Package 'pgee.mixed'

December 21, 2016

Type Package

Title Penalized Generalized Estimating Equations for Bivariate Mixed Outcomes

Version 0.1.0

Description Perform simultaneous estimation and variable selection for correlated bivariate mixed outcomes (one continuous outcome and one binary outcome per cluster) using penalized generalized estimating equations. In addition, clustered Gaussian and binary outcomes can also be modeled. The SCAD, MCP, and LASSO penalties are supported. Cross-validation can be performed to find the optimal regularization parameter(s).

License GPL-2

Encoding UTF-8

LazyData true

Imports mvtnorm (>= 1.0-5), copula (>= 0.999-15), Rcpp (>= 0.12.6), methods (>= 3.3.2)

LinkingTo Rcpp, RcppArmadillo

RoxygenNote 5.0.1

URL http://github.com/kaos42/pgee.mixed

BugReports http://github.com/kaos42/pgee.mixed/issues

NeedsCompilation yes

Author Ved Deshpande [aut, cre]

Maintainer Ved Deshpande <veddeshpande@gmail.com>

Repository CRAN

Date/Publication 2016-12-21 08:30:40

R topics documented:

cv.pgee *Cross validation for Penalized Generalized Estimating Equations*

Description

Performs k-fold cross-validation for Penalized Generalized Estimating Equations (PGEEs) over grid(s) of tuning parameters lambda. Linear and binary logistic models are supported. In particular, can handle the case of bivariate correlated mixed outcomes, in which each cluster consists of one continuous outcome and one binary outcome.

Usage

```
cv.pgee(N, m, X, Z = NULL, y = NULL, yc = NULL, yb = NULL, K = 5,
  grid1, grid2 = NULL, wctype = "Ind", family = "Gaussian", eps = 1e-06,
  maxiter = 1000, tol.coef = 0.001, tol.score = 0.001, init = NULL,
  standardize = TRUE, penalty = "SCAD", warm = TRUE, weights = rep(1,
  N), type_c = "square", type_b = "deviance", marginal = 0, FDR = FALSE,
  fdr.corr = NULL, fdr.type = "all")
```

Arguments

N	Number of clusters.
m	Cluster size. Assumed equal across all clusters. Should be set to 2 for family=="Mixed".
X	Design matrix. If family=="Mixed", then design matrix for continuous responses. For family!="Mixed", should have N*m rows. For family=="Mixed", should have N rows.
Z	Design matrix for binary responses for family=="Mixed". Should not be provided for other family types. If not provided for family=="Mixed", is set equal to X. For family!="Mixed", should have N*m rows. For family=="Mixed", should have N rows.
y	Response vector. Don't use this argument for family == "Mixed". Instead, use arguments yc and yb. Since the cluster size is assumed equal across clusters, the vector is assumed to have the form c(y_1, y_2,...,y_N), with y_i = c(y_i1,...,y_im).
yc	Continuous response vector. Use only for family=="Mixed".
yb	Binary (0/1) response vector. Use only for family=="Mixed".
K	Number of folds.
grid1	For family!="Mixed", the grid of tuning parameters. For family=="Mixed", the grid of tuning parameters for coefficients corresponding to the continuous outcomes.
grid2	For family=="Mixed", the grid of tuning parameters for coefficients corresponding to the binary outcomes. Not used for family!="Mixed".

