Package 'BNSP'

February 27, 2017

Title Bayesian Non- And Semi-Parametric Model Fitting

Version 1.1.1

Date 2017-02-24

Author Georgios Papageorgiou

Maintainer Georgios Papageorgiou <gpapageo@gmail.com>

Description MCMC for Dirichlet process mixtures.

Depends R (>= 3.1.0)

Suggests mvtnorm

License GPL (>= 2)

URL http://www.bbk.ac.uk/ems/faculty/papageorgiou/BNSP

NeedsCompilation yes

Repository CRAN

Date/Publication 2017-02-27 08:29:45

R topics documented:

`BNSP-package` *Bayesian non- and semi-parametric model fitting*

Description

MCMC for Dirichlet process mixtures

Details

Package:	BNSP
Type:	Package
Version:	1.1.0
Date:	2015-12-04
License:	GPL (>=2)

Acknowledgments

This work was partly supported by the Medical Research Council grant number G09018401.

Author(s)

Georgios Papageorgiou (2014)

Maintainer: Georgios Papageorgiou <gpapageo@gmail.com>

References

Papageorgiou, G., Richardson, S. and Best, N. (2015). Bayesian nonparametric models for spatially indexed data of mixed type. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 77:973-999.

---

bnpglm *Bayesian nonparametric generalized linear models*

Description

Fits Dirichlet process mixtures of joint response-covariate models, where the covariates are continuous while the discrete responses are represented utilizing continuous latent variables. See 'Details' section for a full model description.

Usage

```
bnpglm(formula,family,data,offset,sampler="slice",StorageDir,
       ncomp,sweeps,burn,thin=1,seed,prec,V,Vdf,Mu.nu,Sigma.nu,
       Mu.mu,Sigma.mu,Alpha.xi,Beta.xi,Alpha.alpha,Beta.alpha,Turnc.alpha,
       Xpred,offsetPred,...)
```

Arguments

formula	a formula defining the response and the covariates e.g. y ~ x.
family	a description of the kernel of the response variable. Currently eight options are supported: 1. "poisson", 2. "negative binomial", 3. "generalized poisson", 4. "hyper-poisson", 5. "ctpd", 6. "com-poisson", 7. "binomial" and 8. "beta binomial". The first six kernels are used for count data analysis while the last two are used for binomial data analysis. Kernels 3.-6. allow for both over- and under-dispersion relative to the Poisson distribution. See 'Details' section for some of the kernel details.
data	an optional data frame, list or environment (or object coercible by 'as.data.frame' to a data frame) containing the variables in the model. If not found in 'data', the variables are taken from 'environment(formula)'.
offset	this can be used to specify an a priori known component to be included in the linear predictor during fitting. This should be 'NULL' or a numeric vector of length equal to the sample size. One 'offset' term can be included in the formula, and if more are required, their sum should be used.
sampler	the MCMC algorithm to be utilized. The two options are sampler="slice" which implements a slice sampler (Walker, 2007; Papaspiliopoulos, 2008) and sampler="truncated" which proceeds by truncating the countable mixture at ncomp components (see argument ncomp).
StorageDir	a directory to store files with the posterior samples of models parameters and other quantities of interest. If a directory is not provided, files are created in the current directory and removed when the sampler completes.
ncomp	number of mixture components. Defines where the countable mixture of densities [in (1) below] is truncated. Even if sampler="slice" is chosen, ncomp needs to be specified as it is used in the initialization process.
sweeps	total number of posterior samples, including those discarded in burn-in period (see argument burn) and those discarded by the thinning process (see argument thin).
burn	length of burn-in period.
thin	thinning parameter.
seed	optional seed for the random generator.
prec	precision parameter. Updating the parameters of the response distribution requires a Metropolis - Hastings step, with proposal distributions centered at current values and with precision equal to this argument. It can be of length one (for "poisson" and "binomial" kernels) or of length two (for "negative binomial", "beta binomial", "generalized-poisson", "hyper-poisson" and "com-poisson" kernels) or of length three (for the "ctpd" kernel).
V	optional scale matrix V of the prior Wishart distribution assigned to precision matrix T_h. See 'Details' section.
Vdf	optional degrees of freedom Vdf of the prior Wishart distribution assigned to precision matrix T_h. See 'Details' section.
Mu.nu	optional prior mean μ_ν of the covariance vector ν_h. See 'Details' section.

`Sigma.nu` optional prior covariance matrix Σ_ν of ν_h. See 'Details' section.

`Mu.mu` optional prior mean μ_μ of the mean vector μ_h. See 'Details' section.

`Sigma.mu` optional prior covariance matrix Σ_μ of μ_h. See 'Details' section.

`Alpha.xi` an optional parameter that depends on the specified family.

