

Package ‘compound.Cox’

March 18, 2017

Type Package

Title Estimation, Gene Selection, and Survival Prediction Based on the Compound Covariate Method Under the Cox Proportional Hazard Model

Version 3.3

Date 2017-3-18

Author Takeshi Emura, Hsuan-Yu Chen, Yi-Hau Chen

Maintainer Takeshi Emura <takeshiemura@gmail.com>

Description Estimation, gene selection, and survival prediction based on the compound covariate method under the Cox model with high-dimensional gene expressions. Available are survival data for non-small-cell lung cancer patients with gene expressions (Chen et al 2007 New Engl J Med) <DOI:10.1056/NEJMoa060096>, statistical methods in Emura et al (2012 PLoS ONE) <DOI:10.1371/journal.pone.0047627> and Emura & Chen (2016 Stat Methods Med Res) <DOI:10.1177/0962280214533378>. Algorithms for generating correlated gene expressions are also available.

License GPL-2

Depends numDeriv, survival

NeedsCompilation no

Repository CRAN

Date/Publication 2017-03-18 17:52:02 UTC

R topics documented:

compound.Cox-package	2
CG.Clayton	3
cindex.CV	4
compound.reg	5
dependCox.reg	7
dependCox.reg.CV	9
Lung	11
PBC	14
uni.selection	16

X.pathway 17

X.tag 18

Index 20

compound.Cox-package	<i>Estimation, Gene Selection, and Survival Prediction Based on the Compound Covariate Method Under the Cox Proportional Hazard Model.</i>
----------------------	--

Description

Estimation, gene selection, and survival prediction based on the compound covariate method under the Cox model with high-dimensional gene expressions. Available are survival data for non-small-cell lung cancer patients with microarrays (Chen et al 2007 New Engl J Med), statistical methods in Emura et al (2012 PLoS ONE) and Emura and Chen (2016 Stat Methods Med Res). Algorithms for generating correlated gene expressions are also available.

Details

Package: compound.Cox
Type: Package
Version: 3.3
Date: 2017-3-18
License: GPL-2

Author(s)

Takeshi Emura, Hsuan-Yu Chen, Yi-Hau Chen; Maintainer: Takeshi Emura <takeshiemura@gmail.com>

References

Matsui S (2006). Predicting Survival Outcomes Using Subsets of Significant Genes in Prognostic Marker Studies with Microarrays. BMC Bioinformatics: 7:156.

Chen HY, Yu SL, Chen CH, et al (2007). A Five-gene Signature and Clinical Outcome in Non-small-cell Lung Cancer, N Engl J Med 356: 11-20.

Emura T, Chen YH, Chen HY (2012). Survival Prediction Based on Compound Covariate under Cox Proportional Hazard Models. PLoS ONE 7(10): e47627. doi:10.1371/journal.pone.0047627

Emura T, Chen YH (2016). Gene Selection for Survival Data Under Dependent Censoring: a Copula-based Approach, Stat Methods Med Res 25(No.6): 2840-57

CG.Clayton*Copula-graphic estimator under the Clayton copula.*

Description

The copula-graphic estimator calculated based on the formula of Rivest & Wells (2001). We used this estimator in Emura & Chen (2016).

Usage

```
CG.Clayton(t.vec, d.vec, alpha, S.plot = TRUE, S.col = "black")
```

Arguments

t.vec	Vector of survival times (time to either death or censoring)
d.vec	Vector of censoring indicators, 1=death, 0=censoring
alpha	Association parameter that is related to Kendall's tau through " $\tau = \alpha / (\alpha + 2)$ "
S.plot	If TRUE, plot the survival curve
S.col	Color of the survival curve

Details

Estimates for survival probability are calculated at given time points of "t.vec". Association parameter "alpha" of the Clayton copula must be given ($\alpha > 0$), where $\alpha = 0$ corresponds to the independence copula.

Value

time	sort(t.vec)
surv	survival probability at "time"

Author(s)

Takeshi Emura

References

Emura T, Chen YH (2016). Gene Selection for Survival Data Under Dependent Censoring: a Copula-based Approach, Stat Methods Med Res 25(No.6): 2840-57.

Rivest LP, Wells MT (2001). A Martingale Approach to the Copula-graphic Estimator for the Survival Function under Dependent Censoring, J Multivar Anal; 79: 138-55.

