Package 'springer'

February 1, 2024

Type Package

Title Sparse Group Variable Selection for Gene-Environment Interactions in the Longitudinal Study

Version 0.1.9

Author Fei Zhou, Yuwen Liu, Xi Lu, Jie Ren, Cen Wu

Maintainer Fei Zhou <fei.zhou@outlook.com>

Description Recently, regularized variable selection has emerged as a powerful tool to identify and dissect gene-environment interactions. Nevertheless, in longitudinal studies with high dimensional genetic factors, regularization methods for G×E interactions have not been systematically developed. In this package, we provide the implementation of sparse group variable selection, based on both the quadratic inference function (QIF) and generalized estimating equation (GEE), to accommodate the bi-level selection for longitudinal G×E studies with high dimensional genomic features. Alternative methods conducting only the group or individual level selection have also been included. The core modules of the package have been developed in C++.

Depends R (>= 3.5.0)

License GPL-2

Encoding UTF-8

URL https://github.com/feizhoustat/springer

BugReports https://github.com/feizhoustat/springer/issues

LazyData true

Imports MASS,Rcpp

LinkingTo Rcpp, RcppArmadillo

RoxygenNote 7.2.3

NeedsCompilation yes

Repository CRAN

Date/Publication 2024-02-01 07:50:02 UTC

R topics documented:

`springer-package` *Sparse Group Variable Selection for Gene-Environment Interactions in the Longitudinal Study*

Description

In this package, we provide a set of regularized variable selection methods tailored for longitudinal studies of gene- environment interactions. The proposed method conducts sparse group variable selection by accounting for bi-level sparsity. Specifically, the individual and group level penalties have been simultaneously imposed to identify important main and interaction effects under three working correlation structures (exchangeable , AR-1 and independence), based on either the quadratic inference function (QIF) or generalized estimating equation (GEE). In addition, only the individual or group level selection in the longitudinal setting can also be conducted using springer. In total, springer provides 18 (=3×3×2) methods. Among them, sparse group variable selection for longitudinal studies have been developed for the first time. Please read the Details below for how to configure the method used.

Details

Users can flexibly choose the methods to fit the model by specifying the three arguments in the user interface **springer()**:

func:	the framework to obtain the score equation. Two choices are available: "GEE" and "QIF".
corr:	working correlation. Three choices are available: "exchangeable", "AR-1" and "independence".
structure:	structural identification. Three choices are available: "bilevel", "group" and "individual".

The function springer() returns a springer object that contains the estimated coefficients.

References

Zhou, F., Liu, Y., Ren, J., Wang, W., and Wu, C. (2023). Springer: An R package for bi-level variable selection of high-dimensional longitudinal data. *Frontiers in Genetics*, 14, 1088223 doi:10.3389/fgene.2023.1088223

Zhou, F., Lu, X., Ren, J., Fan, K., Ma, S. and Wu, C. (2022). Sparse Group Variable Selection for Gene-Environment Interactions in the Longitudinal Study. *Genetic Epidemiology*, 46(5-6), 317-340 doi:10.1002/gepi.22461

Zhou, F., Ren, J., Lu, X., Ma, S. and Wu, C. (2021). Gene-Environment Interaction: a Variable Selection Perspective. *Epistasis: Methods and Protocols, Springer US* doi:10.1007/978107160947-7_13

Zhou, F., Ren, J., Li, G., Jiang, Y., Li, X., Wang, W. and Wu, C. (2019). Penalized Variable Selection for Lipid-Environment Interactions in a Longitudinal Lipidomics Study. *Genes*, 10(12), 1002 doi:10.3390/genes10121002

Zhou, F., Ren, J., Li, X., Wu, C. and Jiang, Y. (2019) interep: Interaction Analysis of Repeated Measure Data. R package version 0.3.1. https://CRAN.R-project.org/package=interep

Ren, J., Du, Y., Li, S., Ma, S., Jiang, Y. and Wu, C. (2019). Robust network-based regularization and variable selection for high-dimensional genomic data in cancer prognosis. *Genetic epidemiology*, 43(3), 276-291 doi:10.1002/gepi.22194

Wu, C., Zhang, Q., Jiang, Y. and Ma, S. (2018). Robust network-based analysis of the associations between (epi) genetic measurements. *Journal of multivariate analysis*, 168, 119-130 doi:10.1016/j.jmva.2018.06.009

Wu, C., Jiang, Y., Ren, J., Cui, Y. and Ma, S. (2018). Dissecting gene-environment interactions: A penalized robust approach accounting for hierarchical structures. *Statistics in Medicine*, 37:437-456 doi:10.1002/sim.7518

Wu, C., Zhong, P.S. and Cui, Y. (2018). Additive varying-coefficient model for nonlinear gene-environment interactions. *Statistical Applications in Genetics and Molecular Biology*, 17(2) doi:10.1515/sagmb20170008

