Package 'glmmPen'

January 18, 2024

Type Package

Title High Dimensional Penalized Generalized Linear Mixed Models (pGLMM)

Version 1.5.4.4

Date 2024-01-18

Description Fits high dimensional penalized generalized linear mixed models using the Monte Carlo Expectation Conditional Minimization (MCECM) algorithm. The purpose of the package is to perform variable selection on both the fixed and random effects simultaneously for generalized linear mixed models. The package supports fitting of Binomial, Gaussian, and Poisson data with canonical links, and supports penalization using the MCP, SCAD, or LASSO penalties. The MCECM algorithm is described in Rashid et al. (2020) <doi:10.1080/01621459.2019.1671197>. The techniques used in the minimization portion of the procedure (the M-step) are derived from the procedures of the 'ncvreg' package (Breheny and Huang (2011) <doi:10.1214/10-AOAS388>) and 'grpreg' package (Breheny and Huang (2015) <doi:10.1007/s11222-013-9424-2>), with appropriate modifications to account for the estimation and penalization of the random effects. The 'ncvreg' and 'grpreg' packages also describe the MCP, SCAD, and LASSO penalties.

License GPL (>= 2)

Encoding UTF-8

Imports ggplot2, Matrix, methods, ncvreg, reshape2, rstan (>= 2.18.1), stringr, mvtnorm, MASS, survival, rstantools, RcppParallel (>= 5.0.1)

Depends lme4, bigmemory, Rcpp (>= 0.12.0), R (>= 3.6.0)

LinkingTo BH (>= 1.66.0), bigmemory, Rcpp (>= 0.12.0), RcppArmadillo, RcppEigen (>= 0.3.3.3.0), rstan (>= 2.18.1), StanHeaders (>= 2.18.0), RcppParallel (>= 5.0.1)

RoxygenNote 7.1.2

NeedsCompilation yes

Author Hillary Heiling [aut, cre],
Naim Rashid [aut],
Quefeng Li [aut],
Joseph Ibrahim [aut]

Suggests testthat, knitr, rmarkdown

Biarch true

SystemRequirements GNU make

Maintainer Hillary Heiling <hmheiling@gmail.com>

Repository CRAN

Date/Publication 2024-01-18 17:40:02 UTC

R topics documented:

`adaptControl` *Control of Metropolis-within-Gibbs Adaptive Random Walk Sampling Procedure Controls the adaptive random walk Metropolis-within-Gibbs sampling procedure.*

Description

Control of Metropolis-within-Gibbs Adaptive Random Walk Sampling Procedure

Controls the adaptive random walk Metropolis-within-Gibbs sampling procedure.

Usage

```
adaptControl(batch_length = 100, offset = 0)
```

Arguments

batch_length	positive integer specifying the number of posterior samples to collect before the proposal variance is adjusted based on the acceptance rate of the last batch_length accepted posterior samples. Default is set to 100. Batch length restricted to be no less than 50.
offset	non-negative integer specifying an offset value for the increment of the proposal variance adjustment. Optionally used to ensure the required diminishing adaptation condition. Default set to 0. increment = min(0.01, 1 / sqrt(batch*batch_length + offset))

Value

Function returns a list (inheriting from class "adaptControl") containing parameter specifications for the adaptive random walk sampling procedure.

---

basal

Basal dataset: A composition of cancer datasets with top scoring pairs (TSPs) as covariates and binary response indicating if the subject's cancer subtype was basal-like. A dataset composed of four datasets combined from studies that contain gene expression data from subjects with several types of cancer. Two of these datasets contain gene expression data for subjects with Pancreatic Ductal Adenocarcinoma (PDAC), one dataset contains data for subjects with Breast Cancer, and the fourth dataset contains data for subjects with Bladder Cancer. The response of interest is whether or not the subject's cancer subtype was the basal-like subtype. See articles Rashid et al. (2020) "Modeling Between-Study Heterogeneity for Improved Replicability in Gene Signature Selection and Clinical Prediction" and Moffitt et al. (2015) "Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma" for further details on these four datasets.

Description

Basal dataset: A composition of cancer datasets with top scoring pairs (TSPs) as covariates and binary response indicating if the subject's cancer subtype was basal-like.

A dataset composed of four datasets combined from studies that contain gene expression data from subjects with several types of cancer. Two of these datasets contain gene expression data for subjects with Pancreatic Ductal Adenocarcinoma (PDAC), one dataset contains data for subjects with Breast Cancer, and the fourth dataset contains data for subjects with Bladder Cancer. The response of interest is whether or not the subject's cancer subtype was the basal-like subtype. See articles Rashid et al. (2020) "Modeling Between-Study Heterogeneity for Improved Replicability in Gene

Signature Selection and Clinical Prediction" and Moffitt et al. (2015) "Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma" for further details on these four datasets.

Usage

```
data("basal")
```

Format

A list containing the following elements:

y binary response vector; 1 indicates that the subject's cancer was of the basal-like subtype, 0 otherwise

X matrix of 50 top scoring pair (TSP) covariates

group factor indicating which cancer study the observation belongs to, which are given the following descriptions: UNC PDAC, TCGA PDAC, TCGA Bladder Cancer, and UNC Breast Cancer

Z model matrix for random effects; organized first by variable, then by group (i.e. by cancer study)

---

glmm *Fit a Generalized Mixed Model via Monte Carlo Expectation Conditional Minimization (MCECM)*

Description

`glmm` is used to fit a single generalized mixed model via Monte Carlo Expectation Conditional Minimization (MCECM). Unlike `glmmPen`, no model selection is performed.

Usage

```
glmm(
  formula,
  data = NULL,
  family = "binomial",
  covar = NULL,
  offset = NULL,
  optim_options = optimControl(),
  adapt_RW_options = adaptControl(),
  trace = 0,
  tuning_options = lambdaControl(),
  progress = TRUE,
  ...
)
```

Arguments

formula	a two-sided linear formula object describing both the fixed effects and random effects part of the model, with the response on the left of a ~ operator and the terms, separated by + operators, on the right. Random-effects terms are distinguished by vertical bars ("|") separating expression for design matrices from the grouping factor. formula should be of the same format needed for glmer in package **lme4**. Only one grouping factor will be recognized. The random effects covariates need to be a subset of the fixed effects covariates. The offset must be specified outside of the formula in the 'offset' argument.
data	an optional data frame containing the variables named in formula. If data is omitted, variables will be taken from the environment of formula.
family	a description of the error distribution and link function to be used in the model. Currently, the glmmPen algorithm allows the Binomial ("binomial" or binomial()), Gaussian ("gaussian" or gaussian()), and Poisson ("poisson" or poisson()) families with canonical links only. See phmmPen for variable selection within proportional hazards mixed models for survival data.
covar	character string specifying whether the covariance matrix should be unstructured ("unstructured") or diagonal with no covariances between variables ("independent"). Default is set to NULL. If covar is set to NULL and the number of random effects predictors (not including the intercept) is greater than or equal to 10 (i.e. high dimensional), then the algorithm automatically assumes an independent covariance structure and covar is set to "independent". Otherwise if covar is set to NULL and the number of random effects predictors is less than 10, then the algorithm automatically assumes an unstructured covariance structure and covar is set to "unstructured".
offset	This can be used to specify an *a priori* known component to be included in the linear predictor during fitting. Default set to NULL (no offset). If the data argument is not NULL, this should be a numeric vector of length equal to the number of cases (the length of the response vector). If the data argument specifies a data.frame, the offset argument should specify the name of a column in the data.frame.
optim_options	a structure of class "optimControl" created from function optimControl that specifies several optimization parameters. See the documentation for optimControl for more details on defaults.
adapt_RW_options	a list of class "adaptControl" from function adaptControl containing the control parameters for the adaptive random walk Metropolis-within-Gibbs procedure. Ignored if optimControl parameter sampler is set to "stan" (default) or "independence".
trace	an integer specifying print output to include as function runs. Default value is 0. See Details for more information about output provided when trace = 0, 1, or 2.
tuning_options	a list of class "selectControl" or "lambdaControl" resulting from selectControl or lambdaControl containing additional control parameters. When function glmm is used,the algorithm may be run using one specific set of penalty parameters lambda0 and lambda1 by specifying such values in lambdaControl(). The default for glmm is to run the model fit with no penalization (lambda0 =

lambda1 = 0). When function glmmPen is run, tuning_options is specified using selectControl(). See the lambdaControl and selectControl documentation for further details.

progress — a logical value indicating if additional output should be given showing the progress of the fit procedure. If TRUE, such output includes iteration-level information for the fit procedure (iteration number EM_iter, number of MCMC samples nMC, average Euclidean distance between current coefficients and coefficients from t–defined in optimControl–iterations back EM_conv, and number of non-zero fixed and random effects covariates not including the intercept). Additionally, progress = TRUE gives some other information regarding the progress of the variable selection procedure, including the model selection criteria and log-likelihood estimates for each model fit. Default is TRUE.

... — additional arguments that could be passed into glmmPen. See glmmPen for further details.

Details

The glmm function can be used to fit a single generalized mixed model. While this approach is meant to be used in the case where the user knows which covariates belong in the fixed and random effects and no penalization is required, one is allowed to specify non-zero fixed and random effects penalties using lambdaControl and the (...) arguments. The (...) allow for specification of penalty-related arguments; see glmmPen for details. For a high dimensional situation, the user may want to fit a minimal penalty model using a small penalty for the fixed and random effects and save the posterior samples from this minimal penalty model for use in any BIC-ICQ calculations during selection within glmmPen. Specifying a file name in the 'BICq_posterior' argument will save the posterior samples from the glmm model into a big.matrix with this file name, see the Details section of glmmPen for additional details.

Value

A reference class object of class pglmmObj for which many methods are available (e.g. methods(class = "pglmmObj"))

---

glmmPen — *Fit Penalized Generalized Mixed Models via Monte Carlo Expectation Conditional Minimization (MCECM)*

Description

glmmPen is used to fit penalized generalized mixed models via Monte Carlo Expectation Conditional Minimization (MCECM). The purpose of the function is to perform variable selection on both the fixed and random effects simultaneously for the generalized linear mixed model. glmmPen selects the best model using BIC-type selection criteria (see selectControl documentation for further details). To improve the speed of the algorithm, consider setting var_restrictions = "fixef" within the optimControl options.

Usage

```
glmmPen(
  formula,
  data = NULL,
  family = "binomial",
  covar = NULL,
  offset = NULL,
  fixef_noPen = NULL,
  penalty = c("MCP", "SCAD", "lasso"),
  alpha = 1,
  gamma_penalty = switch(penalty[1], SCAD = 4, 3),
  optim_options = optimControl(),
  adapt_RW_options = adaptControl(),
  trace = 0,
  tuning_options = selectControl(),
  BICq_posterior = NULL,
  progress = TRUE,
  ...
)
```

Arguments

formula	a two-sided linear formula object describing both the fixed effects and random effects part of the model, with the response on the left of a ~ operator and the terms, separated by + operators, on the right. Random-effects terms are distinguished by vertical bars ("|") separating expression for design matrices from the grouping factor. formula should be of the same format needed for glmer in package **lme4**. Only one grouping factor will be recognized. The random effects covariates need to be a subset of the fixed effects covariates. The offset must be specified outside of the formula in the 'offset' argument.
data	an optional data frame containing the variables named in formula. If data is omitted, variables will be taken from the environment of formula.
family	a description of the error distribution and link function to be used in the model. Currently, the glmmPen algorithm allows the Binomial ("binomial" or binomial()), Gaussian ("gaussian" or gaussian()), and Poisson ("poisson" or poisson()) families with canonical links only. See phmmPen for variable selection within proportional hazards mixed models for survival data.
covar	character string specifying whether the covariance matrix should be unstructured ("unstructured") or diagonal with no covariances between variables ("independent"). Default is set to NULL. If covar is set to NULL and the number of random effects predictors (not including the intercept) is greater than or equal to 10 (i.e. high dimensional), then the algorithm automatically assumes an independent covariance structure and covar is set to "independent". Otherwise if covar is set to NULL and the number of random effects predictors is less than 10, then the algorithm automatically assumes an unstructured covariance structure and covar is set to "unstructured".