Examples

```
t.vec=c(1,3,5,4,7,8,10,13)
d.vec=c(1,0,0,1,1,0,1,0)
CG.Clayton(t.vec,d.vec,alpha=18,S.col="blue")
### CG.Clayton gives identical results with the Kaplan-Meier estimator with alpha=0 ###
CG.Clayton(t.vec,d.vec,alpha=0.000000001,S.plot=FALSE)$surv
survfit(Surv(t.vec,d.vec)~1)$surv
```

cindex.CV	<i>Cross-validated c-index for measuring the predictive accuracy of a prognostic index under a copula-based dependent censoring model.</i>
-----------	--

Description

This function calculates the cross-validated c-index (concordance index) for measuring the predictive accuracy of a prognostic index under a copula-based dependent censoring model. Here the prognostic index is calculated as a compound covariate predictor based on the univariate Cox regression estimates. The expression and details are given in Section 3.2 of Emura and Chen (2016). The association between survival time and censoring time is modeled via the Clayton copula.

Usage

```
cindex.CV(t.vec, d.vec, X.mat, alpha, K = 5)
```

Arguments

t.vec	Vector of survival times (time to death or time to censoring, whichever comes first)
d.vec	Vector of censoring indicators, 1=death, 0=censoring
X.mat	n by p matrix of covariates, where n is the sample size and p is the number of covariates
alpha	Association parameter of the Clayton copula; Kendall's tau = $\alpha/(\alpha+2)$
K	The number of cross-validation folds (K=5 is the default)

Details

Currently, only the Clayton copula is implemented for modeling association between survival time and censoring time. The Clayton model yields positive association between survival time and censoring time with the Kendall's tau being equal to $\alpha/(\alpha+2)$, where $\alpha > 0$. The independent copula corresponds to $\alpha = 0$.

If the number of covariates p is large (e.g., $p \geq 100$), the computational time becomes very long. Pre-filtering for covariates is recommended to reduce p.

Value

concordant	Cross-validated c-index
------------	-------------------------

Author(s)

Takeshi Emura

References

Emura T, Chen YH (2016). Gene Selection for Survival Data Under Dependent Censoring: a Copula-based Approach, Stat Methods Med Res 25(No.6): 2840-57.

Examples

```
n=25 ### sample size ###
p=3  ### the number of covariates ###
set.seed(1)
T=rexp(n) ### survival time
U=rexp(n) ### censoring time
t.vec=pmin(T,U) ### minimum of survival time and censoring time
d.vec=as.numeric( c(T<=U) ) ### censoring indicator
X.mat=matrix(runif(n*p),n,p) ### covariates matrix

cindex.CV(t.vec,d.vec,X.mat,alpha=2) ### alpha=2 corresponds to Kendall's tau=0.5
```

compound.reg

Compound shrinkage estimation under the Cox proportional hazard model

Description

This function utilizes the so-called "compound shrinkage estimator" to calculate the regression coefficients under the Cox proportional hazard model, which is proposed by Emura, Chen & Chen (2012). The method is a variant of the Cox partial likelihood estimator such that the regression coefficients are shrunk toward the univariate Cox regression estimators. The resultant estimator is applicable even when the number of covariates is greater than the number of samples (high dimensional and low sample size setting). The standard errors are calculated based on the asymptotic theory (see Emura et al., 2012).

Usage

```
compound.reg(t.vec, d.vec, X.mat, K = 5, delta_a = 0.025, a_0 = 0, var = FALSE,
plot=TRUE, randomize = TRUE, var.detail = FALSE)
```

Arguments

t.vec	Vector of survival times (time to either death or censoring)
d.vec	Vector of censoring indicators, 1=death, 0=censoring
X.mat	Design matrix of n by p, where n is the sample size and p is the number of covariates

K	The number of cross validation folds, $K=n$ corresponds to a leave-one-out cross validation (default=5)
delta_a	The step size for a grid search for the maximum of the cross-validated likelihood (default=0.025)
a_0	The starting value of a grid search for the maximum of the cross-validated likelihood (default=0)
var	If TRUE, the standard deviations and confidence intervals are given (default=FALSE, to reduce the computational cost)
plot	If TRUE, the cross validated likelihood curve and its maximized point are drawn
randomize	If TRUE, randomize the subject ID's so that the subjects in the cross validation folds are randomly chosen (default=TRUE). Otherwise, the cross validation folds are constructed in the ascending sequence
var.detail	Detailed information about the covariance matrix, which is mainly used for theoretical purposes. Please consult Takeshi Emura for more details (default=FALSE)

Details

$K=5$ cross validation is recommended for computational efficiency, though the results appear to be robust against the choice of the number K . If the number of covariates is greater than 200, the computational time becomes very long. In such a case, the univariate pre-selection is recommended to reduce the number of covariates.