Ren, J., He, T., Li, Y., Liu, S., Du, Y., Jiang, Y. and Wu, C. (2017). Network-based regularization for high dimensional SNP data in the case-control study of Type 2 diabetes. *BMC genetics*, 18(1), 44 doi:10.1186/s1286301704955

Wu, C., Shi, X., Cui, Y. and Ma, S. (2015). A penalized robust semiparametric approach for gene-environment interactions. *Statistics in Medicine*, 34 (30): 4016-4030 doi:10.1002/sim.6609

Wu, C., Cui, Y., and Ma, S. (2014). Integrative analysis of gene-environment interactions under a multi-response partially linear varying coefficient model. *Statistics in Medicine*, 33(28), 4988-4998 doi:10.1002/sim.6287

Wu, C. and Cui, Y. (2014). Boosting signals in gene-based association studies via efficient SNP selection. *Briefings in bioinformatics*, 15(2), 279-291 doi:10.1093/bib/bbs087

Wu, C., Zhong, P.S. and Cui, Y. (2013). High dimensional variable selection for gene-environment interactions. *Technical Report. Michigan State University.*

See Also

springer

`cv.springer` *k-folds cross-validation for springer*

Description

This function conducts k-fold cross-validation for springer and returns the optimal values of the tuning parameters.

Usage

```
cv.springer(
  clin = NULL,
  e,
  g,
  y,
  beta0,
  lambda1,
  lambda2,
  nfolds,
  func,
  corr,
  structure,
  maxits = 30,
  tol = 0.001
)
```

Arguments

`clin`	a matrix of clinical covariates. The default value is NULL. Whether to include the clinical covariates is decided by user.
`e`	a matrix of environment factors.
`g`	a matrix of genetic factors.
`y`	the longitudinal response.
`beta0`	the initial value for the coefficient vector.
`lambda1`	a user-supplied sequence of λ_1 values, which serves as a tuning parameter for the individual-level penalty.
`lambda2`	a user-supplied sequence of λ_2 values, which serves as a tuning parameter for the group-level penalty.
`nfolds`	the number of folds for cross-validation.
`func`	the framework to obtain the score equation. Two choices are available: "GEE" and "QIF".
`corr`	the working correlation structure adopted in the estimation algorithm. The springer provides three choices for the working correlation structure: exchangeable, AR-1,and independence.

`structure`	Three choices are available for structured variable selection. "bilevel" for sparse-group selection on both group-level and individual-level. "group" for selection on group-level only. "individual" for selection on individual-level only.
`maxits`	the maximum number of iterations that is used in the estimation algorithm. The default value is 30.
`tol`	The tolerance level. Coefficients with absolute values that are smaller than the tolerance level will be set to zero. The adhoc value can be chosen as 0.001.

Details

For bi-level sparse group selection, cv.springer returns two optimal tuning parameters, λ_1 and λ_2; for group-level selection, this function returns the optimal λ_2 with λ_1=0; for individual-level selection, this function returns the optimal λ_1 with λ_2=0.

Value

an object of class "cv.springer" is returned, with is a list with components below:

`lam1`	the optimal λ_1.
`lam2`	the optimal λ_2.

---

dat *simulated data for demonstrating the usage of springer*

Description

Simulated gene expression data for demonstrating the usage of springer.

Usage

```
data("dat")
```

Format

The dat file consists of five components: e, g, y, clin and coeff. The coefficients are the true values of parameters used for generating Y.

Details

The data model for generating Y

Consider a longitudinal case study with n subjects and k_i measurements over time for the ith subject ($i = 1, \ldots, n$). Let Y_{ij} be the response of the jth observation for the ith subject ($i = 1, \ldots, n$, $j = 1, \ldots, k_i$), $X_{ij} = (X_{ij1}, ..., X_{ijp})^\top$ be a p-dimensional vector of covariates denoting p genetic factors, $E_{ij} = (E_{ij1}, ..., E_{ijq})^\top$ be a q-dimensional environmental factor and $Clin_{ij} = (Clin_{ij1}, ..., Clin_{ijt})^\top$ be a t-dimensional clinical factor. There is time dependence among measurements on the same subject, but we assume that the measurements between different subjects are

independent. The model we used for hierarchical variable selection for gene–environment interactions is given as:

$$Y_{ij} = \alpha_0 + \sum_{m=1}^{t} \theta_m Clin_{ijm} + \sum_{u=1}^{q} \alpha_u E_{iju} + \sum_{v=1}^{p} (\gamma_v X_{ijv} + \sum_{u=1}^{q} h_{uv} E_{iju} X_{ijv}) + \epsilon_{ij},$$

