Package 'rmp'

December 13, 2023

Type Package

Title Rounded Mixture Package

Version 2.2

Date 2023-12-12

Author Antonio Canale [aut, cre],
Nicola Lunardon [ctb]

Maintainer Antonio Canale <antonio.canale@unipd.it>

Description
Performs univariate probability mass function estimation via Bayesian nonparametric mixtures of rounded kernels as in Canale and Dunson (2011) <doi:10.1198/jasa.2011.tm10552>.

License GPL-2

NeedsCompilation yes

Repository CRAN

Date/Publication 2023-12-13 11:50:02 UTC

R topics documented:

`dp.post.est` *Posterior probability mass function estimation with DP prior*

Description

Performs Bayesian probability mass function estimation under DP prior with Poisson base measure.

Usage

```
dp.post.est(x, y, alpha, lambda)
```

Arguments

x	Values on which to compute the pmf.
y	Vector of observed data.
alpha	DP precision parameter
lambda	Mean parameter for the Poisson base measure.

Details

Performs probability mass function estimation under th following model

$$y_i \mid P \sim P, i = 1, \ldots, n$$

$$P \sim DP(\alpha, P_0),$$

where P_0 is Poisson with mean `lambda`.

Value

A vector of size `length(x)` containing the probability masses

Author(s)

Antonio Canale

References

Carota, C., and Parmigiani, G. (2002), "Semiparametric Regression for Count Data," *Biometrika*, **89**, 265–281.

Examples

```
data(ethylene)
y <- tapply(ethylene$impl,FUN=mean,INDEX=ethylene$id)
z <- tapply(ethylene$dose,FUN=mean,INDEX=ethylene$id)

# Estimate the pmf of the number of implants in the control group
y0  <- y[z==0]
pmf.control = dp.post.est(0:30, y0, alpha = 1)
```

ethylene *Developmental toxicity study of ethylene glycol in mice*

Description

Data from the developmental toxicity study of ethylene glycol in mice conducted by the National Toxicology Program. Pregnant mice were assigned to dose groups of 0, 750, 1,500 or 3,000 mg/kg per day, with the number of implants measured for each mouse at the end of the experiment. Group sizes are 25, 24, 23 and 23, respectively.

Usage

```
data(ethylene)
```

Format

A data frame with 1192 observations on the following 7 variables.

`id` identifyer for pregnant mouse

`dose` dose of ethylene glycol

`weight` weight of the fetus

`sex` sex of the fetus

`impl` number of implants in the pregnant mouce

`litsz` size of the relative litter

`malf` presence of malformation in the fetus

References

Price, C. J., Kimmel, C. A., Tyl, R. W., and Marr, M. C. (1985) "The developemental toxicity of ethylene glycol in rats and mice" *Toxicological and Applied Pharmacology* **81**, 113-127.

Examples

```
data(ethylene)
implants <- tapply(ethylene$impl, FUN=mean, INDEX=ethylene$id)

summary(implants)
m <- mean(implants)
v <- var(implants)
hist(implants, main=paste("Histogram of the number of implants (Mean = ",
round(m,2), ", Var = ", round(v,2),")"))
```

npmp *DP mixtures of Poissons*

Description

Performs probability mass function estimation under nonparametric mixture of Poisson kernels.

Usage

```
npmp(y, k, nrep, nb, alpha=1, theta=alpha, sigma=0,
mixing_hyperprior= FALSE, basemeasure_hyperprior = FALSE, mixing_type="DP",
algo="slice",  prior="gamma", a, b, a_a=1, b_a=1, lb=NULL, ub=NULL,
print = 1, ndisplay = nrep/4, plot.it = FALSE, pdfwrite = FALSE, ... )

dpmpoiss(y, k, nrep, nb, alpha = 1, a, b, lb = NULL, ub = NULL, print = 1,
ndisplay = nrep/4, plot.it = FALSE, pdfwrite = FALSE, ...)
```

Arguments

y	Vector of count data
k	Truncation level for the number of cluster in the mixture. Default is `length(ydis)`.
nrep	Number of MCMC iterations
nb	Number of burn-in iteration in the MCMC to discard
alpha	Value of the precision parameter of the Dirichlet process prior
theta	Value of the strength parameter of the Two-parameters-Poisson-Dirichlet process prior
sigma	Value of the discount parameter of the Two-parameters-Poisson-Dirichlet process prior
mixing_hyperprior	Logical. If TRUE `alpha` is random with gamma hyperprior
basemeasure_hyperprior	Logical. If TRUE also the parameters of the base measure are random, see details below.
mixing_type	Type of mixing distribution. Default is "DP" for Dirichlet process but also "2PD" for Two-parameters-Poisson-Dirichlet process is allowed.
algo	Type of algorithm. Current choices are: slice sampler (`algo="slice"`) or polya-urn-type sampler (`algo="polya-urn"`).
prior	String for the base measure prior. Default is "gamma" for lambda ~ Gamma(a,b). The other choice is "normal" for exp(lambda) ~ N(a,b)
a	Shape (mean) hyperparameter for the gamma (normal) base measure
b	Scale (sd) hyperparameter for the gamma (normal) prior
a_a	Shape hyperparameter for `alpha`