`offset` This can be used to specify an *a priori* known component to be included in the linear predictor during fitting. Default set to `NULL` (no offset). If the data argument is not `NULL`, this should be a numeric vector of length equal to the number of cases (the length of the response vector). If the data argument specifies a data.frame, the offset argument should specify the name of a column in the data.frame.

`fixef_noPen` Optional vector of 0's and 1's of the same length as the number of fixed effects covariates used in the model. Value 0 indicates the variable should not have its fixed effect coefficient penalized, 1 indicates that it can be penalized. Order should correspond to the same order of the fixed effects given in the formula.

`penalty` character describing the type of penalty to use in the variable selection procedure. Options include 'MCP', 'SCAD', and 'lasso'. Default is MCP penalty. If the random effect covariance matrix is "unstructured", then a group MCP, group SCAD, or group LASSO penalty is used on the random effects coefficients. See Breheny and Huang (2011) <doi:10.1214/10-AOAS388> and Breheny and Huang (2015) <doi:10.1007/s11222-013-9424-2> for details of these penalties.

`alpha` Tuning parameter for the Mnet estimator which controls the relative contributions from the MCP/SCAD/LASSO penalty and the ridge, or L2, penalty. `alpha=1` is equivalent to the MCP/SCAD/LASSO penalty, while `alpha=0` is equivalent to ridge regression. However, `alpha=0` is not supported; `alpha` may be arbitrarily small, but not exactly zero

`gamma_penalty` The scaling factor of the MCP and SCAD penalties. Not used by LASSO penalty. Default is 4.0 for SCAD and 3.0 for MCP. See Breheny and Huang (2011) <doi:10.1214/10-AOAS388> and Breheny and Huang (2015) <doi:10.1007/s11222-013-9424-2> for further details.

`optim_options` a structure of class "optimControl" created from function `optimControl` that specifies several optimization parameters. See the documentation for `optimControl` for more details on defaults.

`adapt_RW_options` a list of class "adaptControl" from function `adaptControl` containing the control parameters for the adaptive random walk Metropolis-within-Gibbs procedure. Ignored if `optimControl` parameter `sampler` is set to "stan" (default) or "independence".

`trace` an integer specifying print output to include as function runs. Default value is 0. See Details for more information about output provided when trace = 0, 1, or 2.

`tuning_options` a list of class "selectControl" or "lambdaControl" resulting from `selectControl` or `lambdaControl` containing additional control parameters. When function `glmm` is used,the algorithm may be run using one specific set of penalty parameters `lambda0` and `lambda1` by specifying such values in `lambdaControl()`. The default for `glmm` is to run the model fit with no penalization (`lambda0 = lambda1 = 0`). When function `glmmPen` is run, `tuning_options` is specified using `selectControl()`. See the `lambdaControl` and `selectControl` documentation for further details.

`BICq_posterior` an optional character string expressing the path and file basename of a file combination that will file-back or currently file-backs a `big.matrix` of the posterior samples from the minimal penalty model used for the BIC-ICQ calculation used for model selection. T (BIC-ICQ reference: Ibrahim et al (2011)

<doi:10.1111/j.1541-0420.2010.01463.x>). If this argument is specified as NULL (default) and BIC-ICQ calculations are requested (see selectControl) for details), the posterior samples will be saved in the file combination 'BICq_Posterior_Draws.bin' and 'BICq_Posterior_Draws.desc' in the working directory. See 'Details' section for additional details about the required format of BICq_posterior and the file-backed big matrix.

progress — a logical value indicating if additional output should be given showing the progress of the fit procedure. If TRUE, such output includes iteration-level information for the fit procedure (iteration number EM_iter, number of MCMC samples nMC, average Euclidean distance between current coefficients and coefficients from t–defined in optimControl–iterations back EM_conv, and number of non-zero fixed and random effects covariates not including the intercept). Additionally, progress = TRUE gives some other information regarding the progress of the variable selection procedure, including the model selection criteria and log-likelihood estimates for each model fit. Default is TRUE.

... — additional arguments that could be passed into glmmPen. See glmmPen_FA and phmmPen_FA for further details.

Details

Argument BICq_posterior details: If the BIC_option in selectControl (tuning_options) is specified to be 'BICq', this requests the calculation of the BIC-ICQ criterion during the selection process. For the BIC-ICQ criterion to be calculated, a minimal penalty model assuming a small valued lambda penalty needs to be fit, and the posterior samples from this minimal penalty model need to be used. In order to avoid repetitive calculations of this minimal penalty model (i.e. if the user wants to re-run glmmPen with a different set of penalty parameters), a big.matrix of these posterior samples will be file-backed as two files: a backing file with extension '.bin' and a descriptor file with extension '.desc'. The BICq_posterior argument should contain a path and a filename of the form "./path/filename" such that the backingfile and the descriptor file would then be saved as "./path/filename.bin" and "./path/filename.desc", respectively. If BICq_posterior is set to NULL, then by default, the backingfile and descriptor file are saved in the working directory as "BICq_Posterior_Draws.bin" and "BICq_Posterior_Draws.desc". If the big matrix of posterior samples is already file-backed, BICq_posterior should specify the path and basename of the appropriate files (again of form "./path/filename"); the minimal penalty model will not be fit again and the big.matrix of posterior samples will be read using the attach.big.matrix function of the bigmemory package and used in the BIC-ICQ calculations. If the appropriate files do not exist or BICq_posterior is specified as NULL, the minimal penalty model will be fit and the minimal penalty model posterior samples will be saved as specified above. The algorithm will save 10^4 posterior samples automatically.

Trace details: The value of 0 (default) does not output any extra information. The value of 1 additionally outputs the updated coefficients, updated covariance matrix values, and the number of coordinate descent iterations used for the M step for each EM iteration. When pre-screening procedure is used and/or if the BIC-ICQ criterion is requested, trace = 1 gives additional information about the penalties used for the 'minimal penalty model' fit procedure. If Stan is not used as the E-step sampling mechanism, the value of 2 outputs all of the above plus gibbs acceptance rate information for the adaptive random walk and independence samplers and the updated proposal standard deviation for the adaptive random walk.

Value

A reference class object of class pglmmObj for which many methods are available (e.g. methods(class = "pglmmObj"), see ?pglmmObj for additional documentation)

---

glmmPen_FA *Fit Penalized Generalized Mixed Models via Monte Carlo Expectation Conditional Minimization (MCECM)*

Description

glmmPen_FA is used to fit penalized generalized linear mixed models via Monte Carlo Expectation Conditional Minimization (MCECM) using a factor model decomposition of the random effects. The purpose of the function is to perform variable selection on both the fixed and random effects simultaneously for the generalized linear mixed model. This function uses a factor model decomposition on the random effects. This assumption reduces the latent space in the E-step (Expectation step) of the algorithm, which reduces the computational complexity of the algorithm. This improves the speed of the algorithm and enables the scaling of the algorithm to larger dimensions. Besides the modeling of the random effects, this function is similar to glmmPen. glmmPen_FA selects the best model using BIC-type selection criteria (see selectControl documentation for further details). Function is currently available for Binomial and Poisson families with canonical links. To improve the speed of the algorithm, consider setting var_restrictions = "fixef" within the optimControl options.

Usage

```
glmmPen_FA(
  formula,
  data = NULL,
  family = "binomial",
  offset = NULL,
  r_estimation = rControl(),
  fixef_noPen = NULL,
  penalty = c("MCP", "SCAD", "lasso"),
  alpha = 1,
  gamma_penalty = switch(penalty[1], SCAD = 4, 3),
  optim_options = optimControl(),
  adapt_RW_options = adaptControl(),
  trace = 0,
  tuning_options = selectControl(),
  BICq_posterior = NULL,
  progress = TRUE,
  ...
)
```

Arguments

formula	a two-sided linear formula object describing both the fixed effects and random effects part of the model, with the response on the left of a ~ operator and the terms, separated by + operators, on the right. Random-effects terms are distinguished by vertical bars ("|") separating expression for design matrices from the grouping factor. formula should be of the same format needed for glmer in package **lme4**. Only one grouping factor will be recognized. The random effects covariates need to be a subset of the fixed effects covariates. The offset must be specified outside of the formula in the 'offset' argument.
data	an optional data frame containing the variables named in formula. If data is omitted, variables will be taken from the environment of formula.
family	a description of the error distribution and link function to be used in the model. Currently, the glmmPen algorithm allows the Binomial ("binomial" or binomial()), Gaussian ("gaussian" or gaussian()), and Poisson ("poisson" or poisson()) families with canonical links only. See phmmPen for variable selection within proportional hazards mixed models for survival data.
offset	This can be used to specify an *a priori* known component to be included in the linear predictor during fitting. Default set to NULL (no offset). If the data argument is not NULL, this should be a numeric vector of length equal to the number of cases (the length of the response vector). If the data argument specifies a data.frame, the offset argument should specify the name of a column in the data.frame.
r_estimation	a list of class "rControl" from function rControl containing the control parameters for the estimation of the number of latent factors to use in the glmmPen_FA and glmm_FA estimation procedures.
fixef_noPen	Optional vector of 0's and 1's of the same length as the number of fixed effects covariates used in the model. Value 0 indicates the variable should not have its fixed effect coefficient penalized, 1 indicates that it can be penalized. Order should correspond to the same order of the fixed effects given in the formula.
penalty	character describing the type of penalty to use in the variable selection procedure. Options include 'MCP', 'SCAD', and 'lasso'. Default is MCP penalty. If the random effect covariance matrix is "unstructured", then a group MCP, group SCAD, or group LASSO penalty is used on the random effects coefficients. See Breheny and Huang (2011) <doi:10.1214/10-AOAS388> and Breheny and Huang (2015) <doi:10.1007/s11222-013-9424-2> for details of these penalties.
alpha	Tuning parameter for the Mnet estimator which controls the relative contributions from the MCP/SCAD/LASSO penalty and the ridge, or L2, penalty. alpha=1 is equivalent to the MCP/SCAD/LASSO penalty, while alpha=0 is equivalent to ridge regression. However, alpha=0 is not supported; alpha may be arbitrarily small, but not exactly zero
gamma_penalty	The scaling factor of the MCP and SCAD penalties. Not used by LASSO penalty. Default is 4.0 for SCAD and 3.0 for MCP. See Breheny and Huang (2011) <doi:10.1214/10-AOAS388> and Breheny and Huang (2015) <doi:10.1007/s11222-013-9424-2> for further details.

`optim_options` a structure of class "optimControl" created from function `optimControl` that specifies several optimization parameters. See the documentation for `optimControl` for more details on defaults.

`adapt_RW_options` a list of class "adaptControl" from function `adaptControl` containing the control parameters for the adaptive random walk Metropolis-within-Gibbs procedure. Ignored if `optimControl` parameter `sampler` is set to "stan" (default) or "independence".

`trace` an integer specifying print output to include as function runs. Default value is 0. See Details for more information about output provided when trace = 0, 1, or 2.

`tuning_options` a list of class "selectControl" or "lambdaControl" resulting from `selectControl` or `lambdaControl` containing additional control parameters. When function `glmm` is used,the algorithm may be run using one specific set of penalty parameters `lambda0` and `lambda1` by specifying such values in `lambdaControl()`. The default for `glmm` is to run the model fit with no penalization (`lambda0` = `lambda1` = 0). When function `glmmPen` is run, `tuning_options` is specified using `selectControl()`. See the `lambdaControl` and `selectControl` documentation for further details.