where α_0 is the intercept and the marginal density of Y_{ij} belongs to a canonical exponential family defined in Liang and Zeger (1986). Define $\eta_v = (\gamma_v, h_{1v}, ..., h_{qv})^\top$, which is a vector of length q+1 and $Z_{ijv} = (X_{ijv}, E_{ij1}X_{ijv}, ..., E_{ijq}X_{ijv})^\top$, which contains the main genetic effect of the vth SNP from the jth measurement on the ith subject and its interactions with all the q environmental factors. The model can be written as:

$$Y_{ij} = \alpha_0 + \sum_{m=1}^{t} \theta_m Clin_{ijm} + \sum_{u=1}^{q} \alpha_u E_{iju} + \sum_{v=1}^{p} \eta_v^\top Z_{ijv} + \epsilon_{ij},$$

where Z_{ijv} is the vth genetic factor and its interactions with the q environment factors for the jth measurement on the ith subject, and η_v is the corresponding coefficient vector of length $1 + q$. The random error $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{ik_i})^T$, which is assumed to follow a multivariate normal distribution with Σ_i as the covariance matrix for the repeated measurements of the ith subject among the k_i time points.

See Also

springer

---

dmcp *The first order derivative function of MCP (Minimax Concave Penalty)*

Description

The first order derivative function of MCP (Minimax Concave Penalty)

Usage

```
dmcp(theta, lambda, gamma = 3)
```

Arguments

theta	a coefficient vector.
lambda	the tuning parameter.
gamma	the regularization parameter for MCP (Minimax Concave Penalty). It balances the unbiasedness and concavity of MCP.

Details

The regularization parameter γ for MCP should be obtained via a data-driven approach in a rigorous way. Among the published studies, it is suggested to check several choices, such as 1.4, 3, 4.2, 5.8, 6.9, and 10, then fix the value. We examined this sequence in our study and found that the results are not sensitive to the choice of value for γ. Therefore, we set the value to 3. To be prudent, other values should also be examined in practice.

Value

the first order derivative of the MCP function.

References

Ren, J., Du, Y., Li, S., Ma, S., Jiang, Y. and Wu, C. (2019). Robust network-based regularization and variable selection for high-dimensional genomic data in cancer prognosis. *Genetic epidemiology*, 43(3), 276-291 doi:10.1002/gepi.22194

Wu, C., Zhang, Q., Jiang, Y. and Ma, S. (2018). Robust network-based analysis of the associations between (epi) genetic measurements. *Journal of multivariate analysis*, 168, 119-130 doi:10.1016/j.jmva.2018.06.009

Ren, J., He, T., Li, Y., Liu, S., Du, Y., Jiang, Y. and Wu, C. (2017). Network-based regularization for high dimensional SNP data in the case-control study of Type 2 diabetes. *BMC genetics*, 18(1), 44 doi:10.1186/s1286301704955

Examples

```
theta=runif(30,-4,4)
lambda=1
dmcp(theta,lambda,gamma=3)
```

`penalty` *This function provides the penalty functions. Users can choose one of the three penalties: sparse group MCP, group MCP and MCP.*

Description

This function provides the penalty functions. Users can choose one of the three penalties: sparse group MCP, group MCP and MCP.

Usage

```
penalty(x, n, t, p, q, beta, lam1, structure, p1, lam2)
```

Arguments

x	the matrix of predictors, consisting of the clinical covariates, environmental factors, genetic factors and gene-environment interactions.
n	the sample size.
t	the number of clinical covariates.
p	the number of predictors, which consists of the clinical covariates, environmental factors, genetic factors and gene-environment interactions.
q	the number of environment factors.
beta	the coefficient vector.
lam1	the tuning parameter λ_1 for individual-level penalty.
structure	Three choices are available for structured variable selection. "bilevel" for sparse-group selection on both group-level and individual-level. "group" for selection on group-level only. "individual" for selection on individual-level only.
p1	the number of genetic factors.
lam2	the tuning parameter λ_2 for group-level penalty.

Details

When structure="bilevel", sparse group MCP is adopted and variable selection for longitudinal data including both genetic main effects and gene-environment interactions will be conducted on both individual and group levels (bi-level selection):

- **Group-level selection:** If the vth genetic factor has any effect at all (associated with the response or not) can be determined by whether $||\eta_v||_2 = 0$.
- **Individual-level selection:** whether the vth genetic variant has main effect, G×E interaction or both can be determined by the nonzero componet.

If structure="group", group MCP will be used and only group-level selection will be conducted on $||\eta_v||_2$; if structure="individual", MCP will be adopted and only individual-level selection will be conducted on each η_{vu}, $(u = 1, \ldots, q)$.

The minimax concave penalty (MCP) is adopted as the baseline penalty function in the springer package. Methods based on other popular choices, such as SCAD and LASSO, will be examined in the future.