wctype	Working correlation type; one of "Ind", "CS", or "AR1". For family=="Mixed", "CS" and "AR1" are equivalent.
family	"Gaussian", "Binomial", or "Mixed" (use the last for bivariate mixed outcomes). Note that for "Binomial", currently only binary outcomes are supported.
eps	Disturbance in the Linear Quadratic Approximation algorithm.
maxiter	The maximum number of iterations the Newton algorithm tries before declaring failure to converge.
tol.coef	Converge of the Newton algorithm is declared if two conditions are met: The L1-norm of the difference of successive iterates should be less than tol.coef AND the L1-norm of the penalized score should be less than tol.score.
tol.score	See tol.coef.
init	Vector of initial values for regression coefficients. For family=="Mixed", should be c(init_c, init_b). Defaults to glm values.
standardize	Standardize the design matrices prior to estimation?
penalty	"SCAD", "MCP", or "LASSO".
warm	Use warm starts?
weights	Vector of cluster weights. All observations in a cluster are assumed to have the same weight.
type_c	Loss function for continuous outcomes. "square" (square error loss, the default) or "absolute" (absolute error loss).
type_b	Loss function for binary outcomes. "deviance" (binomial deviance, the default) or "classification" (prediction error).
marginal	For the mixed outcomes case, set to 0 (the default) to account for both the continuous loss and the binary loss, set to 1 to only account for the continuous loss, and set to 2 to only account for the binary loss.
FDR	Should the false discovery rate be estimated for family=="Mixed"? Currently, FDR cannot be estimated for other family types.
fdr.corr	Association parameter to use in FDR estimation. The default is to use the association parameter estimated from the PGEEs.
fdr.type	Estimate the FDR for only the coefficients corresponding to the continuous outcomes ("continuous"), for only the coefficients corresponding to the binary outcomes ("binary"), or for all coefficients ("all", the default).

Details

The function calls pgee.fit K times, each time leaving out 1/K of the data. The cross-validation error is determined by the arguments type_c and type_b. For family=="Mixed", the cross-validation error is (by default) the sum of the continuous error and the binary error.

Value

A list

coefficients	Vector of estimated regression coefficients. For family=="Mixed", this takes the form c(coef_c, coef_b).

vcov	Sandwich formula based covariance matrix of estimated regression coefficients (other than the intercept(s)). The row/column names correspond to elements of coefficients.
phi	Estimated dispersion parameter.
alpha	Estimated association parameter.
num_iterations	Number of iterations the Newton algorithm ran.
converge	0=converged, 1=did not converge.
PenScore	Vector of penalized score functions at the estimated regression coefficients. If the algorithm converges, then these should be close to zero.
FDR	Estimated FDR for family=="Mixed", if requested.
lambda.loss	Cross validation loss (error) for the optimal tuning parameter(s) lambda, averaged across folds.
LossMat	Matrix of cross validation losses. Rows denote tuning parameter values, columns denote folds.

Examples

```
## Not run:
# Gaussian
N <- 100
m <- 10
p <- 50
y <- rnorm(N * m)
# If you want standardize = TRUE, you must provide an intercept.
X <- cbind(1, matrix(rnorm(N * m * (p - 1)), N * m, p - 1))
gr1 <- seq(0.001, 0.1, length.out = 100)
fit <- cv.pgee(X = X, y = y, N = N, m = m, grid1 = gr1, wctype = "CS",
            family = "Gaussian")

# Binary
y <- sample(0:1, N*m, replace = TRUE)
fit <- cv.pgee(X = X, y = y, N = N, m = m, grid1 = gr1, wctype = "CS",
            family = "Binomial")

# Bivariate mixed outcomes
# Generate some data
Bc <- c(2.0, 3.0, 1.5, 2.0, rep(0,times=p-4))
Bb <- c(0.7, -0.7, -0.4, rep(0,times=p-3))
dat <- gen_mixed_data(Bc, Bc, N, 0.5)
# We require two grids of tuning parameters
gr2 <- seq(0.0001, 0.01, length.out = 100)
# Estimate regression coefficients and false discovery rate
fit <- cv.pgee(X = dat$X, Z = dat$Z, yc = dat$yc, yb = dat$yb, N = N, m = 2,
               wctype = "CS", family = "Mixed", grid1 = gr1, grid2 = gr2,
               FDR = TRUE)

## End(Not run)
```

gen_mixed_data *Generate correlated bivariate mixed outcome data*

Description

gen_mixed_data returns randomly generated correlated bivariate mixed outcomes, and covariate matrices to model them, based on design parameters set in the function.