1. If `family="poisson"`, this argument is parameter α_ξ of the prior of the Poisson rate: $\xi \sim$ Gamma(α_ξ, β_ξ).
2. If `family="negative binomial"`, this argument is a two-dimensional vector that includes parameters $\alpha_{1\xi}$ and $\alpha_{2\xi}$ of the priors: $\xi_1 \sim$ Gamma$(\alpha_{1\xi}, \beta_{1\xi})$ and $\xi_2 \sim$ Gamma$(\alpha_{2\xi}, \beta_{2\xi})$, where ξ_1 and ξ_2 are the two parameters of the Negative Binomial pmf.
3. If `family="generalized-poisson"`, this argument is a two-dimensional vector that includes parameters $\alpha_{1\xi}$ and $\alpha_{2\xi}$ of the priors: $\xi_1 \sim$ Gamma$(\alpha_{1\xi}, \beta_{1\xi})$ and $\xi_2 \sim$ N$(\alpha_{2\xi}, \beta_{2\xi})I[\xi_2 \in R_{\xi_2}]$, where ξ_1 and ξ_2 are the two parameters of the Generalized Poisson pmf. Parameter ξ_2 has to be in the range R_{ξ_2} (which is automatically done during posterior sampling).
4. If `family="hyper-poisson"`, this argument is a two-dimensional vector that includes parameters $\alpha_{1\xi}$ and $\alpha_{2\xi}$ of the priors: $\xi_1 \sim$ Gamma$(\alpha_{1\xi}, \beta_{1\xi})$ and $\xi_2 \sim$ Gamma$(\alpha_{2\xi}, \beta_{2\xi})$, where ξ_1 and ξ_2 are the two parameters of the hyper Poisson pmf.
5. If `family="ctpd"`, this argument is a three-dimensional vector that includes parameters $\alpha_{1\xi}, \alpha_{2\xi}$ and $\alpha_{3\xi}$ of the priors: $\xi_i \sim$ Gamma$(\alpha_{i\xi}, \beta_{i\xi}), i = 1, 2,$ and $\xi_3 \sim$ N$(\alpha_{3\xi}, \beta_{3\xi})I[\xi_3 \in R_{\xi_3}]$, where $\xi_i, i = 1, 2, 3$, are the three parameters of the complex triparametric Pearson distribution. Parameter ξ_3 has to be in the range R_{ξ_3} (which is automatically done during posterior sampling).
6. If `family="com-poisson"`, this argument is a two-dimensional vector that includes parameters $\alpha_{1\xi}$ and $\alpha_{2\xi}$ of the priors: $\xi_1 \sim$ Gamma$(\alpha_{1\xi}, \beta_{1\xi})$ and $\xi_2 \sim$ Gamma$(\alpha_{2\xi}, \beta_{2\xi})$, where ξ_1 and ξ_2 are the two parameters of the COM-Poisson pmf.
7. If `family="binomial"`, this argument is parameter α_ξ of the prior of the Binomial probability: $\xi \sim$ Beta(α_ξ, β_ξ).
8. If `family="beta binomial"`, this argument is a two-dimensional vector that includes parameters $\alpha_{1\xi}$ and $\alpha_{2\xi}$ of the priors: $\xi_1 \sim$ Gamma$(\alpha_{1\xi}, \beta_{1\xi})$ and $\xi_2 \sim$ Gamma$(\alpha_{2\xi}, \beta_{2\xi})$, where ξ_1 and ξ_2 are the two parameters of the Beta Binomial pmf.

See 'Details' section.

`Beta.xi` an optional parameter that depends on the specified family.

1. If `family="poisson"`, this argument is parameter β_ξ of the prior of the Poisson rate: $\xi \sim$ Gamma(α_ξ, β_ξ).
2. If `family="negative binomial"`, this argument is a two-dimensional vector that includes parameters $\beta_{1\xi}$ and $\beta_{2\xi}$ of the priors: $\xi_1 \sim$ Gamma$(\alpha_{1\xi}, \beta_{1\xi})$ and $\xi_2 \sim$ Gamma$(\alpha_{2\xi}, \beta_{2\xi})$, where ξ_1 and ξ_2 are the two parameters of the Negative Binomial pmf.
3. If `family="generalized poisson"`, this argument is a two-dimensional vector that includes parameters $\beta_{1\xi}$ and $\beta_{2\xi}$ of the priors: $\xi_1 \sim$ Gamma$(\alpha_{1\xi}, \beta_{1\xi})$

and $\xi_2 \sim$ Normal$(\alpha_{2\xi}, \beta_{2\xi})I[\xi_2 \in R_{\xi_2}]$, where ξ_1 and ξ_2 are the two parameters of the Generalized Poisson pmf. Parameter ξ_2 has to be in the range R_{ξ_2} (which is automatically done during posterior sampling). Note that $\beta_{2\xi}$ is a standard deviation.

4. If `family="hyper-poisson"`, this argument is a two-dimensional vector that includes parameters $\beta_{1\xi}$ and $\beta_{2\xi}$ of the priors: $\xi_1 \sim$ Gamma$(\alpha_{1\xi}, \beta_{1\xi})$ and $\xi_2 \sim$ Gamma$(\alpha_{2\xi}, \beta_{2\xi})$, where ξ_1 and ξ_2 are the two parameters of the hyper Poisson pmf.
5. If `family="ctpd"`, this argument is a three-dimensional vector that includes parameters $\beta_{1\xi}, \beta_{2\xi}$ and $\beta_{3\xi}$ of the priors: $\xi_i \sim$ Gamma$(\alpha_{i\xi}, \beta_{i\xi}), i = 1, 2,$ and $\xi_3 \sim$ N$(\alpha_{3\xi}, \beta_{3\xi})I[\xi_3 \in R_{\xi_3}]$, where $\xi_i, i = 1, 2, 3$, are the three parameters of the complex triparametric Pearson distribution. Note that $\beta_{3\xi}$ is a standard deviation.
6. If `family="com-poisson"`, this argument is a two-dimensional vector that includes parameters $\beta_{1\xi}$ and $\beta_{2\xi}$ of the priors: $\xi_1 \sim$ Gamma$(\alpha_{1\xi}, \beta_{1\xi})$ and $\xi_2 \sim$ Gamma$(\alpha_{2\xi}, \beta_{2\xi})$, where ξ_1 and ξ_2 are the two parameters of the COM-Poisson pmf.
7. If `family="binomial"`, this argument is parameter β_ξ of the prior of the Binomial probability: $\xi \sim$ Beta(α_ξ, β_ξ).
8. If `family="beta binomial"`, this argument is a two-dimensional vector that includes parameters $\beta_{1\xi}$ and $\beta_{2\xi}$ of the priors: $\xi_1 \sim$ Gamma$(\alpha_{1\xi}, \beta_{1\xi})$ and $\xi_2 \sim$ Gamma$(\alpha_{2\xi}, \beta_{2\xi})$, where ξ_1 and ξ_2 are the two parameters of the Beta Binomial pmf.