`BICq_posterior` an optional character string expressing the path and file basename of a file combination that will file-back or currently file-backs a `big.matrix` of the posterior samples from the minimal penalty model used for the BIC-ICQ calculation used for model selection. T (BIC-ICQ reference: Ibrahim et al (2011) <doi:10.1111/j.1541-0420.2010.01463.x>). If this argument is specified as `NULL` (default) and BIC-ICQ calculations are requested (see `selectControl`) for details), the posterior samples will be saved in the file combination 'BICq_Posterior_Draws.bin' and 'BICq_Posterior_Draws.desc' in the working directory. See 'Details' section for additional details about the required format of `BICq_posterior` and the file-backed big matrix.

`progress` a logical value indicating if additional output should be given showing the progress of the fit procedure. If `TRUE`, such output includes iteration-level information for the fit procedure (iteration number EM_iter, number of MCMC samples nMC, average Euclidean distance between current coefficients and coefficients from t–defined in `optimControl`–iterations back EM_conv, and number of non-zero fixed and random effects covariates not including the intercept). Additionally, `progress = TRUE` gives some other information regarding the progress of the variable selection procedure, including the model selection criteria and log-likelihood estimates for each model fit. Default is `TRUE`.

`...` additional arguments that could be passed into `glmmPen_FA`. See `phmmPen_FA` for further details (e.g. `survival_options` argument).

Details

Argument `BICq_posterior` details: If the `BIC_option` in `selectControl` (`tuning_options`) is specified to be 'BICq', this requests the calculation of the BIC-ICQ criterion during the selection process. For the BIC-ICQ criterion to be calculated, a full model assuming a small valued lambda penalty needs to be fit, and the posterior draws from this full model need to be used. In order to avoid repetitive calculations of this full model (i.e. if the user wants to re-run `glmmPen` with a different

set of penalty parameters), a `big.matrix` of these posterior draws will be file-backed as two files: a backing file with extention '.bin' and a descriptor file with extension '.desc'. The `BICq_posterior` argument should contain a path and a filename with no extension of the form "./path/filename" such that the backingfile and the descriptor file would then be saved as "./path/filename.bin" and "./path/filename.desc", respectively. If `BICq_posterior` is set to `NULL`, then by default, the backingfile and descriptor file are saved in the working directory as "BICq_Posterior_Draws.bin" and "BICq_Posterior_Draws.desc". If the big matrix of posterior draws is already file-backed, `BICq_posterior` should specify the path and basename of the appropriate files (again of form "./path/filename"); the full model will not be fit again and the big.matrix of posterior draws will be read using the `attach.big.matrix` function of the `bigmemory` package and used in the BIC-ICQ calcuations. If the appropriate files do not exist or `BICq_posterior` is specified as `NULL`, the full model will be fit and the full model posterior draws will be saved as specified above. The algorithm will save 10^4 posterior draws automatically.

Trace details: The value of 0 (default) does not output any extra information. The value of 1 additionally outputs the updated coefficients, updated covariance matrix values, and the number of coordinate descent iterations used for the M step for each EM iteration. When pre-screening procedure is used and/or if the BIC-ICQ criterion is requested, trace = 1 gives additional information about the penalties used for the 'full model' fit procedure. If Stan is not used as the E-step sampling mechanism, the value of 2 outputs all of the above plus gibbs acceptance rate information for the adaptive random walk and independence samplers and the updated proposal standard deviation for the adaptive random walk.

Value

A reference class object of class `pglmmObj` for which many methods are available (e.g. `methods(class = "pglmmObj")`, see ?pglmmObj for additional documentation)

`glmm_FA` *Fit a Generalized Mixed Model via Monte Carlo Expectation Conditional Minimization (MCECM)*

Description

`glmm_FA` is used to fit a single generalized linear mixed model via Monte Carlo Expectation Conditional Minimization (MCECM) using a factor model decomposition of the random effects. No model selection is performed. This function uses a factor model decomposition on the random effects. This assumption reduces the latent space in the E-step (Expectation step) of the algorithm, which reduces the computational complexity of the algorithm. This improves the speed of the algorithm and enables the scaling of the algorithm to larger dimensions. Besides the modeling of the random effects, this function is similar to `glmm`.

Usage

```
glmm_FA(
  formula,
  data = NULL,
  family = "binomial",
```

```
  offset = NULL,
  r_estimation = rControl(),
  optim_options = optimControl(),
  adapt_RW_options = adaptControl(),
  trace = 0,
  tuning_options = lambdaControl(),
  progress = TRUE,
  ...
)
```

Arguments

formula	a two-sided linear formula object describing both the fixed effects and random effects part of the model, with the response on the left of a ~ operator and the terms, separated by + operators, on the right. Random-effects terms are distinguished by vertical bars ("|") separating expression for design matrices from the grouping factor. formula should be of the same format needed for glmer in package **lme4**. Only one grouping factor will be recognized. The random effects covariates need to be a subset of the fixed effects covariates. The offset must be specified outside of the formula in the 'offset' argument.
data	an optional data frame containing the variables named in formula. If data is omitted, variables will be taken from the environment of formula.
family	a description of the error distribution and link function to be used in the model. Currently, the glmmPen algorithm allows the Binomial ("binomial" or binomial()), Gaussian ("gaussian" or gaussian()), and Poisson ("poisson" or poisson()) families with canonical links only. See phmmPen for variable selection within proportional hazards mixed models for survival data.
offset	This can be used to specify an *a priori* known component to be included in the linear predictor during fitting. Default set to NULL (no offset). If the data argument is not NULL, this should be a numeric vector of length equal to the number of cases (the length of the response vector). If the data argument specifies a data.frame, the offset argument should specify the name of a column in the data.frame.
r_estimation	a list of class "rControl" from function rControl containing the control parameters for the estimation of the number of latent factors to use in the glmmPen_FA and glmm_FA estimation procedures.
optim_options	a structure of class "optimControl" created from function optimControl that specifies several optimization parameters. See the documentation for optimControl for more details on defaults.
adapt_RW_options	a list of class "adaptControl" from function adaptControl containing the control parameters for the adaptive random walk Metropolis-within-Gibbs procedure. Ignored if optimControl parameter sampler is set to "stan" (default) or "independence".
trace	an integer specifying print output to include as function runs. Default value is 0. See Details for more information about output provided when trace = 0, 1, or 2.

tuning_options — a list of class "selectControl" or "lambdaControl" resulting from selectControl or lambdaControl containing additional control parameters. When function glmm is used,the algorithm may be run using one specific set of penalty parameters lambda0 and lambda1 by specifying such values in lambdaControl(). The default for glmm is to run the model fit with no penalization (lambda0 = lambda1 = 0). When function glmmPen is run, tuning_options is specified using selectControl(). See the lambdaControl and selectControl documentation for further details.

progress — a logical value indicating if additional output should be given showing the progress of the fit procedure. If TRUE, such output includes iteration-level information for the fit procedure (iteration number EM_iter, number of MCMC samples nMC, average Euclidean distance between current coefficients and coefficients from t–defined in optimControl–iterations back EM_conv, and number of non-zero fixed and random effects covariates not including the intercept). Additionally, progress = TRUE gives some other information regarding the progress of the variable selection procedure, including the model selection criteria and log-likelihood estimates for each model fit. Default is TRUE.

... — additional arguments that could be passed into glmmPen_FA. See glmmPen_FA for further details (e.g. arguments related to variable selection that could be used to fit a single penalized GLMM model).

Details

The glmm_FA function can be used to fit a single generalized linear mixed model. While this approach is meant to be used in the case where the user knows which covariates belong in the fixed and random effects and no penalization is required, one is allowed to specify non-zero fixed and random effects penalties using lambdaControl and the (...) arguments. The (...) allow for specification of penalty-related arguments; see glmmPen_FA for details. For a high dimensional situation, the user may want to fit a full model using a small penalty for the fixed and random effects and save the posterior draws from this full model for use in any BIC-ICQ calculations during selection within glmmPen_FA. Specifying a file name in the 'BICq_posterior' argument will save the posterior draws from the glmm_FA model into a big.matrix with this file name, see the Details section of glmmPen_FA for additional details.

Value

A reference class object of class pglmmObj for which many methods are available (e.g. methods(class = "pglmmObj"))

---

lambdaControl *Control of Penalization Parameters and Selection Criteria*

Description

Constructs control structures for penalized mixed model fitting.

Usage

```
lambdaControl(lambda0 = 0, lambda1 = 0)

selectControl(
  lambda0_seq = NULL,
  lambda1_seq = NULL,
  nlambda = 10,
  search = c("abbrev", "full_grid"),
  BIC_option = c("BICq", "BICh", "BIC", "BICNgrp"),
  logLik_calc = switch(BIC_option[1], BICq = FALSE, TRUE),
  lambda.min = NULL,
  pre_screen = TRUE,
  lambda.min.presc = NULL
)
```

Arguments

`lambda0`	a non-negative numeric penalty parameter for the fixed effects coefficients
`lambda1`	a non-negative numeric penalty parameter for the (grouped) random effects covariance coefficients
`lambda0_seq, lambda1_seq`	a sequence of non-negative numeric penalty parameters for the fixed and random effect coefficients (`lambda0_seq` and `lambda1_seq`, respectively). If `NULL`, then a sequence will be automatically calculated. See 'Details' section for more details on these default calculations.
`nlambda`	positive integer specifying number of penalty parameters to use for the fixed and random effects penalty parameters. Default set to 10. Only used if either `lambda0_seq` or `lambda1_seq` remain unspecified by the user (one or both of these arguments set to `NULL`) and, consequently, one or more default sequences need to be calculated.
`search`	character string of "abbrev" (default) or "full_grid" indicating if the search of models over the penalty parameter space should be the full grid search (total number of models equals ‘nlambda‘^2 or length(‘lambda0_seq‘)*length(‘lambda1_seq‘)) or an abbreviated grid search. The abbreviated grid search is described in more detail in the Details section. Te authors highly recommend the abbreviated grid search.
`BIC_option`	character string specifying the selection criteria used to select the 'best' model. Default "BICq" option specifies the BIC-ICQ criterion (Ibrahim et al (2011) <doi:10.1111/j.1541-0420.2010.01463.x>), which requires a fit of a 'minimum penalty' model; a small penalty (the minimum of the penalty sequence) is used for the fixed and random effects. See "Details" section for what these small penalties will be. The "BICh" option utilizes the hybrid BIC value described in Delattre, Lavielle, and Poursat (2014) <doi:10.1214/14-EJS890>. The regular "BIC" option penalty term uses (total non-zero coefficients)*(length(y) = total number observations). The "BICNgrp" option penalty term uses (total non-zero coefficients)*(nlevels(group) = number groups).

`logLik_calc` logical value specifying if the log likelihood (and log-likelihood based calculations BIC, BICh, and BICNgrp) should be calculated for all of the models in the selection procedure. If BIC-ICQ is used for selection, the log-likelihood is not needed for each model. However, if users are interested in comparing the best models from BIC-ICQ and other BIC-type selection criteria, setting `logLik_calc` to `TRUE` will calculate these other quantities for all of the models.

`lambda.min` numeric fraction between 0 and 1. The sequence of the lambda penalty parameters ranges from the maximum lambda where all fixed and random effects are penalized to 0 and a minimum lambda value, which equals a small fraction of the maximum lambda. The parameter `lambda.min` specifies this fraction. Default value is set to `NULL`, which automatically selects `lambda.min` to equal 0.01 when the number of observations is greater than the number of fixed effects predictors and 0.05 otherwise. Only used if either `lambda0_seq` or `lambda1_seq` remain unspecified by the user (one or both of these sequence arguments set to `NULL`) and, consequently, one or more default sequences need to be calculated.

`pre_screen` logical value indicating whether pre-screening should be performed before model selection (default `TRUE`). If the number of random effects covariates considered is 4 or less, then no pre-screening will be performed. Pre-screening removes random effects from consideration during the model selection process, which can significantly speed up the algorithm. See "Details" section for a further discussion.