Value

H	the penalty function.

`print.springer` *print a springer result*

Description

Print a springer result

Usage

```
## S3 method for class 'springer'
print(x, digits = max(3, getOption("digits") - 3), ...)
```

Arguments

`x`	springer result
`digits`	significant digits in printout.
`...`	other print arguments

See Also

`springer`

`reformat` *This function changes the format of the longitudinal data from wide format to long format*

Description

This function changes the format of the longitudinal data from wide format to long format

Usage

```
reformat(k, y, x)
```

Arguments

`k`	the number of repeated measurements/time points.
`y`	the longitudinal response.
`x`	a matrix of predictors, consisting of clinical covariates, genetic and environment factors, as well as gene-environment interactions.

springer *fit the model with given tuning parameters*

Description

This function performs penalized variable selection for longitudinal data based on generalized estimating equation (GEE) or quadratic inference functions (QIF) with a given value of lambda. Typical usage is to first obtain the optimal lambda using cross validation, then provide it to the springer function.

Usage

```
springer(
  clin = NULL,
  e,
  g,
  y,
  beta0,
  func,
  corr,
  structure,
  lam1,
  lam2,
  maxits = 30,
  tol = 0.001
)
```

Arguments

clin	a matrix of clinical covariates. The default value is NULL. Whether to include the clinical covariates is decided by user.
e	a matrix of environment factors.
g	a matrix of genetic factors.
y	the longitudinal response.
beta0	the initial coefficient vector
func	the framework to obtain the score equation. Two choices are available: "GEE" and "QIF".
corr	the working correlation structure adopted in the estimation algorithm. The springer provides three choices for the working correlation structure: exchangeable, AR-1,and independence.
structure	Three choices are available for structured variable selection. "bilevel" for sparse-group selection on both group-level and individual-level. "group" for selection on group-level only. "individual" for selection on individual-level only.
lam1	the tuning parameter λ_1 for individual-level penalty applied to genetic factors.

lam2	the tuning parameter λ_2 for group-level penalty applied to gene-environment interactions.
maxits	the maximum number of iterations that is used in the estimation algorithm. The default value is 30.
tol	The tolerance level. Coefficients with absolute values that are smaller than the tolerance level will be set to zero. The adhoc value can be chosen as 0.001.

Details

Look back to the data model described in "dat":

$$Y_{ij} = \alpha_0 + \sum_{m=1}^{t} \theta_m Clin_{ijm} + \sum_{u=1}^{q} \alpha_u E_{iju} + \sum_{v=1}^{p} \eta_v^\top Z_{ijv} + \epsilon_{ij},$$

where Z_{ijv} contains the vth genetic main factor and its interactions with the q environment factors for the jth measurement on the ith subject and η_v is the corresponding coefficient vector of length $1 + q$.

When structure="bilevel", variable selection for genetic main effects and gene-environment interactions under the longitudinal response will be conducted on both individual and group levels (bi-level selection):

- **Group-level selection:** by determining whether $||\eta_v||_2 = 0$, we can know if the vth genetic variant has any effect at all.
- **Individual-level selection:** investigate whether the vth genetic variant has main effect, G×E interaction or both, by determining which components in η_v has non-zero values.

If structure="group", only group-level selection will be conducted on $||\eta_v||_2$; if structure="individual", only individual-level selection will be conducted on each η_{vu}, $(u = 1, \ldots, q)$.

This function also provides choices for the framework that is used. If func="QIF", variable selection will be conducted within the quadratic inference functions framework; if func="GEE", variable selection will be conducted within the generalized estimating equation framework.

There are three options for the choice of the working correlation. If corr="exchangeable", the exchangeable working correlation will be applied; if corr="AR-1", the AR-1 working correlation will be adopted; if corr="independence", the independence working correlation will be used. Please check the references for more details.

Value

coef	the coefficient vector.

Examples

```
data("dat")
##load the clinical covariates, environment factors, genetic factors and response from the
##"dat" file
clin=dat$clin
if(is.null(clin)){t=0} else{t=dim(clin)[2]}
e=dat$e
u=dim(e)[2]
```

```
g=dat$g
y=dat$y
##initial coefficient
beta0=dat$coef
##true nonzero coefficients
index=dat$index
beta = springer(clin=clin, e, g, y,beta0,func="GEE",corr="independence",structure="bilevel",
lam1=dat$lam1, lam2=dat$lam2,maxits=30,tol=0.01)
##only focus on the genetic main effects and gene-environment interactions
beta[1:(1+t+u)]=0
##effects that have nonzero coefficients
pos = which(beta != 0)
##true positive and false positive
tp = length(intersect(index, pos))
fp = length(pos) - tp
list(tp=tp, fp=fp)
```

Index