Package 'FLORAL'

January 20, 2025

Type Package

Title Fit Log-Ratio Lasso Regression for Compositional Data

Version 0.3.0

Date 2024-08-19

Description Log-ratio Lasso regression for continuous, binary, and survival outcomes with compositional features. See Fei and others (2023) <doi:10.1101/2023.05.02.538599>.

License GPL (>= 3)

URL https://vdblab.github.io/FLORAL/

BugReports https://github.com/vdblab/FLORAL/issues

Depends R (>= 3.5.0)

SystemRequirements C++17

Imports Rcpp (>= 1.0.9), stats, survival, ggplot2, survcomp, reshape, dplyr, glmnet, caret, grDevices, utils, mvtnorm, doParallel, doRNG, foreach, msm

LinkingTo Rcpp, RcppArmadillo, RcppProgress, ast2ast

RoxygenNote 7.3.2

Encoding UTF-8

Suggests covr, knitr, rmarkdown, spelling, testthat (>= 3.0.0), patchwork

Language en-US

Config/testthat/edition 3

VignetteBuilder knitr

NeedsCompilation yes

Author Teng Fei [aut, cre, cph] (<https://orcid.org/0000-0001-7888-1715>), Tyler Funnell [aut] (<https://orcid.org/0000-0003-1612-5644>), Nicholas Waters [aut] (<https://orcid.org/0000-0002-9035-2143>), Sandeep Raj [aut] (<https://orcid.org/0000-0003-4629-0528>)

Maintainer Teng Fei <feit1@mskcc.org>

Repository CRAN

Date/Publication 2024-08-20 04:50:02 UTC

Contents

`a.FLORAL` *Comparing prediction performances under different choices of weights for lasso/ridge penalty*

Description

Summarizing `FLORAL` outputs from various choices of `a`

Usage

```
a.FLORAL(
  a = c(0.1, 0.5, 1),
  ncore = 1,
  seed = NULL,
  x,
  y,
  ncov = 0,
  family = "gaussian",
  longitudinal = FALSE,
  id = NULL,
  tobs = NULL,
  failcode = NULL,
  corstr = "exchangeable",
  scalefix = FALSE,
  scalevalue = 1,
  pseudo = 1,
  length.lambda = 100,
  lambda.min.ratio = NULL,
  ncov.lambda.weight = 0,
  mu = 1,
  maxiter = 100,
  ncv = 5,
  intercept = FALSE,
  step2 = FALSE,
  progress = TRUE
)
```

Arguments

a	vector of scalars between 0 and 1 for comparison.
ncore	Number of cores used for parallel computation. Default is to use only 1 core.
seed	A random seed for reproducibility of the results. By default the seed is the numeric form of Sys.Date().
x	Feature matrix, where rows specify subjects and columns specify features. The first ncov columns should be patient characteristics and the rest columns are microbiome absolute counts corresponding to various taxa. If x contains longitudinal data, the rows must be sorted in the same order of the subject IDs used in y.
y	Outcome. For a continuous or binary outcome, y is a vector. For survival outcome, y is a Surv object.
ncov	An integer indicating the number of first ncov columns in x that will not be subject to the zero-sum constraint.
family	Available options are gaussian, binomial, cox, finegray.
longitudinal	TRUE or FALSE, indicating whether longitudinal data matrix is specified for input x. (Longitudinal=TRUE and family="cox" or "finegray" will fit a time-dependent covariate model. Longitudinal=TRUE and family="gaussian" or "binomial" will fit a GEE model.)
id	If longitudinal is TRUE, id specifies subject IDs corresponding to the rows of input x.
tobs	If longitudinal is TRUE, tobs specifies time points corresponding to the rows of input x.
failcode	If family = finegray, failcode specifies the failure type of interest. This must be a positive integer.
corstr	If a GEE model is specified, then corstr is the corresponding working correlation structure. Options are independence, exchangeable, AR-1 and unstructured.
scalefix	TRUE or FALSE, indicating whether the scale parameter is estimated or fixed if a GEE model is specified.
scalevalue	Specify the scale parameter if scalefix=TRUE.
pseudo	Pseudo count to be added to x before taking log-transformation
length.lambda	Number of penalty parameters used in the path
lambda.min.ratio	Ratio between the minimum and maximum choice of lambda. Default is NULL, where the ratio is chosen as 1e-2.
ncov.lambda.weight	Weight of the penalty lambda applied to the first ncov covariates. Default is 0 such that the first ncov covariates are not penalized.
mu	Value of penalty for the augmented Lagrangian
maxiter	Number of iterations needed for the outer loop of the augmented Lagrangian algorithm.
ncv	Folds of cross-validation. Use NULL if cross-validation is not wanted.