`lambda.min.presc` numeric fraction between 0 and 1. During pre-screening and the minimal penalty model fit for the BIC-ICQ calculation, the small penalty used on the random effect is the fraction `lambda.min.presc` multiplied by the maximum penalty parameter that penalizes all fixed and random effects to 0. If left as `NULL`, the default value is 0.01 when the number of random effect covariates is 10 or less and 0.05 otherwise. Only used if `lambda1_seq` remains unspecified by the user (this argument set to `NULL` so the random effects penalty parameter sequence needs to be automatically calculated) AND either the pre-screening procedure is selected by the argument `pre_screen` or the BIC-ICQ is selected as the model selection criteria, i.e., `BIC_option` = "BICq". See the "Details" section for a further discussion.

Details

If left as the default `NULL` values, the `lambda0_seq` and `lambda1_seq` numeric sequences are automatically calculated. The sequence will be calculated in the same manner as `ncvreg` calculates the range: the max value (let's denote this as `lambda_max`) penalizes all fixed and random effects to 0, the min value is a small portion of max (`lambda.min*lambda_max`), and the sequence is composed of `nlambda` values ranging from these min and max values spread evenly on the log scale. Unlike `ncvreg`, the order of penalty values used in the algorithm must run from the min lambda to the max lambda (as opposed to running from max lambda to min lambda). The length of the sequence is specified by `nlambda`. By default, these sequences are calculated using `LambdaSeq`.

The `lambda0` and `lambda1` arguments used within the `glmm` function allow for a user to fit a model with a single non-zero penalty parameter combination. However, this is generally not recommended.

Abbreviated grid search: The abbreviated grid search proceeds in two stages. In stage 1, the algorithm fits the following series of models: the fixed effects penalty parameter remains a fixed value evaluated at the minimum of the fixed effects penalty parameters, and all random effects penalty parameters are examined. The 'best' model from this first stage of models determines the optimum random effect penalty parameter. In stage 2, the algorithm fits the following series of models: the random effects penalty parameter remains fixed at the value of the optimum random effect penalty parameter (from stage 1) and all fixed effects penalty parameters are considered. The best overall model is the best model from stage 2. This reduces the number of models considered to length(‘lambda0_seq‘) + length(‘lambda1_seq‘). The authors found that this abbreviated grid search worked well in simulations, and performed considerably faster than the full grid search that examined all possible fixed and random effect penalty parameter combinations.

The arguments `nlambda` and `lambda.min` are only used if one or both of the `lambda0_seq` and `lambda1_seq` penalty sequences (corresponding to the fixed and random effects penalty sequences, respectively) remain unspecified by the user (one or both of these arguments left as `NULL`), indicating that the algorithm needs to calculate default penalty sequences.

The argument `lambda.min.presc` is only used under the following condition: `lambda1_seq` remains unspecified by the user (this argument set to `NULL` so the random effects penalty parameter sequence needs to be calculated) AND either the pre-screening procedure is selected by the argument `pre_screen` or the BIC-ICQ is selected as the model selection criteria, i.e., `BIC_option = "BICq"`. If `lambda1_seq` is specified by the user, the minimum value in that sequence will be used as the random effect penalty in the pre-screening procedure and/or the minimal penalty model for the BIC-ICQ calculation.

BIC-ICQ calculation: This model selection criteria requires the fitting of a 'minimal penalty' model, which fits a model with a small penalty on the fixed and random effects. For the fixed effects penalty, the minimal penalty is: (a) 0 if the number of fixed effects covariates is 4 or less or (b) the minimum fixed effect penalty from the fixed effects penalty sequence (either from the default sequence or from the sequence specified by the user). For the random effects penalty, the minimal penalty is (a) 0 if the number of random effects covariates is 4 or less; (b) the minimum random effect penalty from the random effects penalty sequence specified by the user, or (c) `lambda.min.presc` multiplied to the `lambda_max` maximum penalty specified above when a default random effects penalty sequence is calculated.

Pre-screening: The minimum fixed effects penalty used in the pre-screening stage will be the minimum penalty of the fixed effects penalty sequence, `lambda0_seq`. The minimum random effects penalty used in the pre-screening stage will be either (a) the minimum random effects penalty in the sequence `lambda1_seq` if this sequence specified by the user, or (b) `lambda.min.pres` x `lambda_max`, where `lambda_max` was described above.

Value

The *Control functions return a list (inheriting from class "`pglmmControl`") containing parameter values that determine settings for variable selection.

`LambdaSeq` *Calculation of Penalty Parameter Sequence (Lambda Sequence)*

Description

Calculates the sequence of penalty parameters used in the model selection procedure. This function calls functions from package ncvreg.

Usage

```
LambdaSeq(
  X,
  y,
  family,
  offset = NULL,
  alpha = 1,
  lambda.min = NULL,
  nlambda = 10,
  penalty.factor = NULL
)
```

Arguments

X	matrix of standardized fixed effects (see std function in ncvreg documenation). X should not include intercept.
y	numeric vector of response values. If "survival" family, y are the event indicator values (0 if censored, 1 if event)
family	a description of the error distribution and link function to be used in the model. Currently, the glmmPen algorithm allows the Binomial ("binomial" or binomial()), Gaussian ("gaussian" or gaussian()), and Poisson ("poisson" or poisson()) families with canonical links only. See phmmPen for variable selection within proportional hazards mixed models for survival data.
offset	numeric vector that 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 number of observations
alpha	Tuning parameter for the Mnet estimator which controls the relative contributions from the MCP/SCAD/LASSO penalty and the ridge, or L2, penalty. alpha=1 is equivalent to the MCP/SCAD/LASSO penalty, while alpha=0 is equivalent to ridge regression. However, alpha=0 is not supported; alpha may be arbitrarily small, but not exactly zero
lambda.min	numeric fraction between 0 and 1. The sequence of the lambda penalty parameters ranges from the maximum lambda where all fixed and random effects are penalized to 0 and a minimum lambda value, which equals a small fraction of the maximum lambda. The parameter lambda.min specifies this fraction. Default value is set to NULL, which automatically selects lambda.min to equal 0.01 when the number of observations is greater than the number of fixed effects predictors and 0.05 otherwise. Only used if either lambda0_seq or lambda1_seq remain unspecified by the user (one or both of these sequence arguments set to NULL) and, consequently, one or more default sequences need to be calculated.
nlambda	positive integer specifying number of penalty parameters (lambda) with which to fit a model.

`penalty.factor` an optional numeric vector equal to the `fixef_noPen` argument in `glmmPen`

Value

numeric sequence of penalty parameters of length `nlambda` ranging from the minimum penalty parameter (first element) equal to fraction `lambda.min` multiplied by the maximum penalty parameter to the maximum penalty parameter (last element)

`optimControl` *Control of Penalized Generalized Linear Mixed Model Fitting*

Description

Constructs the control structure for the optimization of the penalized mixed model fit algorithm.

Usage

```
optimControl(
  var_restrictions = c("none", "fixef"),
  conv_EM = 0.0015,
  conv_CD = 5e-04,
  nMC_burnin = NULL,
  nMC_start = NULL,
  nMC_max = NULL,
  nMC_report = 5000,
  maxitEM = NULL,
  maxit_CD = 50,
  M = 10000,
  t = 2,
  mcc = 2,
  sampler = c("stan", "random_walk", "independence"),
  var_start = "recommend",
  step_size = 1,
  standardization = TRUE,
  convEM_type = c("AvgEuclid1", "maxdiff", "AvgEuclid2", "Qfun"),
  B_init_type = c("deterministic", "data", "random")
)
```

Arguments

`var_restrictions`

character string indicating how the random effect covariance matrix should be initialized at the beginning of the algorithm when penalties are applied to the coefficients. If "none" (default), all random effect predictors are initialized to have non-zero variances. If "fixef", the code first examines the initialized fixed effects (initialized using a regular penalized GLM), and only the random effect predictors that are initialized with non-zero fixed effects are initialized with non-zero variances.

conv_EM — a non-negative numeric convergence criteria for the convergence of the EM algorithm. Default is 0.0015. EM algorithm is considered to have converge if the average Euclidean distance between the current coefficient estimates and the coefficient estimates from t EM iterations back is less than conv_EM mcc times in a row. See t and mcc for more details.

conv_CD — a non-negative numeric convergence criteria for the convergence of the grouped coordinate descent loop within the M step of the EM algorithm. Default 0.0005.

nMC_burnin — positive integer specifying the number of posterior samples to use as burn-in for each E step in the EM algorithm. If set to NULL, the algorithm inputs the following defaults: Default 250 when the number of random effects predictors is less than or equal to 10; default 100 otherwise. Function will not allow nMC_burnin to be less than 100.

nMC_start — a positive integer for the initial number of Monte Carlo draws. If set to NULL, the algorithm inputs the following defaults: Default 250 when the number of random effects predictors is less than or equal to 10; default 100 otherwise.

nMC_max — a positive integer for the maximum number of allowed Monte Carlo draws used in each step of the EM algorithm. If set to NULL, the algorithm inputs the following defaults: When the number of random effect covariates is greater than 10, the default is set to 1000; when the number of random effect covariates is 10 or less, the default is set to 2500.

nMC_report — a positive integer for the number of posterior samples to save from the final model. These posterior samples can be used for diagnostic purposes, see plot_mcmc. Default set to 5000.

maxitEM — a positive integer for the maximum number of allowed EM iterations. If set to NULL, then the algorithm inputs the following defaults: Default equals 50 for the Binomial and Poisson families, 65 for the Gaussian family.

maxit_CD — a positive integer for the maximum number of allowed iterations for the coordinate descent algorithms used within the M-step of each EM iteration. Default equals 50.

M — positive integer specifying the number of posterior samples to use within the Pajor log-likelihood calculation. Default is 10^4; minimum allowed value is 5000.

t — the convergence criteria is based on the average Euclidean distance between the most recent coefficient estimates and the coefficient estimates from t EM iterations back. Positive integer, default equals 2.

mcc — the number of times the convergence criteria must be met before the algorithm is seen as having converged (mcc for 'meet condition counter'). Default set to 2. Value restricted to be no less than 2.

sampler — character string specifying whether the posterior samples of the random effects should be drawn using Stan (default, from package rstan) or the Metropolis-within-Gibbs procedure incorporating an adaptive random walk sampler ("random_walk") or an independence sampler ("independence"). If using the random walk sampler, see adaptControl for some additional control structure parameters.

`var_start` either the character string "recommend" or a positive number specifying the starting values to initialize the variance of the covariance matrix. For `glmmPen`, the default "recommend" first fits a simple model with a fixed and random intercept only using the lme4 package. The random intercept variance estimate from this model is then multiplied by 2 and used as the starting variance. For `glmmPen_FA`, the default is set to 0.10 (see `B_init_type` for further information).

`step_size` positive numeric value indicating the starting step size to use in the Majorization-Minimization scheme of the M-step. Only relevant when the distributional assumption used is not Binomial or Gaussian with canonical links (e.g. Poisson with log link)

`standardization`
logical value indicating whether covariates should standardized (`TRUE`, default) or unstandardized (`FALSE`) before being used within the algorithm. If `standardization = TRUE`, then the standardized covariates will also be used to create the Z matrix used in the estimation of the random effects.

`convEM_type` character string indicating the type of convergence criteria to use within the EM algorithm to determine when a model has converged. The default is "AvgEuclid1", which calculates the average Euclidean distance between the most recent coefficient vector and the coefficient vector `t` EM iterations back (Euclidean distance divided by the number of non-zero coefficients `t` EM iterations back). Alternative convergence options include "maxdiff", which determines convergence based on the maximum difference between the coefficient vectors; "AvgEuclid2", which is similar to "AvgEuclid1" except it divides the Euclidean distance by the square-root of the number of non-zero coefficients; and "Qfun", which determines convergence based on the relative difference in the Q-function estimates calculated with the most recent coefficient vector and the coefficient vector `t` EM iterations back.

`B_init_type` character string indicating how the B matrix within the `glmmPen_FA` method should be initialized. (This argument is not used within the `glmmPen` function.) The default "deterministic" initializes all non-zero variance and covariance values of the random effect covariance matrix to the value of `var_start`, such that each non-zero element of the B matrix is `sqrt(var_start / r)` (where `r` is the number of latent factors). Option "data" is similar to "deterministic", but the `var_start` value is the default data-driven variance estimate used in `glmmPen` (see argument `var_start` for more details).