Usage

```
gen_mixed_data(Beta.cont, Beta.bin, N, rho, intercept = TRUE, cov = "same",
  xcor = 0.25, sigma_yc = 1)
```

Arguments

Beta.cont	Vector of true regression coefficients for the continuous outcome.
Beta.bin	Vector of true regression coefficients for the binary outcome.
N	Number of pairs of correlated outcomes.
rho	Gaussian copula parameter.
intercept	Assume an intercept (for both outcomes)? (default TRUE). If TRUE, then the first coefficient in Beta.cont and Beta.bin are assumed to correspond to intercepts.
cov	Specify if the covariate matrices for the continuous outcome and the binary outcome should share all covariates (set to "same"), share some covariates (set to "shared"), or share no covariates (set to "separate").
xcor	Correlation parameter for AR(1) correlation structure of covariate design matrices (assumed same for both).
sigma_yc	Marginal variance of continuous responses.

Details

A Gaussian copula is used to generate the correlated outcomes. Marginally, the continuous outcome follows a normal distribution with identity link to covariates, while the binary outcome follows a Bernoulli distribution with logit link to covariates. Covariates are generated from a zero-mean unit variance multivariate normal distribution, with an AR(1) correlation structure.

Value

A list of generated data

yc	Vector of continuous outcomes.
yb	Vector of binary outcomes.
X	Covariate matrix for the continuous outcomes.
Z	Covariate matrix for the binary outcomes.

Examples

```
# default settings
gen_mixed_data(rnorm(5), rnorm(5), 10, 0.5)
# separate covariate matrices, non-unit continuous variance
gen_mixed_data(rnorm(5), rnorm(5), 10, 0.5, cov = "separate", sigma_yc = 2)
```

---

pgee.fit *Penalized Generalized Estimating Equations*

Description

Estimate regression coefficients using Penalized Generalized Estimating Equations (PGEEs). Linear and binary logistic models are currently supported. In particular, can handle the case of bivariate correlated mixed outcomes, in which each cluster consists of one continuous outcome and one binary outcome.

Usage

```
pgee.fit(N, m, X, Z = NULL, y = NULL, yc = NULL, yb = NULL,
  wctype = "Ind", family = "Gaussian", lambda = 0, eps = 1e-06,
  maxiter = 1000, tol.coef = 0.001, tol.score = 0.001, init = NULL,
  standardize = TRUE, penalty = "SCAD", weights = rep(1, N),
  FDR = FALSE, fdr.corr = NULL, fdr.type = "all")
```

Arguments

N	Number of clusters.
m	Cluster size. Assumed equal across all clusters. Should be set to 2 for family=="Mixed".
X	Design matrix. If family=="Mixed", then design matrix for continuous responses. For family!="Mixed", should have N*m rows. For family=="Mixed", should have N rows. For standardize=TRUE, the first column should be a column vector of ones, corresponding to the intercept.
Z	Design matrix for binary responses for family=="Mixed". Should not be provided for other family types. If not provided for family=="Mixed", is set equal to X. For family!="Mixed", should have N*m rows. For family=="Mixed", should have N rows. For standardize=TRUE, the first column should be a column vector of ones, corresponding to the intercept.
y	Response vector. Don't use this argument for family == "Mixed". Instead, use arguments yc and yb. Since the cluster size is assumed equal across clusters, the vector is assumed to have the form c(y_1, y_2,...,y_N), with y_i = c(y_i1,...,y_im).
yc	Continuous response vector. Use only for family=="Mixed".
yb	Binary (0/1) response vector. Use only for family=="Mixed".