intercept	TRUE or FALSE, indicating whether an intercept should be estimated.
step2	TRUE or FALSE, indicating whether a second-stage feature selection for specific ratios should be performed for the features selected by the main lasso algorithm. Will only be performed if cross validation is enabled.
progress	TRUE or FALSE, indicating whether printing progress bar as the algorithm runs.

Value

A ggplot2 object of cross-validated prediction metric versus lambda, stratified by a. Detailed data can be retrieved from the ggplot2 object itself.

Author(s)

Teng Fei. Email: feit1@mskcc.org

References

Fei T, Funnell T, Waters N, Raj SS et al. Scalable Log-ratio Lasso Regression Enhances Microbiome Feature Selection for Predictive Models. bioRxiv 2023.05.02.538599.

Examples

```
set.seed(23420)

dat <- simu(n=50,p=30,model="linear")
pmetric <- a.FLORAL(a=c(0.1,1),ncore=1,x=dat$xcount,y=dat$y,family="gaussian",ncv=2,progress=FALSE)
```

FLORAL *Fit Log-ratio lasso regression for compositional covariates*

Description

Conduct log-ratio lasso regression for continuous, binary and survival outcomes.

Usage

```
FLORAL(
  x,
  y,
  ncov = 0,
  family = "gaussian",
  longitudinal = FALSE,
  id = NULL,
  tobs = NULL,
  failcode = NULL,
  corstr = "exchangeable",
  scalefix = FALSE,
```

```
  scalevalue = 1,
  pseudo = 1,
  length.lambda = 100,
  lambda.min.ratio = NULL,
  ncov.lambda.weight = 0,
  a = 1,
  mu = 1,
  maxiter = 100,
  ncv = 5,
  ncore = 1,
  intercept = FALSE,
  foldid = NULL,
  step2 = TRUE,
  progress = TRUE,
  plot = TRUE
)
```

Arguments

x	Feature matrix, where rows specify subjects and columns specify features. The first ncov columns should be patient characteristics and the rest columns are microbiome absolute counts corresponding to various taxa. If x contains longitudinal data, the rows must be sorted in the same order of the subject IDs used in y.
y	Outcome. For a continuous or binary outcome, y is a vector. For survival outcome, y is a Surv object.
ncov	An integer indicating the number of first ncov columns in x that will not be subject to the zero-sum constraint.
family	Available options are gaussian, binomial, cox, finegray.
longitudinal	TRUE or FALSE, indicating whether longitudinal data matrix is specified for input x. (Longitudinal=TRUE and family="cox" or "finegray" will fit a time-dependent covariate model. Longitudinal=TRUE and family="gaussian" or "binomial" will fit a GEE model.)
id	If longitudinal is TRUE, id specifies subject IDs corresponding to the rows of input x.
tobs	If longitudinal is TRUE, tobs specifies time points corresponding to the rows of input x.
failcode	If family = finegray, failcode specifies the failure type of interest. This must be a positive integer.
corstr	If a GEE model is specified, then corstr is the corresponding working correlation structure. Options are independence, exchangeable, AR-1 and unstructured.
scalefix	TRUE or FALSE, indicating whether the scale parameter is estimated or fixed if a GEE model is specified.
scalevalue	Specify the scale parameter if scalefix=TRUE.
pseudo	Pseudo count to be added to x before taking log-transformation. If unspecified, then the log-transformation will not be performed.

`length.lambda` Number of penalty parameters used in the path

`lambda.min.ratio`
Ratio between the minimum and maximum choice of lambda. Default is `NULL`, where the ratio is chosen as 1e-2.

`ncov.lambda.weight`
Weight of the penalty lambda applied to the first `ncov` covariates. Default is 0 such that the first `ncov` covariates are not penalized.

`a` A scalar between 0 and 1: `a` is the weight for lasso penalty while `1-a` is the weight for ridge penalty.

`mu` Value of penalty for the augmented Lagrangian

`maxiter` Number of iterations needed for the outer loop of the augmented Lagrangian algorithm.