Details

Several arguments are set to a default value of `NULL`. If these arguments are left as `NULL` by the user, then these values will be filled in with appropriate default values as specified above, which may depend on the number of random effects or the family of the data. If the user specifies particular values for these arguments, no additional modifications to these arguments will be done within the algorithm.

Value

Function returns a list inheriting from class `optimControl` containing fit and optimization criteria values used in optimization routine.

`pglmmObj-class` *Class* `pglmmObj` *of Fitted Penalized Generalized Mixed-Effects Models for package* `glmmPen`

Description

The functions `glmm`, `glmmPen`, `glmm_FA`, and `glmmPen_FA` from the package `glmmPen` output the reference class object of type `pglmmObj`.

Usage

```
## S3 method for class 'pglmmObj'
fixef(object, ...)

## S3 method for class 'pglmmObj'
ranef(object, ...)

## S3 method for class 'pglmmObj'
sigma(object, ...)

## S3 method for class 'pglmmObj'
coef(object, ...)

## S3 method for class 'pglmmObj'
family(object, ...)

## S3 method for class 'pglmmObj'
nobs(object, ...)

## S3 method for class 'pglmmObj'
ngrps(object, ...)

## S3 method for class 'pglmmObj'
formula(x, fixed.only = FALSE, random.only = FALSE, ...)

## S3 method for class 'pglmmObj'
model.frame(formula, fixed.only = FALSE, ...)

## S3 method for class 'pglmmObj'
model.matrix(object, type = c("fixed", "random"), ...)

## S3 method for class 'pglmmObj'
```

```
fitted(object, fixed.only = TRUE, ...)

## S3 method for class 'pglmmObj'
predict(
  object,
  newdata = NULL,
  type = c("link", "response"),
  fixed.only = TRUE,
  ...
)

## S3 method for class 'pglmmObj'
residuals(object, type = c("deviance", "pearson", "response", "working"), ...)

## S3 method for class 'pglmmObj'
print(x, digits = c(fef = 4, ref = 4), ...)

## S3 method for class 'pglmmObj'
summary(
  object,
  digits = c(fef = 4, ref = 4),
  resid_type = switch(object$family$family, gaussian = "pearson", "deviance"),
  ...
)

## S3 method for class 'pglmmObj'
logLik(object, ...)

## S3 method for class 'pglmmObj'
BIC(object, ...)

## S3 method for class 'pglmmObj'
plot(x, fixed.only = FALSE, type = NULL, ...)
```

Arguments

object	pglmmObj object output from glmm, glmmPen, or glmmPen_FineSearch
...	potentially further arguments passed from other methods
x	an R object of class pglmmObj
fixed.only	logical value; default TRUE indicates that only the fixed effects should be used in the fitted value/prediction, while FALSE indicates that both the fixed and random effects posterior modes should be used in the fitted value/prediction
random.only	logical value used in formula; TRUE indicates that only the formula elements relating to the random effects should be returned
formula	in the case of model.frame, a pglmmObj object
type	See details of type options for each function under "Functions" section.

newdata	optional new data.frame containing the same variables used in the model fit procedure
digits	number of significant digits for printing; default of 4
resid_type	type of residuals to summarize in output. See predict.pglmmObj for residual options available.

Value

The pglmmObj object returns the following items:

fixef	vector of fixed effects coefficients
ranef	matrix of random effects coefficients for each explanatory variable for each level of the grouping factor
sigma	random effects covariance matrix
scale	if family is Gaussian, returns the residual error variance
posterior_samples	Samples from the posterior distribution of the random effects, taken at the end of the model fit (after convergence or after maximum iterations allowed). Can be used for diagnositics purposes. Note: These posterior samples are from a single chain.
sampling	character string for type of sampling used to calculate the posterior samples in the E-step of the algorithm
results_all	matrix of results from all model fits during variable selection (if selection performed). Output for each model includes: penalty parameters for fixed (lambda0) and random (lambda1) effects, BIC-derived quantities and the log-likelihood (note: the arguments BIC_option and logLik_calc in selectControl determine which of these quantities are calculated for each model), the number of non-zero fixed and random effects (includes intercept), number of EM iterations used for model fit, whether or not the model converged (0 for no vs 1 for yes), and the fixed and random effects coefficients
results_optim	results from the 'best' model fit; see results_all for details. BICh, BIC, BICN-grp, and LogLik computed for this best model if not previously calculated.
family	Family
penalty_info	list of penalty information
call	arguments plugged into glmm, glmmPen, glmm_FA, or glmmPen_FA
formula	formula
fixed_vars	names of fixed effects variables
data	list of data used in model fit, including the response y, the fixed effects covariates matrix X, the random effects model matrix Z (which is composed of values from the standardized fixed effects model matrix), the grouping factor, offset, model frame, and standarization information used to standardize the fixed effects covariates
optinfo	Information about the optimization of the 'best' model
control_info	optimization parameters used for the model fit

`Estep_init`	materials that can be used to initialize another E-step, if desired
`Gibbs_info`	list of materials to perform diagnositics on the Metropolis-within-Gibbs sample chains, including the Gibbs acceptance rates (included for both the independence and adaptive random walk samplers) and the final proposal standard deviations (included for the adaptive random walk sampler only))
`r_estimation`	list of output related to estimation of number of latent common factors, r. Only relevant for the output of functions `glmm_FA` and `glmmPen_FA`, which are currently in development and are not yet ready for general use.

showClass("pglmmObj") methods(class = "pglmmObj")

Functions

- `fixef.pglmmObj`: Provides the fixed effects coefficients
- `ranef.pglmmObj`: Provides the random effects posterior modes for each explanatory variable for each level of the grouping factor
- `sigma.pglmmObj`: Provides the random effect covariance matrix. If family is Gaussian, also returns the standard deviation of the residual error.
- `coef.pglmmObj`: Computes the sum of the fixed effects coefficients and the random effect posterior modes for each explanatory variable for each level of each grouping factor.
- `family.pglmmObj`: Family of the fitted GLMM
- `nobs.pglmmObj`: Number of observations used in the model fit
- `ngrps.pglmmObj`: Number of levels in the grouping factor
- `formula.pglmmObj`: Formula used for the model fit. Can return the full formula, or just the formula elements relating to the fixed effects (fixed.only = TRUE) or random effects (random.only = TRUE)
- `model.frame.pglmmObj`: Returns the model frame
- `model.matrix.pglmmObj`: Returns the model matrix of either the fixed (type = "fixed") or random effects (type = "random")
- `fitted.pglmmObj`: Fitted values, i.e., the linear predictor of the model.
- `predict.pglmmObj`: Predictions for the model corresponding to the pglmmObj output object from the glmmPen package functions. The function `predict` can predict either the linear predictor of the model or the expected mean of the response, as specified by the `type` argument. Argument `type`: character string for type of predictors: "link" (default), which generates the linear predictor, and "response", which generates the expected mean values of the response.
- `residuals.pglmmObj`: Residuals for the pglmmObj output object from the glmmPen package functions. Argument `type`: character string for type of residuals to report. Options include "deviance" (default), "pearson", "response", and "working", which specify the deviance residuals, Pearson residuals, the difference between the actual response y and the expected mean response (y - mu), and the working residuals (y - mu) / mu
- `print.pglmmObj`: Prints a selection of summary information of fitted model
- `summary.pglmmObj`: Returns a list of summary statistics of the fitted model.

- `logLik.pglmmObj`: Returns the log-likelihood using the Corrected Arithmetic Mean estimator with importance sampling weights developed by Pajor (2017). Degrees of freedom give the sum of the non-zero fixed and random effects coefficients. Citation: Pajor, A. (2017). Estimating the marginal likelihood using the arithmetic mean identity. Bayesian Analysis, 12(1), 261-287.
- `BIC.pglmmObj`: Returns BIC, BICh (hybrid BIC developed by Delattre et al., citation: Delattre, M., Lavielle, M., & Poursat, M. A. (2014). A note on BIC in mixed-effects models. Electronic journal of statistics, 8(1), 456-475.), BICNgrps (BIC using N = number of groups in the penalty term), and possibly BIC-ICQ (labeled as "BICq") if the argument `BIC_option` was set to "BICq" in `selectControl` (citation for BIC-ICQ: Ibrahim, J. G., Zhu, H., Garcia, R. I., & Guo, R. (2011). Fixed and random effects selection in mixed effects models. Biometrics, 67(2), 495-503.)
- `plot.pglmmObj`: Plot residuals for the pglmmObj output object from the glmmPen package. Argument `type`: character string for type of residuals to report. Options include "deviance" (default for non-Gaussian family), "pearson" (default for Gaussian family), "response", and "working", which specify the deviance residuals, Pearson residuals, the difference between the actual response y and the expected mean response (y - mu), and the working residuals (y - mu) / mu

---

`phmm` *Fit a Proportional Hazards Mixed Model via Monte Carlo Expectation Conditional Minimization (MCECM) using a Piecewise Constant Hazard Mixed Model Approximation*

Description

`phmm` is used to approximate a single proportional hazards mixed model using a piecewise constant hazard mixed model approximation via Monte Carlo Expectation Conditional Minimization (MCECM). No model selection is performed.

Usage

```
phmm(
  formula,
  data = NULL,
  covar = NULL,
  offset = NULL,
  optim_options = optimControl(),
  adapt_RW_options = adaptControl(),
  trace = 0,
  tuning_options = lambdaControl(),
  survival_options = survivalControl(),
  progress = TRUE,
  ...
)
```

Arguments

formula	a two-sided linear formula object describing both the fixed effects and random effects part of the model, with the response on the left of a ~ operator and the terms, separated by + operators, on the right. The response must be a Surv object (see Surv from the survival package). Random-effects terms are distinguished by vertical bars ("|") separating expression for design matrices from the grouping factor. formula should be of the same format needed for glmer in package **lme4**. Only one grouping factor will be recognized. The random effects covariates need to be a subset of the fixed effects covariates. The offset must be specified outside of the formula in the 'offset' argument.
data	an optional data frame containing the variables named in formula. If data is omitted, variables will be taken from the environment of formula.
covar	character string specifying whether the covariance matrix should be unstructured ("unstructured") or diagonal with no covariances between variables ("independent"). Default is set to NULL. If covar is set to NULL and the number of random effects predictors (not including the intercept) is greater than or equal to 10 (i.e. high dimensional), then the algorithm automatically assumes an independent covariance structure and covar is set to "independent". Otherwise if covar is set to NULL and the number of random effects predictors is less than 10, then the algorithm automatically assumes an unstructured covariance structure and covar is set to "unstructured".
offset	This can be used to specify an *a priori* known component to be included in the linear predictor during fitting. Default set to NULL (no offset). If the data argument is not NULL, this should be a numeric vector of length equal to the number of cases (the length of the response vector). If the data argument specifies a data.frame, the offset argument should specify the name of a column in the data.frame.
optim_options	a structure of class "optimControl" created from function optimControl that specifies several optimization parameters. See the documentation for optimControl for more details on defaults.
adapt_RW_options	a list of class "adaptControl" from function adaptControl containing the control parameters for the adaptive random walk Metropolis-within-Gibbs procedure. Ignored if optimControl parameter sampler is set to "stan" (default) or "independence".
trace	an integer specifying print output to include as function runs. Default value is 0. See Details for more information about output provided when trace = 0, 1, or 2.
tuning_options	a list of class "selectControl" or "lambdaControl" resulting from selectControl or lambdaControl containing additional control parameters. When function glmm is used,the algorithm may be run using one specific set of penalty parameters lambda0 and lambda1 by specifying such values in lambdaControl(). The default for glmm is to run the model fit with no penalization (lambda0 = lambda1 = 0). When function glmmPen is run, tuning_options is specified using selectControl(). See the lambdaControl and selectControl documentation for further details.

survival_options
: a structure of class "survivalControl" created from function survivalControl that specifies several parameters needed to properly fit the input survival data using a piecewise constant hazard mixed model. See the documentation for survivalControl for more details on defaults.

progress
: a logical value indicating if additional output should be given showing the progress of the fit procedure. If TRUE, such output includes iteration-level information for the fit procedure (iteration number EM_iter, number of MCMC samples nMC, average Euclidean distance between current coefficients and coefficients from t–defined in optimControl–iterations back EM_conv, and number of non-zero fixed and random effects covariates not including the intercept). Additionally, progress = TRUE gives some other information regarding the progress of the variable selection procedure, including the model selection criteria and log-likelihood estimates for each model fit. Default is TRUE.