wctype	Working correlation type; one of "Ind", "CS", or "AR1". For family=="Mixed", "CS" and "AR1" are equivalent.
family	"Gaussian", "Binomial", or "Mixed" (use the last for bivariate mixed outcomes). Note that for "Binomial", currently only binary outcomes are supported.
lambda	Tuning parameter(s). A vector of two tuning parameters should be provided for family=="Mixed" (one for the continuous outcome coefficients, and one of the binary outcome coefficients). Otherwise, a single tuning parameter should be provided.
eps	Disturbance in the Linear Quadratic Approximation algorithm.
maxiter	The maximum number of iterations the Newton algorithm tries before declaring failure to converge.
tol.coef	Converge of the Newton algorithm is declared if two conditions are met: The L1-norm of the difference of successive iterates should be less than tol.coef AND the L1-norm of the penalized score should be less than tol.score.
tol.score	See tol.coef.
init	Vector of initial values for regression coefficients. For family=="Mixed", should be c(init_c, init_b). Defaults to glm values.
standardize	Standardize the design matrices prior to estimation?
penalty	"SCAD", "MCP", or "LASSO".
weights	Vector of cluster weights. All observations in a cluster are assumed to have the same weight.
FDR	Should the false discovery rate be estimated for family=="Mixed"? Currently, FDR cannot be estimated for other family types.
fdr.corr	Association parameter to use in FDR estimation. The default is to use the association parameter estimated from the PGEEs.
fdr.type	Estimate the FDR for only the coefficients corresponding to the continuous outcomes ("continuous"), for only the coefficients corresponding to the binary outcomes ("binary"), or for all coefficients ("all", the default).

Details

pgee.fit estimates the regression coefficients for a single value of the tuning paramter (or a single pair of tuning parameters in the mixed outcomes case). To select optimal tuning parameter(s) via k-fold cross validation, see cv.pgee.

For bivariate mixed outcomes, the false discovery rate can be estimated.

Value

A list

coefficients	Vector of estimated regression coefficients. For family=="Mixed", this takes the form c(coef_c, coef_b).
vcov	Sandwich formula based covariance matrix of estimated regression coefficients (other than the intercept(s)). The row/column names correspond to elements of coefficients.

phi	Estimated dispersion parameter.
alpha	Estimated association parameter.
num_iterations	Number of iterations the Newton algorithm ran.
converge	0=converged, 1=did not converge.
PenScore	Vector of penalized score functions at the estimated regression coefficients. If the algorithm converges, then these should be close to zero.
FDR	Estimated FDR for family=="Mixed", if requested.

Examples

```
set.seed(100)
# Gaussian
N <- 100
m <- 10
p <- 10
y <- rnorm(N * m)
# If you want standardize = TRUE, you must provide an intercept.
X <- cbind(1, matrix(rnorm(N * m * (p - 1)), N * m, p - 1))
fit <- pgee.fit(X = X, y = y, N = N, m = m, lambda = 0.5, wctype = "CS",
            family = "Gaussian")
str(fit)
fit$coefficients
fit$vcov

# Binary
y <- sample(0:1, N*m, replace = TRUE)
fit <- pgee.fit(X = X, y = y, N = N, m = m, lambda = 0.1, wctype = "CS",
            family = "Binomial")
str(fit)
fit$coefficients
fit$vcov

# Bivariate mixed outcomes
# Generate some data
Bc <- c(2.0, 3.0, 1.5, 2.0, rep(0, times = p - 4))
Bb <- c(0.7, -0.7, -0.4, rep(0, times = p - 3))
dat <- gen_mixed_data(Bc, Bc, N, 0.5)
# Estimate regression coefficients and false discovery rate
fit <- pgee.fit(X = dat$X, yc = dat$yc, yb = dat$yb, N = N, m = 2,
            wctype = "CS", family = "Mixed", lambda = c(0.1, 0.05),
            FDR = TRUE)
str(fit)
fit$coefficients
fit$vcov
```

---

pgee.mixed *pgee.mixed: Penalized Generalized Estimating Equations for Bivariate Mixed Outcomes*

Description

Perform simultaneous estimation and variable selection for correlated bivariate mixed outcomes (one continuous outcome and one binary outcome per cluster) using penalized generalized estimating equations. In addition, clustered Gaussian and binary outcomes can also be modeled. The SCAD, MCP, and LASSO penalties are supported. Cross-validation can be performed to find the optimal regularization parameter(s).

References

Deshpande, V., Dey, D. K., and Schifano, E. D. (2016). Variable selection for correlated bivariate mixed outcomes using penalized generalized estimating equations. Technical Report 16-23, Department of Statistics, University of Connecticut, Storrs, CT.

Wang, L., Zhou, J., and Qu, A. (2012). Penalized generalized estimating equations for high-dimensional longitudinal data analysis. Biometrics, **68**, 353–360.

Index