`ncv` Folds of cross-validation. Use `NULL` if cross-validation is not wanted.

`ncore` Number of cores for parallel computing for cross-validation. Default is 1.

`intercept` `TRUE` or `FALSE`, indicating whether an intercept should be estimated.

`foldid` A vector of fold indicator. Default is `NULL`.

`step2` `TRUE` or `FALSE`, indicating whether a second-stage feature selection for specific ratios should be performed for the features selected by the main lasso algorithm. Will only be performed if cross validation is enabled.

`progress` `TRUE` or `FALSE`, indicating whether printing progress bar as the algorithm runs.

`plot` `TRUE` or `FALSE`, indicating whether returning plots of model fitting.

Value

A list with path-specific estimates (beta), path (lambda), and others. Details can be found in `README.md`.

Author(s)

Teng Fei. Email: feit1@mskcc.org

References

Fei T, Funnell T, Waters N, Raj SS et al. Enhanced Feature Selection for Microbiome Data using FLORAL: Scalable Log-ratio Lasso Regression bioRxiv 2023.05.02.538599.

Examples

```
set.seed(23420)

# Continuous outcome
dat <- simu(n=50,p=30,model="linear")
fit <- FLORAL(dat$xcount,dat$y,family="gaussian",ncv=2,progress=FALSE,step2=TRUE)

# Binary outcome
# dat <- simu(n=50,p=30,model="binomial")
# fit <- FLORAL(dat$xcount,dat$y,family="binomial",progress=FALSE,step2=TRUE)
```

```
# Survival outcome
# dat <- simu(n=50,p=30,model="cox")
# fit <- FLORAL(dat$xcount,survival::Surv(dat$t,dat$d),family="cox",progress=FALSE,step2=TRUE)

# Competing risks outcome
# dat <- simu(n=50,p=30,model="finegray")
# fit <- FLORAL(dat$xcount,survival::Surv(dat$t,dat$d,type="mstate"),failcode=1,
#               family="finegray",progress=FALSE,step2=FALSE)
```

`mcv.FLORAL` *Summarizing selected compositional features over multiple cross validations*

Description

Summarizing `FLORAL` outputs from multiple random k-fold cross validations

Usage

```
mcv.FLORAL(
  mcv = 10,
  ncore = 1,
  seed = NULL,
  x,
  y,
  ncov = 0,
  family = "gaussian",
  longitudinal = FALSE,
  id = NULL,
  tobs = NULL,
  failcode = NULL,
  corstr = "exchangeable",
  scalefix = FALSE,
  scalevalue = 1,
  pseudo = 1,
  length.lambda = 100,
  lambda.min.ratio = NULL,
  ncov.lambda.weight = 0,
  a = 1,
  mu = 1,
  maxiter = 100,
  ncv = 5,
  intercept = FALSE,
  step2 = TRUE,
  progress = TRUE,
  plot = TRUE
)
```

Arguments

mcv	Number of random 'ncv'-fold cross-validation to be performed.
ncore	Number of cores used for parallel computation. Default is to use only 1 core.
seed	A random seed for reproducibility of the results. By default the seed is the numeric form of Sys.Date().
x	Feature matrix, where rows specify subjects and columns specify features. The first ncov columns should be patient characteristics and the rest columns are microbiome absolute counts corresponding to various taxa. If x contains longitudinal data, the rows must be sorted in the same order of the subject IDs used in y.
y	Outcome. For a continuous or binary outcome, y is a vector. For survival outcome, y is a Surv object.
ncov	An integer indicating the number of first ncov columns in x that will not be subject to the zero-sum constraint.
family	Available options are gaussian, binomial, cox, finegray.
longitudinal	TRUE or FALSE, indicating whether longitudinal data matrix is specified for input x. (Longitudinal=TRUE and family="cox" or "finegray" will fit a time-dependent covariate model. Longitudinal=TRUE and family="gaussian" or "binomial" will fit a GEE model.)
id	If longitudinal is TRUE, id specifies subject IDs corresponding to the rows of input x.
tobs	If longitudinal is TRUE, tobs specifies time points corresponding to the rows of input x.
failcode	If family = finegray, failcode specifies the failure type of interest. This must be a positive integer.
corstr	If a GEE model is specified, then corstr is the corresponding working correlation structure. Options are independence, exchangeable, AR-1 and unstructured.
scalefix	TRUE or FALSE, indicating whether the scale parameter is estimated or fixed if a GEE model is specified.
scalevalue	Specify the scale parameter if scalefix=TRUE.
pseudo	Pseudo count to be added to x before taking log-transformation
length.lambda	Number of penalty parameters used in the path
lambda.min.ratio	Ratio between the minimum and maximum choice of lambda. Default is NULL, where the ratio is chosen as 1e-2.
ncov.lambda.weight	Weight of the penalty lambda applied to the first ncov covariates. Default is 0 such that the first ncov covariates are not penalized.
a	A scalar between 0 and 1: a is the weight for lasso penalty while 1-a is the weight for ridge penalty.
mu	Value of penalty for the augmented Lagrangian
maxiter	Number of iterations needed for the outer loop of the augmented Lagrangian algorithm.

