster
`b`	a numeric vector containing the posterior parameter b for each cluster
`m`	a list of numeric vectors containing the posterior vector m for each cluster
`s`	a list of numeric matrices containing the posterior matrix S for each cluster
`logp`	the unnormalized log value of the marginal posterior mass function for the cluster partition evaluated at `clust`
`model`	a model frame, resulting from a call to `model.frame`

Author(s)

Matt Shotwell

References

Matthew S. Shotwell (2013). **profdpm**: An R Package for MAP Estimation in a Class of Conjugate Product Partition Models. *Journal of Statistical Software*, **53(8)**, 1-18. URL http://www.jstatsoft.org/v53/i08/.

Ward, J. H. (1963) Heirarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association 58:236-244 MacEachern, S. N. (1994) Estimating Normal Means with Conjugate Style Dirichlet Process Prior. Communications in Statistics B 23:727-741

See Also

`pci`

Examples

```
library(profdpm)
set.seed(42)

# set up some data
# linear model 0
x0  <- rnorm(50, 0, 3)
y0  <- x0 + rnorm(50, 0, 1)

# linear model 1
x1  <- rnorm(50, 0, 3)
y1  <- 10 - x1 + rnorm(50, 0, 1)
```

```
# add a column of ones to the covariate matrix (intercept)
dat <- data.frame(x=c(x0, x1), y=c(y0,y1))

# indicate grouping within each linear model
grp <- c(rep(seq(0,4),10), rep(seq(5,9),10))

# fit the DPMLM
fit <- profLinear(y ~ x, data=dat, group=grp)

# plot the resulting fit(s)
plot(dat$x, dat$y, xlab='x', ylab='y')
for(i in 1:length(fit$m)) {
  abline( a=fit$m[[i]][1], b=fit$m[[i]][2] )
}
```

---

summary.profBinary *Summarize objects of class* profBinary.

Description

summary.profBinary is an S3 method to summarize objects of the class profBinary.

Usage

```
## S3 method for class 'profBinary'
summary(object, ...)
```

Arguments

object	an instance of class profBinary
...	additional arguments (not used)

Details

The summary.profBinary function outputs summary information using the cat and print functions. For each unique value of x$clust, the summary.profBinary function outputs the number of observation groups assigned to the corresponding cluster. The estimated outcome probabilities and their 95% credible intervals are also printed for each cluster. The 95% credible intervals are computed using the marginal posterior distribution, conditional on the estimated data partition. See the package vignette for additional information.

Value

A list of lists, one for each unique cluster, each with the following elements:

groups	The number of observation groups assigned to the corresponding cluster
summary	A data frame containing the estimate and 95% credible limits for each outcome probability

Author(s)

Matt Shotwell <matt.shotwell@vanderbilt.edu>

References

Matthew S. Shotwell (2013). **profdpm**: An R Package for MAP Estimation in a Class of Conjugate Product Partition Models. *Journal of Statistical Software*, **53(8)**, 1-18. URL http://www.jstatsoft.org/v53/i08/.

See Also

profBinary

`summary.profLinear` *Summarize objects of class* `profLinear`.

Description

`summary.profLinear` is an S3 method to summarize objects of the class `profLinear`.

Usage

```
## S3 method for class 'profLinear'
summary(object, ...)
```

Arguments

`object`	an instance of class `profLinear`
`...`	additional arguments (not used)

Details

The `summary.profLinear` function outputs summary information using the `cat` and `print` functions. For each unique value of `x$clust`, the `summary.profLinear` function outputs the number of observations, and observation groups assigned to the corresponding cluster. The estimated linear coefficients and their approximate 95% credible intervals are also printed for each cluster. The 95% credible intervals are computed using a Laplace approximation to the marginal posterior distribution for the linear coefficients, conditional on the estimated data partition. See the package vignette for additional information.

Value

A list of lists, one for each unique cluster, each with the following elements:

`groups`	The number of observation groups assigned to the corresponding cluster
`obs`	The number of observations assigned to the corresponding cluster
`summary`	A data frame containing the estimate and 95% credible limits for each linear coefficient

Author(s)

Matt Shotwell <matt.shotwell@vanderbilt.edu>

References

Matthew S. Shotwell (2013). **profdpm**: An R Package for MAP Estimation in a Class of Conjugate Product Partition Models. *Journal of Statistical Software*, **53(8)**, 1-18. URL http://www.jstatsoft.org/v53/i08/.

See Also

profLinear

Index