See 'Details' section.

`Alpha.alpha` optional shape parameter α_α of the Gamma prior assigned to the concentration parameter α. See 'Details' section.

`Beta.alpha` optional rate parameter β_α of the Gamma prior assigned to concentration parameter α. See 'Details' section.

`Turnc.alpha` optional truncation point c_α of the Gamma prior assigned to concentration parameter α. See 'Details' section.

`Xpred` an optional design matrix the rows of which include the covariates x for which the conditional distribution of $Y|x, D$ (where D denotes the data) is calculated. These are treated as 'new' covariates i.e. they do not contribute to the likelihood. The matrix shouldn't include a column of 1's.

`offsetPred` the offset term associated with the new covariates `Xpred`. `offsetPred` is a vector of length equal to the rows of `Xpred`. If `family` is one of `poisson` or `negative binomial` or `generalized poisson`, its entries are the associated Poisson offsets. If `family` is one of `binomial` or `beta binomial`, its entries are the Binomial number of trials. If `offsetPred` is missing, it is taken to be the mean of `offset`, rounded to the nearest integer.

`...` Other options that will be ignored.

Details

Function bnpglm returns samples from the posterior distributions of the parameters of the model:

$$f(y_i, x_i) = \sum_{h=1}^{\infty} \pi_h f(y_i, x_i|\theta_h), \qquad (1)$$

where y_i is a univariate discrete response, x_i is a p-dimensional vector of continuous covariates, and $\pi_h, h \geq 1$, are obtained according to Sethuraman's (1994) stick-breaking construction: $\pi_1 = v_1$, and for $l \geq 2, \pi_l = v_l \prod_{j=1}^{l-1}(1 - v_j)$, where v_k are iid samples $v_k \sim$Beta $(1, \alpha), k \geq 1$.

The discrete responses y_i are represented as discretized versions of continuous latent variables y_i^*. Observed discrete and continuous latent variables are connected by:

$$y_i = q \iff c_{i,q-1} < y_i^* < c_{i,q}, q = 0, 1, 2, \ldots,$$

where the cut-points are obtained as: $c_{i,-1} = -\infty$, while for $q \geq 0$, $c_{i,q} = c_q(\lambda_i) = \Phi^{-1}\{F(q; \lambda_i)\}$. Here $\Phi(.)$ is the cumulative distribution function (cdf) of a standard normal variable and $F()$ denotes an appropriate cdf. Further, latent variables are assumed to independently follow a $N(0, 1)$ distribution, where the mean and variance are restricted to be zero and one as they are non-identifiable by the data. Choices for $F()$ are described next.

For counts, currently six options are supported. First, $F(.; \lambda_i)$ can be specified as the cdf of a Poisson$(H_i\xi_h)$ variable. Here $\lambda_i = (\xi_h, H_i)^T$, ξ_h denotes the Poisson rate associated with cluster h, and H_i the offset term associated with sampling unit i. Second, $F(.; \lambda_i)$ can be specified as the negative binomial cdf, where $\lambda_i = (\xi_{1h}, \xi_{2h}, H_i)^T$. This option allows for overdispersion within each cluster relative to the Poisson distribution. Third, $F(.; \lambda_i)$ can be specified as the Generalized Poisson cdf, where, again, $\lambda_i = (\xi_{1h}, \xi_{2h}, H_i)^T$. This option allows for both over- and under-dispersion within each cluster. The other three options, that also allow for both over- and under-dispersion relative to the Poisson distribution, are the Hyper Poisson (HP), COM-Poisson and the Complex Triparametric Pearson (CTP) kernels. The HP and COM-Poisson kernels have 2 parameters and the CTPD kernel has 3 parameters.

For Binomial data, currently two options are supported. First, $F(.; \lambda_i)$ may be taken to be the cdf of a Binomial(H_i, ξ_h) variable, where ξ_h denotes the success probability of cluster h and H_i the number of trials associated with sampling unit i. Second, $F(.; \lambda_i)$ may be specified to be the beta-binomial cdf, where $\lambda = (\xi_{1h}, \xi_{2h}, H_i)^T$.

Details on all kernels are provided in the tables below. The first table provides the probability mass functions and the mean in the presence of an offset term (which may be taken to be one). The column 'Sample' indicates for which parameters the routine provides posterior samples. The second table provides information on the assumed priors along with the default values of the parameters of the prior distributions and it also indicates the function arguments that allow the user to alter these. Lastly, the third tables provides some details on the less frequently used kernels.