ncv	Folds of cross-validation. Use NULL if cross-validation is not wanted.
intercept	TRUE or FALSE, indicating whether an intercept should be estimated.
step2	TRUE or FALSE, indicating whether a second-stage feature selection for specific ratios should be performed for the features selected by the main lasso algorithm. Will only be performed if cross validation is enabled.
progress	TRUE or FALSE, indicating whether printing progress bar as the algorithm runs.
plot	TRUE or FALSE, indicating whether returning summary plots of selection probability for taxa features.

Value

A list with relative frequencies of a certain feature being selected over mcv ncv-fold cross-validations.

Author(s)

Teng Fei. Email: feit1@mskcc.org

References

Fei T, Funnell T, Waters N, Raj SS et al. Scalable Log-ratio Lasso Regression Enhances Microbiome Feature Selection for Predictive Models. bioRxiv 2023.05.02.538599.

Examples

```
set.seed(23420)

dat <- simu(n=50,p=30,model="linear")
fit <- mcv.FLORAL(mcv=2,ncore=1,x=dat$xcount,y=dat$y,ncv=2,progress=FALSE,step2=TRUE,plot=FALSE)
```

---

simu *Simulate data following log-ratio model*

Description

Simulate a dataset from log-ratio model.

Usage

```
simu(
  n = 100,
  p = 200,
  model = "linear",
  weak = 4,
  strong = 6,
  weaksize = 0.125,
  strongsize = 0.25,
```

```
  pct.sparsity = 0.5,
  rho = 0,
  timedep_slope = NULL,
  timedep_cor = NULL,
  longitudinal_stability = TRUE,
  ncov = 0,
  betacov = 0,
  intercept = FALSE
)
```

Arguments

n	An integer of sample size
p	An integer of number of features (taxa).
model	Type of models associated with outcome variable, can be "linear", "binomial", "cox", "finegray", or "timedep" (survival endpoint with time-dependent features).
weak	Number of features with weak effect size.
strong	Number of features with strong effect size.
weaksize	Actual effect size for weak effect size. Must be positive.
strongsize	Actual effect size for strong effect size. Must be positive.
pct.sparsity	Percentage of zero counts for each sample.
rho	Parameter controlling the correlated structure between taxa. Ranges between 0 and 1.
timedep_slope	If model is "timedep", this parameter specifies the slope for the feature trajectories. Please refer to the Simulation section of the manuscript for more details.
timedep_cor	If model is "timedep", this parameter specifies the sample-wise correlations between longitudinal features. Please refer to the Simulation section of the manuscript for more details.
longitudinal_stability	If model is "timedep", this is a binary indicator which determines whether the trajectories are more stable (TRUE) or more volatile (FALSE).
ncov	Number of covariates that are not compositional features.
betacov	Coefficients corresponding to the covariates that are not compositional features.
intercept	Boolean. If TRUE, then a random intercept will be generated in the model. Only works for linear or binomial models.

Value

A list with simulated count matrix xcount, log1p-transformed count matrix x, outcome (continuous y, continuous centered y0, binary y, or survival t, d), true coefficient vector beta, list of non-zero features idx, value of intercept intercept (if applicable).

Author(s)

Teng Fei. Email: feit1@mskcc.org

References

Fei T, Funnell T, Waters N, Raj SS et al. Enhanced Feature Selection for Microbiome Data using FLORAL: Scalable Log-ratio Lasso Regression bioRxiv 2023.05.02.538599.

Examples

```
set.seed(23420)
dat <- simu(n=50,p=30,model="linear")
```

Index