Value

a_hat	Estimated value of the shrinkage parameter
beta_hat	Estimated regression coefficients
SE	Standard errors for estimated regression coefficients
Lower95CI	Lower ends of 95 percent confidence intervals ($\beta_hat - 1.96 * SE$)
Upper95CI	Upper ends of 95 percent confidence intervals ($\beta_hat + 1.96 * SE$)
Sigma_hat	Covariance matrix for estimated regression coefficients
V_hat	Estimates of the information matrix ($-[Hessian \text{ of the loglikelihood}]/n$)
Hessian_CV	Second derivative of the cross-validated likelihood. Normally negative since the cross-validated curve is concave
h_dot	Derivative of Equation (8) of Emura et al. (2012) with respect to a shrinkage parameter "a"

Author(s)

Takeshi Emura & Yi-Hau Chen

References

Emura T, Chen Y-H, Chen H-Y (2012) Survival Prediction Based on Compound Covariate under Cox Proportional Hazard Models. PLoS ONE 7(10): e47627. doi:10.1371/journal.pone.0047627

Examples

```

n=50 ### sample size
beta_true=c(1,1,0,0,0)
p=length(beta_true)
t.vec=d.vec=numeric(n)
X.mat=matrix(0,n,p)

set.seed(1)
for(i in 1:n){
  X.mat[i,]=rnorm(p,mean=0,sd=1)
  eta=sum( as.vector(X.mat[i,])*beta_true )
  T=rexp(1,rate=exp(eta))
  C=runif(1,min=0,max=5)
  t.vec[i]=min(T,C)
  d.vec[i]=(T<=C)
}

mean(1-d.vec) ### censored percentage ###

res=compound.reg(t.vec,d.vec,X.mat,delta_a=0.1)
# delta_a=0.10 is used to reduce computational time in this example #
# in general, delta_a=0.025 (default) or delta_a=0.05 is recommended #
res
#### compare the estimate (beta_hat) with the true value ###
beta_true

### Lung cancer data analysis (Emura et al. 2012 PLoS ONE) ###
data(Lung)
temp=Lung[, "train"]==TRUE
t.vec=Lung[temp, "t.vec"]
d.vec=Lung[temp, "d.vec"]
X.mat=as.matrix( Lung[temp, -c(1,2,3)] )
#compound.reg(t.vec=t.vec,d.vec=d.vec,X.mat=X.mat,delta_a=0.025) ## take time to finish ##

```

dependCox.reg

Univariate Cox regression under dependent censoring.

Description

This function performs the univariate Cox regression under dependent censoring based on Chen (2010) and Emura and Chen (2016). The association between survival time and censoring time is modeled via the Clayton copula.

Usage

```
dependCox.reg(t.vec, d.vec, X.vec, alpha, var = TRUE)
```


Arguments

t.vec	Vector of survival times (time to either death or censoring)
d.vec	Vector of censoring indicators, 1=death, 0=censoring
X.vec	Vector of univariate covariates
alpha	Association parameter that is related to Kendall's tau through " $\tau = \alpha/(\alpha+2)$ "
var	If TRUE, the standard deviations are given (use FALSE to reduce the computational cost)

Details

Currently, only the Clayton copula is implemented for association models. The Clayton model yields positive association between failure and censoring times with the Kendall's tau being equal to $\alpha/(\alpha+2)$, where $\alpha > 0$. The independent copula corresponds to $\alpha = 0$.

Value

beta_hat	Estimated regression coefficients
SE	Standard error for the estimated regression coefficients
Z	Z-value for testing $H_0: \beta=0$ (Wald test)
P	P-value for testing $H_0: \beta=0$ (Wald test)

Author(s)

Takeshi Emura

References

Chen YH (2010). Semiparametric Marginal Regression Analysis for Dependent Competing Risks under an Assumed Copula, JRSSB 72: 235-251.

Emura T, Chen YH (2016). Gene Selection for Survival Data Under Dependent Censoring: a Copula-based Approach, Stat Methods Med Res 25(No.6): 2840-57.