Kernel	PMF	Offset	Mean	Sample
Poisson	$\exp(-H\xi)(H\xi)^y/y!$	H	$H\xi$	ξ
Negative Binomial	$\frac{\Gamma(y+\xi_1)}{\Gamma(\xi_1)\Gamma(y+1)}(\frac{\xi_2}{H+\xi_2})^{\xi_1}(\frac{H}{H+\xi_2})^y$	H	$H\xi_1/\xi_2$	ξ_1, ξ_2
Generalized Poisson	$\xi_1\{\xi_1 + (\xi_2 - 1)y\}^{y-1}\xi_2^{-y}\times$ $\exp\{-[\xi_1 + (\xi_2 - 1)y]/\xi_2\}/y!$	H	$H\xi_1$	ξ_1, ξ_2
Hyper Poisson	$\frac{1}{{}_1F_1(1,\xi_2,\xi_3)}\frac{\xi_3^y}{(\xi_2)_y}$	H	$H\xi_1 = \xi_3 -$ $(\xi_2 - 1)\frac{{}_1F_1(1,\xi_2,\xi_3)-1}{{}_1F_1(1,\xi_2,\xi_3)}$	ξ_1, ξ_2

CTP	$f_0 \frac{(\xi_3+\xi_4 i)_y (\xi_3-\xi_4 i)_y}{(\xi_2)_y y!}$	H	$H\xi_1 = \frac{\xi_3^2+\xi_4^2}{\xi_2-2\xi_3-1}$	ξ_1, ξ_2, ξ_3
COM-Poisson	$\frac{\xi_3^y}{Z(\xi_2,\xi_3)(y!)^{\xi_2}}$	H	$H\xi_1 = \xi_3 \frac{\partial \log(Z)}{\partial \xi_3}$	ξ_1, ξ_2
Binomial	$\binom{N}{y}\xi^y(1-\xi)^{N-y}$	N	$N\xi$	ξ
Beta Binomial	$\binom{N}{y}\frac{Beta(y+\xi_1, N-y+\xi_2)}{Beta(\xi_1,\xi_2)}$	N	$N\xi_1/(\xi_1+\xi_2)$	ξ_1, ξ_2

Kernel	Priors	Default Values
Poisson	$\xi \sim$ Gamma(α_ξ, β_ξ)	Alpha.xi = 1.0, Beta.xi = 0.1
Negative Binomial	$\xi_i \sim$ Gamma$(\alpha_{\xi_i}, \beta_{\xi_i}), i = 1, 2$	Alpha.xi = c(1.0,1.0), Beta.xi = c(0.1,0.1)
Generalized Poisson	$\xi_1 \sim$ Gamma$(\alpha_{\xi_1}, \beta_{\xi_1})$	
	$\xi_2 \sim TN(\alpha_{\xi_2}, \beta_{\xi_2})$ ($\beta_{\xi_2} \equiv$ st.dev.)	Alpha.xi = c(1.0,1.0), Beta.xi = c(0.1,1.0)
	TN: truncated normal	
Hyper Poisson	$\xi_i \sim$ Gamma$(\alpha_{\xi_i}, \beta_{\xi_i}), i = 1, 2$	Alpha.xi = c(1.0,0.5), Beta.xi = c(0.1,0.5)
CTP	$\xi_i \sim$ Gamma$(\alpha_{\xi_i}, \beta_{\xi_i}), i = 1, 2$	
	$\xi_3 \sim TN(\alpha_{\xi_3}, \beta_{\xi_3})$ ($\beta_{\xi_3} \equiv$ st.dev.)	Alpha.xi = c(1.0,1.0,0.0)
	TN: truncated normal	Beta.xi = c(0.1,0.1,100.0)
COM-Poisson	$\xi_i \sim$ Gamma$(\alpha_{\xi_i}, \beta_{\xi_i}), i = 1, 2$	Alpha.xi = c(1.0,0.5), Beta.xi = c(0.1,0.5)
Binomial	$\xi \sim$ Beta(α_ξ, β_ξ)	Alpha.xi = 1.0, Beta.xi = 1.0
Beta Binomial	$\xi_i \sim$ Gamma$(\alpha_{\xi_i}, \beta_{\xi_i}), i = 1, 2$	Alpha.xi = c(1.0,1.0), Beta.xi = c(0.1,0.1)

Kernel	Notes
Generalized Poisson	$\xi_1 > 0$ is the mean and $\xi_2 > 1/2$ is a dispersion parameter. When $\xi_2 = 1$, the pmf reduces to the Poisson. Parameter values $\xi_2 > 1$ suggest over-dispersion and parameter values $1/2 < \xi_2 < 1$ suggest under-dispersion relative to the Poisson.
Hyper Poisson	$\xi_1 > 0$ is the mean and $\xi_2 > 0$ is a dispersion parameter. When $\xi_2 = 1$, the pmf reduces to the Poisson. When $\xi_2 > 1$ the pmf is over-dispersed and when $\xi_2 < 1$ the pmf is under-dispersed relative to the Poisson.
COM-Poisson	The mean is $\xi_1 (> 0)$ and the variance approximately ξ_1/ξ_2, so similar comments as for the hyper Poisson hold.
CTPD	Things are a bit more complex here. See Rodriguez-Avi et al. (2004) for the details.

Further, joint vectors (y_i^*, x_i) are modeled utilizing Gaussian distributions. Then, with θ_h denoting model parameters associated with the hth cluster, the joint density $f(y_i, x_i|\theta_h)$ takes the form

$$f(y_i, x_i|\theta_h) = \int_{c_{i,y_i-1}}^{c_{i,y_i}} N_{p+1}(y_i^*, x_i|\mu_h, C_h) dy_i^*,$$

where μ_h and C_h denote the mean vector and covariance matrix, respectively.