...
: additional arguments that could be passed into phmmPen. See phmmPen for further details.

Details

The phmm function can be used to approximate a single proportional hazards mixed model using a piecewise constant hazard mixed model. While this approach is meant to be used in the case where the user knows which covariates belong in the fixed and random effects and no penalization is required, one is allowed to specify non-zero fixed and random effects penalties using lambdaControl and the (...) arguments. The (...) allow for specification of penalty-related arguments; see glmmPen and glmmPen_FAfor details. For a high dimensional situation, the user may want to fit a full model using a small penalty for the fixed and random effects and save the posterior draws from this full model for use in any BIC-ICQ calculations during selection within phmmPen. Specifying a file name in the 'BICq_posterior' argument will save the posterior draws from the phmm model into a big.matrix with this file name, see the Details section of phmmPen for additional details.

Value

A reference class object of class pglmmObj for which many methods are available (e.g. methods(class = "pglmmObj"))

---

phmmPen *Fit Penalized Proportional Hazards Mixed Models via Monte Carlo Expectation Conditional Minimization (MCECM) using a Piecewise Constant Hazard Mixed Model Approximation*

Description

phmmPen_FA is used to fit penalized proportional hazards mixed models using a piecewise constant hazard mixed model approximation via Monte Carlo Expectation Conditional Minimization (MCECM). The purpose of the function is to perform variable selection on both the fixed and random effects simultaneously for the piecewise constant hazard mixed model. phmmPen selects the

best model using BIC-type selection criteria (see selectControl documentation for further details). To improve the speed of the algorithm, consider setting var_restrictions = "fixef" within the optimControl options.

Usage

```
phmmPen(
  formula,
  data = NULL,
  covar = NULL,
  offset = NULL,
  fixef_noPen = NULL,
  penalty = c("MCP", "SCAD", "lasso"),
  alpha = 1,
  gamma_penalty = switch(penalty[1], SCAD = 4, 3),
  optim_options = optimControl(),
  adapt_RW_options = adaptControl(),
  trace = 0,
  tuning_options = selectControl(),
  survival_options = survivalControl(),
  BICq_posterior = NULL,
  progress = TRUE
)
```

Arguments

formula	a two-sided linear formula object describing both the fixed effects and random effects part of the model, with the response on the left of a ~ operator and the terms, separated by + operators, on the right. The response must be a Surv object (see Surv from the survival package). Random-effects terms are distinguished by vertical bars ("|") separating expression for design matrices from the grouping factor. formula should be of the same format needed for glmer in package **lme4**. Only one grouping factor will be recognized. The random effects covariates need to be a subset of the fixed effects covariates. The offset must be specified outside of the formula in the 'offset' argument.
data	an optional data frame containing the variables named in formula. If data is omitted, variables will be taken from the environment of formula.
covar	character string specifying whether the covariance matrix should be unstructured ("unstructured") or diagonal with no covariances between variables ("independent"). Default is set to NULL. If covar is set to NULL and the number of random effects predictors (not including the intercept) is greater than or equal to 10 (i.e. high dimensional), then the algorithm automatically assumes an independent covariance structure and covar is set to "independent". Otherwise if covar is set to NULL and the number of random effects predictors is less than 10, then the algorithm automatically assumes an unstructured covariance structure and covar is set to "unstructured".
offset	This can be used to specify an *a priori* known component to be included in the linear predictor during fitting. Default set to NULL (no offset). If the data

argument is not NULL, this should be a numeric vector of length equal to the number of cases (the length of the response vector). If the data argument specifies a data.frame, the offset argument should specify the name of a column in the data.frame.

`fixef_noPen` Optional vector of 0's and 1's of the same length as the number of fixed effects covariates used in the model. Value 0 indicates the variable should not have its fixed effect coefficient penalized, 1 indicates that it can be penalized. Order should correspond to the same order of the fixed effects given in the formula.

`penalty` character describing the type of penalty to use in the variable selection procedure. Options include 'MCP', 'SCAD', and 'lasso'. Default is MCP penalty. If the random effect covariance matrix is "unstructured", then a group MCP, group SCAD, or group LASSO penalty is used on the random effects coefficients. See Breheny and Huang (2011) <doi:10.1214/10-AOAS388> and Breheny and Huang (2015) <doi:10.1007/s11222-013-9424-2> for details of these penalties.

`alpha` Tuning parameter for the Mnet estimator which controls the relative contributions from the MCP/SCAD/LASSO penalty and the ridge, or L2, penalty. `alpha=1` is equivalent to the MCP/SCAD/LASSO penalty, while `alpha=0` is equivalent to ridge regression. However, `alpha=0` is not supported; `alpha` may be arbitrarily small, but not exactly zero

`gamma_penalty` The scaling factor of the MCP and SCAD penalties. Not used by LASSO penalty. Default is 4.0 for SCAD and 3.0 for MCP. See Breheny and Huang (2011) <doi:10.1214/10-AOAS388> and Breheny and Huang (2015) <doi:10.1007/s11222-013-9424-2> for further details.

`optim_options` a structure of class "optimControl" created from function `optimControl` that specifies several optimization parameters. See the documentation for `optimControl` for more details on defaults.

`adapt_RW_options` a list of class "adaptControl" from function `adaptControl` containing the control parameters for the adaptive random walk Metropolis-within-Gibbs procedure. Ignored if `optimControl` parameter `sampler` is set to "stan" (default) or "independence".

`trace` an integer specifying print output to include as function runs. Default value is 0. See Details for more information about output provided when trace = 0, 1, or 2.

`tuning_options` a list of class "selectControl" or "lambdaControl" resulting from `selectControl` or `lambdaControl` containing additional control parameters. When function `glmm` is used,the algorithm may be run using one specific set of penalty parameters `lambda0` and `lambda1` by specifying such values in `lambdaControl()`. The default for `glmm` is to run the model fit with no penalization (`lambda0 = lambda1 = 0`). When function `glmmPen` is run, `tuning_options` is specified using `selectControl()`. See the `lambdaControl` and `selectControl` documentation for further details.

`survival_options` a structure of class "survivalControl" created from function `survivalControl` that specifies several parameters needed to properly fit the input survival data using a piecewise constant hazard mixed model. See the documentation for `survivalControl` for more details on defaults.

BICq_posterior an optional character string expressing the path and file basename of a file combination that will file-back or currently file-backs a big.matrix of the posterior samples from the minimal penalty model used for the BIC-ICQ calculation used for model selection. T (BIC-ICQ reference: Ibrahim et al (2011) <doi:10.1111/j.1541-0420.2010.01463.x>). If this argument is specified as NULL (default) and BIC-ICQ calculations are requested (see selectControl) for details), the posterior samples will be saved in the file combination 'BICq_Posterior_Draws.bin' and 'BICq_Posterior_Draws.desc' in the working directory. See 'Details' section for additional details about the required format of BICq_posterior and the file-backed big matrix.

progress a logical value indicating if additional output should be given showing the progress of the fit procedure. If TRUE, such output includes iteration-level information for the fit procedure (iteration number EM_iter, number of MCMC samples nMC, average Euclidean distance between current coefficients and coefficients from t–defined in optimControl–iterations back EM_conv, and number of non-zero fixed and random effects covariates not including the intercept). Additionally, progress = TRUE gives some other information regarding the progress of the variable selection procedure, including the model selection criteria and log-likelihood estimates for each model fit. Default is TRUE.

Details

Argument BICq_posterior details: If the BIC_option in selectControl (tuning_options) is specified to be 'BICq', this requests the calculation of the BIC-ICQ criterion during the selection process. For the BIC-ICQ criterion to be calculated, a full model assuming a small valued lambda penalty needs to be fit, and the posterior draws from this full model need to be used. In order to avoid repetitive calculations of this full model (i.e. if the user wants to re-run phmmPen with a different set of penalty parameters), a big.matrix of these posterior draws will be file-backed as two files: a backing file with extention '.bin' and a descriptor file with extension '.desc'. The BICq_posterior argument should contain a path and a filename with no extension of the form "./path/filename" such that the backingfile and the descriptor file would then be saved as "./path/filename.bin" and "./path/filename.desc", respectively. If BICq_posterior is set to NULL, then by default, the backingfile and descriptor file are saved in the working directory as "BICq_Posterior_Draws.bin" and "BICq_Posterior_Draws.desc". If the big matrix of posterior draws is already file-backed, BICq_posterior should specify the path and basename of the appropriate files (again of form "./path/filename"); the full model will not be fit again and the big.matrix of posterior draws will be read using the attach.big.matrix function of the bigmemory package and used in the BIC-ICQ calcuations. If the appropriate files do not exist or BICq_posterior is specified as NULL, the full model will be fit and the full model posterior draws will be saved as specified above. The algorithm will save 10^4 posterior draws automatically.

Trace details: The value of 0 (default) does not output any extra information. The value of 1 additionally outputs the updated coefficients, updated covariance matrix values, and the number of coordinate descent iterations used for the M step for each EM iteration. When pre-screening procedure is used and/or if the BIC-ICQ criterion is requested, trace = 1 gives additional information about the penalties used for the 'full model' fit procedure. If Stan is not used as the E-step sampling mechanism, the value of 2 outputs all of the above plus gibbs acceptance rate information for the adaptive random walk and independence samplers and the updated proposal standard deviation for the adaptive random walk.

Value

A reference class object of class pglmmObj for which many methods are available (e.g. methods(class = "pglmmObj"), see ?pglmmObj for additional documentation)

---

phmmPen_FA *Fit a Penalized Proportional Hazards Mixed Model via Monte Carlo Expectation Conditional Minimization (MCECM) using a Piecewise Constant Hazard Mixed Model Approximation*

Description

phmmPen_FA is used to approximate penalized proportional hazards mixed models using using a piecewise constant hazard mixed survival model approximation via Monte Carlo Expectation Conditional Minimization (MCECM) and a factor model decomposition of the random effects. The purpose of the function is to perform variable selection on both the fixed and random effects simultaneously for the piecewise constant hazard mixed model. This function uses a factor model decomposition on the random effects. This assumption reduces the latent space in the E-step (Expectation step) of the algorithm, which reduces the computational complexity of the algorithm. This improves the speed of the algorithm and enables the scaling of the algorithm to larger dimensions. phmmPen_FA selects the best model using BIC-type selection criteria (see selectControl documentation for further details). To improve the speed of the algorithm, consider setting var_restrictions = "fixef" within the optimControl options.