b_a	Scale hyperparameter for alpha
lb	Scalar integer. Lower bound for the argument of the pmf. Default is max(0,min(ydis)-10).
ub	Scalar integer. Upper bound for the argument of the pmf. Default is max(ydis)+10.
print	Vector of integers (from 1 to 5) indicating whether to print each step of the Gibbs sampler. Specifically, 1 for current iteration, 2 for the DP cluster allocation, 3 for the posterior parameters of the mixture components, 4 for the precision of the DP, 5 for the posterior pmf.
ndisplay	Scalar integer. It gives the number of iterations to be displayed on screen (the function reports on the screen when every ndisplay iterations have been carried out)
plot.it	Logical, default FALSE. If TRUE a plot with empirical and estimated posterior probability mass functions is plotted.
pdfwrite	Logical, default FALSE. If TRUE a pdf file is written in the current working directory. Traceplots and other posterior quantities are drown.
...	Additional arguments (for future implemetantions).

Details

The function npmp performs probability mass function estimation under nonparametric mixture of Poisson kernels, i.e.

$$y_i \mid \lambda_i \sim \text{Poi}(\lambda_i), i = 1, \ldots, n$$
$$\lambda_i \mid G \sim G$$
$$G \sim \Pi(P_0),$$

where Π is a nonparametric prior (Dirichlet process or Two-parameters-Poisson-Dirichlet process) with base measure P_0. The function dppoiss is a wrapper to npmp with mixing_type="DP" for back portability with version 1.0 of the package. The main part of the code is written in C language to gain computational speed. Plots and posterior summaries are in plain R code. From version 2.0 on, the blocked Gibbs sampler has been removed in place of slice samper (Kalli et al., 2011) and polya-urn sampler. Two different base measures P_0 are implemented: prior="gamma" and prior="normal" for

$$\lambda_h \sim \text{Gamma}(a, b), \log(\lambda_h) \sim N(a, b), h = 1, \ldots$$

respectively.

Value

name	Name of the model
mixing_type	Name of the mixing prior
mcmc	Quantities about MCMC sampling
mcmc.chains	MCMC chains of the parameters
pmf	A list containing several quantities related to the probability mass function (emprical pmf, posterior mean pmf and pointwise 95% credible intervals) computed for the values from lb to ub

parameters	A list containing the posterior mean of the cluster specific parameters (be careful of label-switching problems)
clustering	A list containing posterior quantities related to the clustering structure of the data

Author(s)

R code and porting by A. Canale, C code by A. Canale with minor contributions by N. Lunardon.

References

Canale, A. and Dunson, D. B. (2011), "Bayesian Kernel Mixtures for Counts", *Journal of American Statistical Association*, **106**, 1528-1539.

Kalli, M., Griffin, J., and Walker, S. (2011), "Slice sampling mixture models," Statistics and Computing, **21**, 93-105.

See Also

rmg

Examples

```
data(ethylene)
y <- tapply(ethylene$impl,FUN=mean,INDEX=ethylene$id)
z <- tapply(ethylene$dose,FUN=mean,INDEX=ethylene$id)

# Estimate the pmf of the number of implants in the control group
y0  <- y[z==0]
pmf.control = dpmpoiss(y0, k=20, nrep=11000, nb=1000, alpha=1, a=1, b=1,
lb=5, ub=24, plot.it=TRUE)
```

---

rmg *Nonparametric mixture of rounded Gaussians*

Description

Performs Bayesian probability mass function estimation under nonparametric mixture of rounded of Gaussian kernels.

Usage

```
rmg(ydis, k=length(ydis), nrep, nb, alpha=1, theta=alpha, sigma=0,
mixing_hyperprior= FALSE, basemeasure_hyperprior = FALSE, mixing_type="DP", algo="slice",
mu0=mean(ydis), kap=var(ydis),
atau=1, btau=2, a_a=1, b_a=1, lb=NULL, ub=NULL, print = 1, ndisplay = nrep/4,
plot.it = FALSE, pdfwrite = FALSE, ...)
```

```
dpmrg(ydis, k, nrep, nb, alpha, alpha_r = FALSE, mu0 = mean(ydis), kap = var(ydis),
atau, btau, a_a = 1, b_a = 1, lb = NULL, ub = NULL,
print = 1, ndisplay = nrep/4,
plot.it = FALSE, pdfwrite = FALSE, ...)
```