The joint distribution of the latent variable y_i^* and the covariates x_i is

$$(y_i^*, x_i^T)^T|\theta_h \sim N_{p+1}\left(\begin{pmatrix} 0 \\ \mu_h \end{pmatrix}, \quad C_h = \begin{bmatrix} 1 & \nu_h^T \\ \nu_h & \Sigma_h \end{bmatrix} \right),$$

where ν_h denotes the vector of covariances $\text{cov}(y_i^*, x_i|\theta_h)$. Sampling from the posterior of constrained covariance matrix C_h is done using methods similar to those of McCulloch et al. (2000). Specifically, the conditional $x_i|y_i^* \sim N_p(\mu_h + y_i^*\nu_h, B_h = \Sigma_h - \nu_h\nu_h^T)$ simplifies matters as there are no constraints on matrix B_h (other than positive definiteness). Given priors for B_h and ν_h, it is easy to sample from their posteriors, and thus obtain samples from the posterior of $\Sigma_h = B_h + \nu_h\nu_h^T$.

Specification of the prior distributions:

1. Define $T_h = B_h^{-1} = (\Sigma_h - \nu_h\nu_h^T)^{-1}, h \geq 1$. We specify that a priori $T_h \sim \text{Wishart}_p(V,\text{Vdf})$, where V is a $p \times p$ scale matrix and Vdf is a scalar degrees of freedom parameter. Default values are: $V = I_p/p$ and Vdf$= p$, however, these can be changed using arguments `V` and Vdf.
2. The assumed prior for ν_h is $N_p(\mu_\nu, \Sigma_\nu), h \geq 1$, with default values $\mu_\nu = 0$ and $\Sigma_\nu = I_p$. Arguments `Mu.nu` and `Sigma.nu` allow the user to change the default values.
3. A priori $\mu_h \sim N_p(\mu_\mu, \Sigma_\mu), h \geq 1$. Here the default values are $\mu_\mu = \bar{x}$ where $\bar{x}$ denotes the sample mean of the covariates, and $\Sigma_\mu = D$ where D denotes a diagonal matrix with diagonal elements equal to the square of the observed range of the covariates. Arguments `Mu.mu` and `Sigma.mu` allow the user to change the default values.
4. For count data, with `family="poisson"`, a priori we take $\xi_h \sim \text{Gamma}(\alpha_\xi, \beta_\xi), h \geq 1$. The default values are $\alpha_\xi = 1.0, \beta_\xi = 0.1$, that define a Gamma distribution with mean $\alpha_\xi/\beta_\xi = 10$ and variance $\alpha_\xi/\beta_\xi^2 = 100$. Defaults can be altered using arguments `Alpha.xi` and `Beta.xi`.

   For count data with `family="negative binomial"` a priori we take $\xi_{jh} \sim \text{Gamma}(\alpha_{j\xi}, \beta_{j\xi})$, $j = 1, 2, h \geq 1$. The default values are $\alpha_{j\xi} = 1.0, \beta_{j\xi} = 0.1$, $j = 1, 2$. Default values for $\{\alpha_{j\xi} : j = 1, 2\}$ can be altered using argument `Alpha.xi`, and default values for $\{\beta_{j\xi} : j = 1, 2\}$ can be altered using argument `Beta.xi`.

   For count data with `family="generalized poisson"`, a priori we take $\xi_{1h} \sim \text{Gamma}(\alpha_{1\xi}, \beta_{1\xi})$, and $\xi_{2h} \sim \text{Normal}(\alpha_{2\xi}, \beta_{2\xi})I[\xi_{2h} \in R_{\xi_2}]$. The default values are $\alpha_{j\xi} = 1.0, j = 1, 2$ and $\beta_{1\xi} = 0.1, \beta_{2\xi} = 1.0$. Default values for $\{\alpha_{j\xi} : j = 1, 2\}$ can be altered using argument `Alpha.xi`, and default values for $\{\beta_{j\xi} : j = 1, 2\}$ can be altered using argument `Beta.xi`.

   For count data with `family="hyper-poisson"` a priori we take $\xi_{jh} \sim \text{Gamma}(\alpha_{j\xi}, \beta_{j\xi})$, $j = 1, 2, h \geq 1$. The default values are $\alpha_{1\xi} = 1.0, \alpha_{2\xi} = 0.5$ and $\beta_{1\xi} = 0.1, \beta_{2\xi} = 0.5$. Default values for $\{\alpha_{j\xi} : j = 1, 2\}$ can be altered using argument `Alpha.xi`, and default values for $\{\beta_{j\xi} : j = 1, 2\}$ can be altered using argument `Beta.xi`.

   For count data with `family="ctpd"`, a priori we take $\xi_{1h} \sim \text{Gamma}(\alpha_{1\xi}, \beta_{1\xi})$, $\xi_{2h} \sim \text{Gamma}(\alpha_{2\xi}, \beta_{2\xi})$ and $\xi_{3h} \sim \text{Normal}(\alpha_{3\xi}, \beta_{3\xi})I[\xi_{3h} \in R_{\xi_3}]$. The default values are $\alpha_{1\xi} = 1.0, \alpha_{2\xi} = 1.0, \alpha_{3\xi} = 0.0$ and $\beta_{1\xi} = 0.1, \beta_{2\xi} = 0.1, \beta_{3\xi} = 100.0$. Default values for $\{\alpha_{j\xi} : j = 1, 2\}$ can be altered using argument `Alpha.xi`, and default values for $\{\beta_{j\xi} : j = 1, 2\}$ can be altered using argument `Beta.xi`.

   For count data with `family="com-poisson"` a priori we take $\xi_{jh} \sim \text{Gamma}(\alpha_{j\xi}, \beta_{j\xi})$, $j = 1, 2, h \geq 1$. The default values are $\alpha_{1\xi} = 1.0, \alpha_{2\xi} = 0.5$ and $\beta_{1\xi} = 0.1, \beta_{2\xi} = 0.5$. Default values for $\{\alpha_{j\xi} : j = 1, 2\}$ can be altered using argument `Alpha.xi`, and default values for $\{\beta_{j\xi} : j = 1, 2\}$ can be altered using argument `Beta.xi`.