Usage

```
phmmPen_FA(
  formula,
  data = NULL,
  offset = NULL,
  r_estimation = rControl(),
  fixef_noPen = NULL,
  penalty = c("MCP", "SCAD", "lasso"),
  alpha = 1,
  gamma_penalty = switch(penalty[1], SCAD = 4, 3),
  optim_options = optimControl(),
  adapt_RW_options = adaptControl(),
  trace = 0,
  tuning_options = selectControl(),
  survival_options = survivalControl(),
  BICq_posterior = NULL,
  progress = TRUE
)
```

Arguments

formula	a two-sided linear formula object describing both the fixed effects and random effects part of the model, with the response on the left of a ~ operator and the

terms, separated by + operators, on the right. The response must be a Surv object (see Surv from the survival package). Random-effects terms are distinguished by vertical bars ("|") separating expression for design matrices from the grouping factor. formula should be of the same format needed for glmer in package **lme4**. Only one grouping factor will be recognized. The random effects covariates need to be a subset of the fixed effects covariates. The offset must be specified outside of the formula in the 'offset' argument.

data — an optional data frame containing the variables named in formula. If data is omitted, variables will be taken from the environment of formula.

offset — This can be used to specify an *a priori* known component to be included in the linear predictor during fitting. Default set to NULL (no offset). If the data argument is not NULL, this should be a numeric vector of length equal to the number of cases (the length of the response vector). If the data argument specifies a data.frame, the offset argument should specify the name of a column in the data.frame.

r_estimation — a list of class "rControl" from function rControl containing the control parameters for the estimation of the number of latent factors to use in the glmmPen_FA and glmm_FA estimation procedures.

fixef_noPen — Optional vector of 0's and 1's of the same length as the number of fixed effects covariates used in the model. Value 0 indicates the variable should not have its fixed effect coefficient penalized, 1 indicates that it can be penalized. Order should correspond to the same order of the fixed effects given in the formula.

penalty — character describing the type of penalty to use in the variable selection procedure. Options include 'MCP', 'SCAD', and 'lasso'. Default is MCP penalty. If the random effect covariance matrix is "unstructured", then a group MCP, group SCAD, or group LASSO penalty is used on the random effects coefficients. See Breheny and Huang (2011) <doi:10.1214/10-AOAS388> and Breheny and Huang (2015) <doi:10.1007/s11222-013-9424-2> for details of these penalties.

alpha — Tuning parameter for the Mnet estimator which controls the relative contributions from the MCP/SCAD/LASSO penalty and the ridge, or L2, penalty. alpha=1 is equivalent to the MCP/SCAD/LASSO penalty, while alpha=0 is equivalent to ridge regression. However, alpha=0 is not supported; alpha may be arbitrarily small, but not exactly zero

gamma_penalty — The scaling factor of the MCP and SCAD penalties. Not used by LASSO penalty. Default is 4.0 for SCAD and 3.0 for MCP. See Breheny and Huang (2011) <doi:10.1214/10-AOAS388> and Breheny and Huang (2015) <doi:10.1007/s11222-013-9424-2> for further details.

optim_options — a structure of class "optimControl" created from function optimControl that specifies several optimization parameters. See the documentation for optimControl for more details on defaults.

adapt_RW_options — a list of class "adaptControl" from function adaptControl containing the control parameters for the adaptive random walk Metropolis-within-Gibbs procedure. Ignored if optimControl parameter sampler is set to "stan" (default) or "independence".

trace — an integer specifying print output to include as function runs. Default value is 0. See Details for more information about output provided when trace = 0, 1, or 2.

`tuning_options` a list of class "selectControl" or "lambdaControl" resulting from `selectControl` or `lambdaControl` containing additional control parameters. When function `glmm` is used,the algorithm may be run using one specific set of penalty parameters `lambda0` and `lambda1` by specifying such values in `lambdaControl()`. The default for `glmm` is to run the model fit with no penalization (`lambda0 = lambda1 = 0`). When function `glmmPen` is run, `tuning_options` is specified using `selectControl()`. See the `lambdaControl` and `selectControl` documentation for further details.

`survival_options` a structure of class "survivalControl" created from function `survivalControl` that specifies several parameters needed to properly fit the input survival data using a piecewise constant hazard mixed model. See the documentation for `survivalControl` for more details on defaults.

`BICq_posterior` an optional character string expressing the path and file basename of a file combination that will file-back or currently file-backs a `big.matrix` of the posterior samples from the minimal penalty model used for the BIC-ICQ calculation used for model selection. T (BIC-ICQ reference: Ibrahim et al (2011) <doi:10.1111/j.1541-0420.2010.01463.x>). If this argument is specified as `NULL` (default) and BIC-ICQ calculations are requested (see `selectControl`) for details), the posterior samples will be saved in the file combination 'BICq_Posterior_Draws.bin' and 'BICq_Posterior_Draws.desc' in the working directory. See 'Details' section for additional details about the required format of `BICq_posterior` and the file-backed big matrix.

`progress` a logical value indicating if additional output should be given showing the progress of the fit procedure. If `TRUE`, such output includes iteration-level information for the fit procedure (iteration number EM_iter, number of MCMC samples nMC, average Euclidean distance between current coefficients and coefficients from t–defined in `optimControl`–iterations back EM_conv, and number of non-zero fixed and random effects covariates not including the intercept). Additionally, `progress = TRUE` gives some other information regarding the progress of the variable selection procedure, including the model selection criteria and log-likelihood estimates for each model fit. Default is `TRUE`.

Details

Argument `BICq_posterior` details: If the `BIC_option` in `selectControl` (`tuning_options`) is specified to be 'BICq', this requests the calculation of the BIC-ICQ criterion during the selection process. For the BIC-ICQ criterion to be calculated, a full model assuming a small valued lambda penalty needs to be fit, and the posterior draws from this full model need to be used. In order to avoid repetitive calculations of this full model (i.e. if the user wants to re-run `phmmPen` with a different set of penalty parameters), a `big.matrix` of these posterior draws will be file-backed as two files: a backing file with extention '.bin' and a descriptor file with extension '.desc'. The `BICq_posterior` argument should contain a path and a filename with no extension of the form "./path/filename" such that the backingfile and the descriptor file would then be saved as "./path/filename.bin" and "./path/filename.desc", respectively. If `BICq_posterior` is set to `NULL`, then by default, the backingfile and descriptor file are saved in the working directory as "BICq_Posterior_Draws.bin" and "BICq_Posterior_Draws.desc". If the big matrix of posterior draws is already file-backed, `BICq_posterior` should specify the path and basename of the appropriate files (again of form "./path/filename");

the full model will not be fit again and the big.matrix of posterior draws will be read using the attach.big.matrix function of the bigmemory package and used in the BIC-ICQ calcuations. If the appropriate files do not exist or BICq_posterior is specified as NULL, the full model will be fit and the full model posterior draws will be saved as specified above. The algorithm will save 10^4 posterior draws automatically.

Trace details: The value of 0 (default) does not output any extra information. The value of 1 additionally outputs the updated coefficients, updated covariance matrix values, and the number of coordinate descent iterations used for the M step for each EM iteration. When pre-screening procedure is used and/or if the BIC-ICQ criterion is requested, trace = 1 gives additional information about the penalties used for the 'full model' fit procedure. If Stan is not used as the E-step sampling mechanism, the value of 2 outputs all of the above plus gibbs acceptance rate information for the adaptive random walk and independence samplers and the updated proposal standard deviation for the adaptive random walk.

Value

A reference class object of class pglmmObj for which many methods are available (e.g. methods(class = "pglmmObj"), see ?pglmmObj for additional documentation)

---

phmm_FA *Fit a Proportional Hazards Mixed Model via Monte Carlo Expectation Conditional Minimization (MCECM) using a Piecewise Constant Hazard Mixed Model Approximation*

Description

phmm_FA is used to fit a single piecewise constant hazard mixed model as an approximation to a proportional hazards mixed model via Monte Carlo Expectation Conditional Minimization (MCECM). This piecewise constant hazard mixed model uses a factor model decomposition of the random effects. No model selection is performed.

Usage

```
phmm_FA(
  formula,
  data = NULL,
  offset = NULL,
  r_estimation = rControl(),
  optim_options = optimControl(),
  adapt_RW_options = adaptControl(),
  trace = 0,
  tuning_options = lambdaControl(),
  survival_options = survivalControl(),
  progress = TRUE,
  ...
)
```

Arguments

formula	a two-sided linear formula object describing both the fixed effects and random effects part of the model, with the response on the left of a ~ operator and the terms, separated by + operators, on the right. The response must be a Surv object (see Surv from the survival package). Random-effects terms are distinguished by vertical bars ("|") separating expression for design matrices from the grouping factor. formula should be of the same format needed for glmer in package **lme4**. Only one grouping factor will be recognized. The random effects covariates need to be a subset of the fixed effects covariates. The offset must be specified outside of the formula in the 'offset' argument.
data	an optional data frame containing the variables named in formula. If data is omitted, variables will be taken from the environment of formula.
offset	This can be used to specify an *a priori* known component to be included in the linear predictor during fitting. Default set to NULL (no offset). If the data argument is not NULL, this should be a numeric vector of length equal to the number of cases (the length of the response vector). If the data argument specifies a data.frame, the offset argument should specify the name of a column in the data.frame.
r_estimation	a list of class "rControl" from function rControl containing the control parameters for the estimation of the number of latent factors to use in the glmmPen_FA and glmm_FA estimation procedures.
optim_options	a structure of class "optimControl" created from function optimControl that specifies several optimization parameters. See the documentation for optimControl for more details on defaults.
adapt_RW_options	a list of class "adaptControl" from function adaptControl containing the control parameters for the adaptive random walk Metropolis-within-Gibbs procedure. Ignored if optimControl parameter sampler is set to "stan" (default) or "independence".
trace	an integer specifying print output to include as function runs. Default value is 0. See Details for more information about output provided when trace = 0, 1, or 2.
tuning_options	a list of class "selectControl" or "lambdaControl" resulting from selectControl or lambdaControl containing additional control parameters. When function glmm is used,the algorithm may be run using one specific set of penalty parameters lambda0 and lambda1 by specifying such values in lambdaControl(). The default for glmm is to run the model fit with no penalization (lambda0 = lambda1 = 0). When function glmmPen is run, tuning_options is specified using selectControl(). See the lambdaControl and selectControl documentation for further details.
survival_options	a structure of class "survivalControl" created from function survivalControl that specifies several parameters needed to properly fit the input survival data using a piecewise constant hazard mixed model. See the documentation for survivalControl for more details on defaults.
progress	a logical value indicating if additional output should be given showing the progress of the fit procedure. If TRUE, such output includes iteration-level information for

the fit procedure (iteration number EM_iter, number of MCMC samples nMC, average Euclidean distance between current coefficients and coefficients from t–defined in optimControl–iterations back EM_conv, and number of non-zero fixed and random effects covariates not including the intercept). Additionally, progress = TRUE gives some other information regarding the progress of the variable selection procedure, including the model selection criteria and log-likelihood estimates for each model fit. Default is TRUE.

... additional arguments that could be passed into phmmPen_FA. See phmmPen_FA for further details.

Details

The phmm_FA function can be used to approximate a single proportional hazards mixed model using a piecewise constant hazard mixed model. While this approach is meant to be used in the case where the user knows which covariates belong in the fixed and random effects and no penalization is required, one is allowed to specify non-zero fixed and random effects penalties using lambdaControl and the (...) arguments. The (...) allow for specification of penalty-related arguments; see phmmPen_FA for details. For a high dimensional situation, the user may want to fit a full model using a small penalty for the fixed and random effects and save the posterior draws from this full model for use in any BIC-ICQ calculations during selection within phmmPen_FA. Specifying a file name in the 'BICq_posterior' argument will save the posterior draws from the phmm_FA model into a big.matrix with this file name, see the Details section of phmmPen_FA for additional details.

Value

A reference class object of class pglmmObj for which many methods are available (e.g. methods(class = "pglmmObj"))

---

plot_mcmc *Plot Diagnostics for MCMC Posterior Draws of the Random Effects*

Description

Provides graphical diagnostics of the random effect posterior draws from the (best) model. Availabile diagnostics include the sample path, histograms, cummulative sums, and autocorrelation.