Arguments

ydis	Vector of count data
k	Truncation level for the number of cluster in the mixture. Default is `length(ydis)`.
nrep	Number of MCMC iterations
nb	Number of burn-in iteration in the MCMC to discard
alpha	Value of the precision parameter of the Dirichlet process prior
theta	Value of the strength parameter of the Two-parameters-Poisson-Dirichlet process prior
sigma	Value of the discount parameter of the Two-parameters-Poisson-Dirichlet process prior
mixing_hyperprior	Logical. If TRUE `alpha` is random with gamma hyperprior
alpha_r	Logical. If TRUE `alpha` is random with gamma hyperprior
basemeasure_hyperprior	Logical. If TRUE also the parameters of the base measure are random, see details below.
mixing_type	Type of mixing distribution. Default is "DP" for Dirichlet process but also "2PD" for Two-parameters-Poisson-Dirichlet process is allowed.
algo	Type of algorithm. Current choices are: slice sampler (`algo="slice"`) or polya-urn-type sampler (`algo="polya-urn"`).
mu0	Location hyperparameter for the latent rounded Gaussian base measure
kap	Precision hyperparameter for the latent rounded Gaussian base measure
atau	Shape hyperparameter for the Gamma distribution
btau	Scale hyperparameter for the Gamma distribution
a_a	Shape hyperparameter for the Gamma distribution for `alpha`
b_a	Scale hyperparameter for the Gamma distribution for `alpha`
lb	Scalar integer. Lower bound for the argument of the pmf. Default is `max(0,min(ydis)-10)`.
ub	Scalar integer. Upper bound for the argument of the pmf. Default is `max(ydis)+10`.
print	Vector of integers (from 1 to 6) indicating whether to print each step of the Gibbs sampler. Specifically, 1 for current iteration, 2 for the data augmentation step simulating the latent continuous variables, 3 for the DP cluster allocation, 4 for the posterior parameters of the mixture components, 5 for the precision of the DP, 6 for the posterior pmf.

ndisplay	Scalar integer. It gives the number of iterations to be displayed on screen (the function reports on the screen when every ndisplay iterations have been carried out)
plot.it	Logical, default FALSE. If TRUE a plot with empirical and estimated posterior probability mass functions is plotted.
pdfwrite	Logical, default FALSE. If TRUE a pdf file is written in the current working directory. Traceplots and other posterior quantities are drown.
...	Additional arguments (for future implemetantions).

Details

The rmg function performs Bayesian probability mass function estimation under the mixture model of Canale and Dunson (2011) with Dirichlet process or Two-parameters-Poisson-Dirichlet process as prior for the mixing measure. The model is

$$y_i \mid \mu_i, \tau_i \sim \text{RG}(\mu_i, \tau_i), i = 1, \ldots, n$$

$$(\mu_i, \tau_i) \mid G \sim G$$

$$G \sim \Pi(P_0),$$

where Π is the Dirichlet process or the Two-parameters-Poisson-Dirichlet process with base measure P_0 and $RG(\mu, \tau)$ is a rounded Gaussian kernel with location μ and precision τ and tresholds $-\infty, 1, 2, \ldots$. The function dpmrg is a wrapper to rmg with mixing_type="DP" for back portability with version 1.0 of the package. The main part of the code is written in C language to gain computational speed. Plots and posterior summaries are in plain R code. From version 2.0 on, the blocked gibbs sampler has been removed in place of slice samper (Kalli et al., 2011) and polya-urn sampler.

Value

name	Name of the model
mixing_type	Name of the mixing prior
mcmc	Quantities about MCMC sampling
mcmc.chains	MCMC chains of the parameters
pmf	A list containing several quantities related to the probability mass function (emprical pmf, posterior mean pmf and pointwise 95% credible intervals) computed for the values from lb to ub
parameters	A list containing the posterior mean of the cluster specific parameters (be careful of label-switching problems)
clustering	A list containing posterior quantities related to the clustering structure of the data

Author(s)

R code and porting by A. Canale, C code by A. Canale with minor contributions by N. Lunardon.

References

Canale, A. and Dunson, D. B. (2011), "Bayesian Kernel Mixtures for Counts", *Journal of American Statistical Association*, **106**, 1528-1539.

Kalli, M., Griffin, J., and Walker, S. (2011), "Slice sampling mixture models," Statistics and Computing, **21**, 93-105.

See Also

dpmpoiss

Examples

```
data(ethylene)
y <- tapply(ethylene$impl,FUN=mean,INDEX=ethylene$id)
z <- tapply(ethylene$dose,FUN=mean,INDEX=ethylene$id)

# Estimate the pmf of the number of implants in the control group
y0  <- y[z==0]
pmf.control = rmg(y0, k=20, nrep=11000, nb=1000, alpha=1, atau=1, btau=1,
lb=5, ub=24, plot.it= TRUE)
```

Index