   For binomial data, with `family="binomial"`, a priori we take $\xi_h \sim \text{Beta}(\alpha_\xi, \beta_\xi), h \geq 1$. The default values are $\alpha_\xi = 1.0, \beta_\xi = 1.0$, that define a uniform distribution. Defaults can be altered using arguments `Alpha.xi` and `Beta.xi`.

   For binomial data with `family="beta binomial"`, a priori we take $\xi_{jh} \sim \text{Gamma}(\alpha_{j\xi}, \beta_{j\xi})$, $j = 1, 2, h \geq 1$. The default values are $\alpha_{j\xi} = 1.0, \beta_{j\xi} = 0.1$. Default values for $\{\alpha_{j\xi} : j =$

$1, 2\}$ can be altered using argument `Alpha.xi`, and default values for $\{\beta_{j\xi} : j = 1, 2\}$ can be altered using argument `Beta.xi`.

5. The concentration parameter α is assigned a Gamma$(\alpha_\alpha, \beta_\alpha)$ prior over the range (c_α, ∞), that is, $f(\alpha) \propto \alpha^{\alpha_\alpha - 1} \exp\{-\alpha\beta_\alpha\} I[\alpha > c_\alpha]$, where $I[.]$ is the indicator function. The default values are $\alpha_\alpha = 2.0, \beta_\alpha = 4.0$, and $c_\alpha = 0.25$. Users can alter the default using using arguments `Alpha.alpha`, `Beta.alpha` and `Turnc.alpha`.

Value

Function `bnpglm` returns the following:

`call`	the matched call.
`seed`	the seed that was used (in case replication of the results is needed).
`meanReg`	if `Xpred` is specified, the function returns the posterior mean of the expectation of the response given each new covariate x.
`modeReg`	if `Xpred` is specified, the function returns the posterior mean of the conditional mode of the response given each new covariate x.
`Q05Reg`	if `Xpred` is specified, the function returns the posterior mean of the conditional 5% quantile of the response given each new covariate x.
`Q10Reg`	if `Xpred` is specified, the function returns the posterior mean of the conditional 10% quantile of the response given each new covariate x.
`Q15Reg`	if `Xpred` is specified, the function returns the posterior mean of the conditional 15% quantile of the response given each new covariate x.
`Q20Reg`	if `Xpred` is specified, the function returns the posterior mean of the conditional 20% quantile of the response given each new covariate x.
`Q25Reg`	if `Xpred` is specified, the function returns the posterior mean of the conditional 25% quantile of the response given each new covariate x.
`Q50Reg`	if `Xpred` is specified, the function returns the posterior mean of the conditional 50% quantile of the response given each new covariate x.
`Q75Reg`	if `Xpred` is specified, the function returns the posterior mean of the conditional 75% quantile of the response given each new covariate x.
`Q80Reg`	if `Xpred` is specified, the function returns the posterior mean of the conditional 80% quantile of the response given each new covariate x.
`Q85Reg`	if `Xpred` is specified, the function returns the posterior mean of the conditional 85% quantile of the response given each new covariate x.
`Q90Reg`	if `Xpred` is specified, the function returns the posterior mean of the conditional 90% quantile of the response given each new covariate x.
`Q95Reg`	if `Xpred` is specified, the function returns the posterior mean of the conditional 95% quantile of the response given each new covariate x.
`denReg`	if `Xpred` is specified, the function returns the posterior mean conditional density of the response given each new covariate x. Results are presented in a matrix the rows of which correspond to the different xs.
`denVar`	if `Xpred` is specified, the function returns the posterior variance of the conditional density of the response given each new covariate x. Results are presented in a matrix the rows of which correspond to the different xs.

Further, function bnpglm creates files where the posterior samples are written. These files are (with all file names preceded by 'BNSP.'):

`alpha.txt` this file contains samples from the posterior of the concentration parameters α. The file is arranged in `(sweeps-burn)/thin` lines and one column, each line including one posterior sample.

`compAlloc.txt` this file contains the allocations or configurations obtained at each iteration of the sampler. It consists of `(sweeps-burn)/thin` lines, that represent the posterior samples, and n columns, that represent the sampling units. Entries in this file range from 0 to $ncomp - 1$.

`MeanReg.txt` this file contains the conditional means of the response y given covariates x obtained at each iteration of the sampler. The rows represent the `(sweeps-burn)/thin` posterior samples. The columns represent the various covariate values x for which the means are obtained.

`muh.txt` this file contains samples from the posteriors of the p-dimensional mean vectors $\mu_h, h = 1, 2, \ldots, ncomp$. The file is arranged in `((sweeps-burn)/thin)*ncomp` lines and p columns. In more detail, each sweep creates `ncomp` lines representing samples $\mu_h^{(sw)}, h = 1, \ldots, ncomp$, where superscript sw represents a particular sweep. The elements of $\mu_h^{(sw)}$ are written in the columns of the file.

`nmembers.txt` this file contains `(sweeps-burn)/thin` lines and `ncomp` columns, where the lines represent posterior samples while the columns represent the components or clusters. The entries represent the number of sampling units allocated to the components.

`nuh.txt` this file contains samples from the posteriors of the p-dimensional covariance vectors $\nu_h, h = 1, 2, \ldots, ncomp$. The file is arranged in `((sweeps-burn)/thin)*ncomp` lines and p columns. In more detail, each sweep creates `ncomp` lines representing samples $\nu_h^{(sw)}, h = 1, \ldots, ncomp$, where superscript sw represents a particular sweep. The elements of $\nu_h^{(sw)}$ are written in the columns of the file.