Usage

```
plot_mcmc(
  object,
  plots = "sample.path",
  grps = "all",
  vars = "all",
  numeric_grp_order = FALSE,
  bin_width = NULL
)
```

Arguments

object	an object of class pglmmObj
plots	a character string or a vector of character strings specifying which graphical diagnostics to provide. Options include a sample path plot (default, "sample.path"), autocorrelation plots ("autocorr"), histograms ("histogram"), cumulative sum plots ("cumsum"), and all four possible plot options ("all"). While the "all" option will produce all four possible plots, subsets of the types of plots (e.g. sample path plots and autocorrelation plots only) can be specified with a vector of the relevant character strings (e.g. c("sample.path","autocorr"))
grps	a character string or a vector of character strings specifying which groups should have diagnostics provided. The names of the groups match the input group factor levels. Default is set to 'all' for all groups.
vars	a character string or a vector of character strings specifying which variables should have diagnostics provided. Default is set to 'all', which picks all variables with non-zero random effects. Tip: can find the names of the random effect variables in the output sigma matrix found in the pglmmObj object, run sigma(object).
numeric_grp_order	if TRUE, specifies that the groups factor should be converted to numeric values. This option could be used to ensure that the organization of the groups is in the proper numeric order (e.g. groups with levels 1-10 are ordered 1-10, not 1, 10, 2-9).
bin_width	optional binwidth argument for geom_histogram from the ggplot2 package. Default set to NULL, which specifies the default geom_histogram binwidth. This argument only applies if the "histogram" plot type is selected.

Value

a list of ggplot graphics, each faceted by group and random effect variable. Type of plots specified in the plots argument.

---

rControl — *Control of Latent Factor Model Number Estimation Constructs the control structure for the estimation of the number of latent factors (r) for use within the* glmmPen_FA *and* glmm_FA *estimation procedures.*

Description

Control of Latent Factor Model Number Estimation

Constructs the control structure for the estimation of the number of latent factors (r) for use within the glmmPen_FA and glmm_FA estimation procedures.

Usage

```
rControl(
  r = NULL,
  r_max = NULL,
  r_est_method = "GR",
  size = 25,
  sample = FALSE
)
```

Arguments

r	positive integer specifying number of latent common factors to assume in the model. If NULL (default), this value estimated from the data. See r_est_method for available estimation procedures, and the Details section for further details on the general estimation procedure. If r is specified, the no estimation procedure is performed and the algorithm uses the input value or r. All other parameters for this function are relevant for the estimation procedure.
r_max	positive integer specifying maximum number of latent factors to consider. If NULL (default), this value is automatically calculated.
r_est_method	character string indicating method used to estimate number of latent factors r. Default "GR" uses the Growth Ratio method of Ahn and Horenstein (2013) (<doi:10.3982/ECTA8968>). Other available options include "ER" for the Eigenvalue Ratio method of Ahn and Horenstein (2013) (<doi:10.3982/ECTA8968>) and "BN1" or "BN2", the Bai and Ng (2002) method (<dio:10.1111/1468-0262.00273>) using one of two penalties: (1) (d + p) / (d p) log(d p/(d+p)) or (2) (d + p) / (d p) log(min(d,p)) where d is the number of groups in the data and p is the number of total random effect covariates (including the intercept)
size	positive integer specifying the total number of pseudo random effect estimates to use in the estimation procedure for the number of latent factors r, which is restricted to be no less than 25. If this size is greater than the number of groups in the data (i.e.~the number of levels of the grouping variable), then a sampling procedure is used to increase the number of pseudo estimates to the value of size if the value of sample is TRUE.
sample	logical value indicating if the total number of pseudo random effect estimates to use in the estimation procedure for the number of latent common factors r should be larger than the number of unique groups in the data, where the number of pseudo estimates are increased to the value of size using a sampling procedure. Default is FALSE. If TRUE, the sampling procedure is only performed if the value of size is greater than the number of groups in the data.

Details

Estimation of r procedure: For each level of the group variable separately, we identify the observations within that group and fit a regular penalized generalized linear model where the penalty value is the minimum fixed effect penalty. These group-specific estimates, which we label as 'pseudo random effects', are placed into a matrix G (rows = number of levels of the grouping variable, columns

= number of random effect covariates), and this pseudo random effects matrix is treated as the observed outcome matrix used in the "GR", "ER", and "BN" estimation procedures described above in the description of `r_est_method`.

`sim.data` *Simulates data to use for the* `glmmPen` *package*

Description

Simulates data to use for testing the `glmmPen` package. `sim.data` simulates data for `glmmPen`, `sim.data.FA` simulates data for `glmmPen_FA`, and `sim.data.piecewise.exp` simulates data for `phmmPen` and `phmmPen_FA`. Possible parameters to specify includes number of total covariates, number of non-zero fixed and random effects, and the magnitude of the random effect covariance values.

Usage

```
sim.data(
  n,
  ptot,
  pnonzero,
  nstudies,
  sd_raneff = 1,
  family = "binomial",
  corr = NULL,
  seed,
  imbalance = 0,
  beta = NULL,
  pnonzerovar = 0,
  sd_x = 1
)

sim.data.FA(
  n,
  ptot,
  pnonzero,
  nstudies,
  sd_raneff = 0,
  family = "binomial",
  B = NULL,
  r = 2,
  corr = NULL,
  seed,
  imbalance = 0,
  beta = NULL,
  pnonzerovar = 0,
  sd_x = 1
)
```

```
sim.data.piecewise.exp(
  n,
  ptot,
  pnonzero,
  nstudies,
  sd_raneff = 0,
  B = NULL,
  r = 2,
  cut_points = c(0, 0.5, 1, 1.5, 2),
  lhaz_vals = c(-1.5, 1, 2.7, 3.7, 6.8),
  cens_type = c("exp", "unif"),
  cens_max = 5,
  exp_rate = 0.15,
  seed,
  imbalance = 0,
  beta = NULL,
  pnonzerovar = 0,
  sd_x = 1
)
```

Arguments

n	integer specifying total number of samples to generate
ptot	integer specifying total number of covariates to generate (values randomly generated from the standard normal distribution)
pnonzero	integer specifying how may of the covariates should have non-zero fixed and random effects
nstudies	number of studies/groups to have in the data
sd_raneff	non-negative value specifying the standard deviation of the random effects covariance matrix (applied to the non-zero random effects)
family	character string specifying which family to generate data from. Family options include "binomial" (default), "poisson", and "gaussian".
corr	optional value to specify correlation between covariates in the model matrix. Default NULL, only available within sim.data.
seed	integer to use for the setting of a random seed
imbalance	integer of 0 or 1 indicating whether the observations should be equally distributed among the groups (0) or unequally distributed (1).
beta	numeric vector of the fixed effects (including intercept)
pnonzerovar	non-negative integer specifying the number of covariates with a zero-valued fixed effect but a non-zero random effect.
sd_x	non-negative value specifying the standard deviation of the simulated covariates (drawn from a normal distribution with mean 0, standard deviation sd_x)
B	matrix specifying the factor loadings matrix for the random effects, only used within sim.data.FA. Dimensions: number of columns is the number of random

	effects (including the intercept), and number of rows is the number of latent common factors (r). The random effect covariance matrix is specified as Sigma = B x t(B) + diag(sd_raneff)
r	positive integer specifying number of latent common factors that describe the random effects, only used within sim.data.FA
cut_points	vector of cut points to use for the time intervals when simulating piecewise exponential data. Length of cut points must equal length of lhaz_vals, and the value of the first cut point must be 0.
lhaz_vals	vector of the log baseline hazard values for each time interval (which correspond to the time intervals defined by the cut_points argument) within piecewise exponential data. Hazards are assumed to be constant within each time interval.
cens_type	character specifying type of censoring to implement. Default "unif" specifies uniform censoring from 0 to cens_max. The option "exp" specifies exponential censoring with rate exp_rate
cens_max	numeric value used to characterize the range of the uniform censoring procedure (from 0 to cens_max)
exp_rate	numeric value used to characterize the exponential censoring rate (where rate is defined as the rate used in rexp)

Value

list containing the following elements:

y	vector of the response
X	model matrix for the fixed effects
Z	model matrix for the random effects, organized first by variable and then by group
pnonzero	number of non-zero fixed effects
z1	values of the random effects for each variable for each level of the grouping factor
group	grouping factor
X0	model matrix for just the non-zero fixed effects

---

survivalControl *Control for Fitting Piecewise Constant Hazard Mixed Models as an Approximation to Fitting Cox Proportional Hazards Mixed Models*

Description

Constructs the control structure for additional parameters needed to properly fit survival data using a piecewise constant hazard mixed model

Usage

```
survivalControl(
  cut_num = 8,
  interval_type = c("equal", "manual", "group"),
  cut_points = NULL,
  time_scale = 1
)
```

Arguments

`cut_num`	positive integer specifying the number of time intervals to include in the piecewise constant hazard model assumptions for the sampling step. Default is 8. General recommendation: use between 5 and 10 intervals. See the Details section for additional information.
`interval_type`	character specifying how the time intervals are calculated. Options include 'equal' (default), 'manual', or 'group'. If 'equal' (default), time intervals are calculated such that there are approximately equal numbers of events per time interval. If 'manual', the user needs to input their own cut points (see `cut_points` for details). If 'group', time intervals are calculated such that there are approximately equal numbers of events per time interval AND there are at least 4 events observed by each group within each time interval. The input number of time intervals `cut_num` may be modified to a lower number in order to accomplish this goal. This option is generally difficult to perform when there are a large number of groups in the data.
`cut_points`	numeric vector specifying the value of the cut points to use in the calculation of the time intervals for the piecewise constant hazard model. If `interval_type` set to 'equal' or 'group', then this argument is not utilized. If `interval_type` set to 'manual', then this argument is required. First value must be 0, and all values must be ordered smallest to largest.
`time_scale`	positive numeric value (greater than 1) used to scale the time variable in the survival data. In order to calculate the piecewise constant hazard mixed model, the log of the time a subject survived within a particular time interval is used as an offset in the model fit. Sometimes multiplying the time scale by a factor greater than 1 improves the stability of the model fit algorithm.

Details

In the piecewise constant hazard model, there is an assumption that the time line of the data can be cut into `cut_num` time intervals and the baseline hazard is constant within each of these time intervals. In the fit algorithm, we estimate these baseline hazard values (specifically, we estimate the log of the baseline hazard values). By default, we determine cut points by specifying the total number of cuts to make (`cut_num`) and then specifying time values for cut points such that each time interval has an approximately equal number of events (`interval_type = equal`). The authors of this package have found simulations to work well using this default `interval_type = equal`, but if desired, users can further specify that each group has at least some (4) events observed within each time interval. Regardless of the `interval_type` choice, users should be aware that having too many cut points could result in having too few events for each time interval needed for a stable estimation

of the baseline hazard estimates. Additionally, data with few events could result in too few events per time interval even for a small number of cut points. Alternatively, having too few cut points could result in a sub-par approximation of the baseline hazard, which can lead to biased estimation for the coefficients corresponding to the input variables of interest. We generally recommend having 8 total time intervals (more broadly, between 5 and 10 total time intervals). Warnings or errors will occur for cases when there are 1 or 0 events for a time interval. If this happens, either adjust the `cut_num` value appropriately, or in the case when the data simply has a very small number of events, consider not using this software for your estimation purposes.

Value

Function returns a list inheriting from class `survivalControl` containing fit and optimization criteria values used in optimization routine.

`survival_data` *Convert Input Survival Data Into Long-Form Data Needed for Fitting a Piecewise Exponential Model*

Description

Converts the input survival data with one row or element corresponding to a single observation or subject into a long-form dataset where one observation or subject contributes `j` rows, where `j` is the number of time intervals that a subject survived at least part-way through.

Usage

```
survival_data(y, X, Z, group, offset_fit = NULL, survival_options)
```

Arguments

`y`	response, which must be a `Surv` object (see `Surv` from the `survival` package)
`X`	matrix of fixed effects covariates
`Z`	matrix of random effects covariates
`group`	vector specifying the group assignment for each subject
`offset_fit`	vector specifying the offset. This can be used to specify an *a priori* known component to be included in the linear predictor during fitting. Default set to `NULL` (no offset). If the data argument is not `NULL`, this should be a numeric vector of length equal to the number of cases (the length of the response vector).
`survival_options`	a structure of class "survivalControl" created from function `survivalControl` that specifies several parameters needed to properly fit the input survival data using a piecewise constant hazard mixed model. See the documentation for `survivalControl` for more details on defaults.

Index