`Q05Reg.txt` this file contains the 5% conditional quantile of the response y given covariates x obtained at each iteration of the sampler. The rows represent the `(sweeps-burn)/thin` posterior samples. The columns represent the various covariate values x for which the quantiles are obtained.

`Q10Reg.txt` this file contains the 10% conditional quantile of the response y given covariates x obtained at each iteration of the sampler. The rows represent the `(sweeps-burn)/thin` posterior samples. The columns represent the various covariate values x for which the quantiles are obtained.

`Q15Reg.txt` this file contains the 15% conditional quantile of the response y given covariates x obtained at each iteration of the sampler. The rows represent the `(sweeps-burn)/thin` posterior samples. The columns represent the various covariate values x for which the quantiles are obtained.

`Q20Reg.txt` this file contains the 20% conditional quantile of the response y given covariates x obtained at each iteration of the sampler. The rows represent the `(sweeps-burn)/thin` posterior samples. The columns represent the various covariate values x for which the quantiles are obtained.

`Q25Reg.txt` this file contains the 25% conditional quantile of the response y given covariates x obtained at each iteration of the sampler. The rows represent the `(sweeps-burn)/thin` posterior samples. The columns represent the various covariate values x for which the quantiles are obtained.

`Q50Reg.txt` this file contains the 50% conditional quantile of the response y given covariates x obtained at each iteration of the sampler. The rows represent the `(sweeps-burn)/thin` posterior samples. The columns represent the various covariate values x for which the quantiles are obtained.

`Q75Reg.txt` this file contains the 75% conditional quantile of the response y given covariates x obtained at each iteration of the sampler. The rows represent the `(sweeps-burn)/thin` posterior samples. The columns represent the various covariate values x for which the quantiles are obtained.

`Q80Reg.txt` this file contains the 80% conditional quantile of the response y given covariates x obtained at each iteration of the sampler. The rows represent the `(sweeps-burn)/thin` posterior samples. The columns represent the various covariate values x for which the quantiles are obtained.

`Q85Reg.txt` this file contains the 85% conditional quantile of the response y given covariates x obtained at each iteration of the sampler. The rows represent the `(sweeps-burn)/thin` posterior samples. The columns represent the various covariate values x for which the quantiles are obtained.

`Q90Reg.txt` this file contains the 90% conditional quantile of the response y given covariates x obtained at each iteration of the sampler. The rows represent the `(sweeps-burn)/thin` posterior samples. The columns represent the various covariate values x for which the quantiles are obtained.

`Q95Reg.txt` this file contains the 95% conditional quantile of the response y given covariates x obtained at each iteration of the sampler. The rows represent the `(sweeps-burn)/thin` posterior samples. The columns represent the various covariate values x for which the quantiles are obtained.

`Sigmah.txt` this file contains samples from the posteriors of the $p \times p$ covariance matrices $\Sigma_h, h = 1, 2, \ldots, ncomp$. The file is arranged in `((sweeps-burn)/thin)*ncomp` lines and p^2 columns. In more detail, each sweep creates `ncomp` lines representing samples $\Sigma_h^{(sw)}, h = 1, \ldots, ncomp$, where superscript sw represents a particular sweep. The elements of $\Sigma_h^{(sw)}$ are written in the columns of the file: the entries in the first p columns of the file are those in the first column (or row) of $\Sigma_h^{(sw)}$, while the entries in the last p columns of the file are those in the last column (or row) of $\Sigma_h^{(sw)}$.

`SigmahI.txt` this file contains samples from the posteriors of the $p \times p$ precision matrices $\Sigma_h^{-1}, h = 1, 2, \ldots, ncomp$. The file is arranged in `((sweeps-burn)/thin)*ncomp` lines and p^2 columns. In more detail, each sweep creates `ncomp` lines representing samples $(\Sigma_h^{-1})^{(sw)}, h = 1, \ldots, ncomp$, where superscript sw represents a particular sweep. The elements of $(\Sigma_h^{-1})^{(sw)}$ are written in the columns of the file: the entries in the first p columns of the file are those in the first column (or row) of $(\Sigma_h^{-1})^{(sw)}$, while the entries in the last p columns of the file are those in the last column (or row) of $(\Sigma_h^{-1})^{(sw)}$.

`Th.txt` this file contains samples from the posteriors of the $p \times p$ precision matrices $T_h, h = 1, 2, \ldots, ncomp$. The file is arranged in `((sweeps-burn)/thin)*ncomp`

lines and p^2 columns. In more detail, each sweep creates ncomp lines representing samples $T_h^{(sw)}, h = 1, \ldots, ncomp$, where superscript sw represents a particular sweep. The elements of $T_h^{(sw)}$ are written in the columns of the file: the entries in the first p columns of the file are those in the first column (or row) of $T_h^{(sw)}$, while the entries in the last p columns of the file are those in the last column (or row) of $T_h^{(sw)}$.

`xih.txt` this file contains samples from the posteriors of parameters $\xi_h, h = 1, 2, \ldots, ncomp$. The file is arranged in ((sweeps-burn)/thin)*ncomp lines and one or two columns, depending on the number of parameters in the selected $F(.; \lambda)$. Sweeps write in the file ncomp lines representing samples $\xi_h^{(sw)}, h = 1, \ldots, ncomp$, where superscript sw represents a particular sweep.

`Updated.txt` this file contains (sweeps-burn)/thin lines with the number of components updated at each iteration of the sampler.

Author(s)

Georgios Papageorgiou <gpapageo@gmail.com>

References

Consul, P. C. & Famoye, G. C. (1992). Generalized Poisson regression model. Communications in Statistics - Theory and Methods, 1992, 89-109.

McCulloch, R. E., Polson, N. G., & Rossi, P. E. (2000). A Bayesian analysis of the multinomial probit model with fully identified parameters. Journal of Econometrics, 99(1), 173-193.

Papageorgiou, G., Richardson, S. and Best, N. (2014). Bayesian nonparametric models for spatially indexed data of mixed type.

Papaspiliopoulos, O. (2008). A note on posterior sampling from Dirichlet mixture models. Technical report, University of Warwick.

Rodriguez-Avi, J., Conde-Sanchez, A., Saez-Castillo, A. J., & Olmo-Jimenez, M. J. (2004). A triparametric discrete distribution with complex parameters. Statistical Papers, 45(1), 81-95.

Saez-Castillo, A. & Conde-Sanchez, A. (2013). A hyper-poisson regression model for overdispersed and underdispersed count data. Computational Statistics & Data Analysis, 61, 148-157.

Sellers, K. F. & Shmueli, G. (2010). A flexible regression model for count data. Annals of Applied Statistics, 4(2), 943-961.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. Statistica Sinica, 4, 639-650.

Shmueli, G., Minka, T. P., Kadane, J. B., Borle, S., & Boatwright, P. (2005). A useful distribution for fitting discrete data: revival of the conwaymaxwellpoisson distribution. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(1), 127-142.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. Communications in Statistics Simulation and Computation, 36(1), 45-54.

Examples

```
# Bayesian nonparametric GLM with Binomial response Y and one predictor X
data(simD)
```

```
pred<-seq(with(simD,min(X))+0.1,with(simD,max(X))-0.1,length.out=30)
npred<-length(pred)
# fit1 and fit2 define the same model but with different numbers of
# components and posterior samples. They both use a slice sampler
# and parameter prec=200 achieves optimal acceptance rate, about 22%.
fit1 <- bnpglm(cbind(Y,(E-Y))~X, family="binomial", data=simD, ncomp=30, sweeps=150,
           burn=100, sampler="slice", prec=c(200), Xpred=pred, offsetPred=rep(30,npred))
fit2 <- bnpglm(cbind(Y,(E-Y))~X, family="binomial", data=simD, ncomp=50, sweeps=5000,
           burn=1000, sampler="slice", prec=c(200), Xpred=pred, offsetPred=rep(30,npred))
plot(with(simD,X),with(simD,Y)/with(simD,E))
lines(pred,fit2$medianReg,col=3,lwd=2)
```

`dnorm.pois` *Bivariate Normal-Poisson distribution*

Description

Normal-Poisson probability density function.

Usage

```
dnorm.pois(x,y,mu,Sigma,rate,E)
```

Arguments

`x`	Real valued sample from a univariate Normal distribution.
`y`	Integer valued sample from a univariate Poisson distribution.
`mu`	Two-dimentional vector with means, denoted as μ below.
`Sigma`	Two-by-two covariance matrix, denoted as Σ below.
`rate`	Mean of the Poisson variable, denoted as λ below.
`E`	Offset term, denoted as E below.

Value

Function `dnorm.pois` returns the joint probability density function of correlated Normal and Poisson random variables

$$f(x, y|\theta) = \int_{c_{y-1}}^{c_y} N(x, y^*|\mu, \Sigma) dy^*,$$

where y^* denotes a continuous random variable that determines the count according to the rule

$$Y = y \iff c_{y-1} < y^* < c_y.$$

Cut-points c_y are defined by

$$c_y = \Phi^{-1}(F(y; E\lambda)),$$

where $\Phi()$ is the Normal cdf, $F()$ the Poisson cdf, E denotes an offset term and λ is the mean of the Poisson. Further, μ and Σ denote the mean vector and covariance matrix of (x, y^*).

The integral is evaluated using the univariate Normal cdf

$$f(y, x|\theta) = N(x|\mu_1, \sqrt{\Sigma}_{11})\{\Phi(\frac{c_y - E(y^*|x)}{sd(y^*|x)}) - \Phi(\frac{c_{y-1} - E(y^*|x)}{sd(y^*|x)})\},$$

where $E(y^*|x) = \mu_2 + \Sigma_{12}(x - \mu_1)/\Sigma_{11}$ and $sd(y^*|x) = \sqrt{\Sigma_{22} - \Sigma_{12}^2/\Sigma_{11}}$.

Note

The mean μ_2 and variannce Σ_{22} of y^* are usually set to zero and one.

Author(s)

Georgios Papageorgiou <gpapageo@gmail.com>

Examples

```
#When the covariance matrix is diagonal dnorm.pois is equal to the product of dnorm and dpois
mu<-c(0,0)
cov.mat<-matrix(c(1,0.0,0.0,1),ncol=2,nrow=2)
dnorm.pois(0,5,mu=mu,Sigma=cov.mat,rate=3,E=2)
dnorm(0,0,1)*dpois(5,6)
#Otherwise not equal
mu<-c(0,0)
cov.mat<-matrix(c(1,-0.8,-0.8,1),ncol=2,nrow=2)
dnorm.pois(0,5,mu=mu,Sigma=cov.mat,rate=3,E=2)
```

---

simD *Simulated dataset*

Description

Just a simulated dataset to illustrate the model. The success probability and the covariate have a non-linear relationship.

Usage

```
data(simD)
```

Format

A data frame with 300 independent observations. Three numerical vectors contain information on

`Y` number of successes.

`E` number of trials.

`X` explanatory variable.

Index