Examples

```
### Univariate Cox regression under dependent censoring ###
n=150
beta_true=1.5
t.vec=d.vec=X.vec=numeric(n)
alpha_true=2 ### Kendall's tau=0.5 ##
set.seed(1)

#### Simulated survival data under the Clayton copula model ####
for(i in 1:n){
  X.vec[i]=runif(1)
  eta=X.vec[i]*beta_true
  U=runif(1)
  V=runif(1)
  T=-1/exp(eta)*log(1-U)
```



```

W=(1-U)^(-alpha_true)
C=1/alpha_true/exp(eta)*log( 1-W+W*(1-V)^(-alpha_true/(alpha_true+1)) )
t.vec[i]=min(T,C)
d.vec[i]=(T<=C)
}

#dependCox.reg(t.vec,d.vec,X.vec,alpha=0)## same result as "coxph(Surv(t.vec,d.vec)~X.vec)" ##
dependCox.reg(t.vec,d.vec,X.vec,alpha=alpha_true)## analysis under dependent censoring ##
dependCox.reg(t.vec,d.vec,X.vec,alpha=alpha_true,var=FALSE)## faster computation by "var=FALSE"##
beta_true

### Reproduce Section 5 of Emura and Chen (2016) ###
data(Lung)
temp=Lung[, "train"]==TRUE
t.vec=Lung[temp, "t.vec"]
d.vec=Lung[temp, "d.vec"]
dependCox.reg(t.vec,d.vec,Lung[temp, "ZNF264"],alpha=18)
dependCox.reg(t.vec,d.vec,Lung[temp, "MMP16"],alpha=18)
dependCox.reg(t.vec,d.vec,Lung[temp, "HGF"],alpha=18)
dependCox.reg(t.vec,d.vec,Lung[temp, "HCK"],alpha=18)

```

dependCox.reg.CV

Cox regression under dependent censoring.

Description

This function perform estimation and significance testing for survival data under a copula-based dependent censoring model proposed in Emura and Chen (2016). The dependency between the failure and censoring times is modeled via the Clayton copula. The method is based on the semiparametric maximum likelihood estimation, where the association parameter is estimated by maximizing the cross-validated c-index (see Emura and Chen 2016 for details).

Usage

```
dependCox.reg.CV(t.vec, d.vec, X.mat, K = 5, G = 20)
```

Arguments

t.vec	Vector of survival times (time to either death or censoring)
d.vec	Vector of censoring indicators, 1=death, 0=censoring
X.mat	n by p matrix of covariates, where n is the sample size and p is the number of covariates
K	The number of cross-validation folds
G	The number of grids in searching for the maximam of the cross-validated c-index

Details

Currently, only the Clayton copula is implemented for association models. The Clayton model yields positive association between failure and censoring times with the Kendall's tau being equal to $\alpha/(\alpha+2)$, where $\alpha > 0$. The independent copula corresponds to $\alpha = 0$.

If the number of covariates p is large ($p \geq 100$), the computational time becomes very long. Pre-filtering is recommended to reduce p .

Value

beta_hat	Estimated regression coefficients
SE	Standard error for the estimated regression coefficients
Z	Z-value for testing $H_0: \beta=0$ (Wald test)
P	P-value for testing $H_0: \beta=0$ (Wald test)
alpha	Estimated association parameter for the Clayton copula by maximizing the cross-validated c-index
c_index	Maximized value of the cross-validated c_index

Author(s)

Takeshi Emura

References

Emura T, Chen YH (2016). Gene Selection for Survival Data Under Dependent Censoring: a Copula-based Approach, Stat Methods Med Res 25(No.6): 2840-57

Examples

```
##### Simulated survival data #####
n=25 ### sample size
p=3 ### the number of covariates
set.seed(1)
T=rexp(n) ### survival time
U=rexp(n) ### censoring time
t.vec=pmin(T,U) ### minimum of survival and censoring times
d.vec=as.numeric( c(T<=U) ) ### censoring indicator
X.mat=matrix(runif(n*p),n,p) ### covariate matrix

dependCox.reg.CV(t.vec,d.vec,X.mat,G=10)

### Reproduce Section 5 of Emura and Chen (2016) ###
data(Lung)
temp=Lung[, "train"]==TRUE
t.vec=Lung[temp, "t.vec"]
d.vec=Lung[temp, "d.vec"]
X.mat=as.matrix(Lung[temp, -c(1,2,3)])
#dependCox.reg.CV(t.vec,d.vec,X.mat,G=20) ### take time to finish ###
```

Lung

Survival data for patients with non-small-cell lung cancer.

Description

A subset of the lung cancer data (Chen et al. 2007) is given. The subset consists of 97 gene expressions from 125 patients with non-small-cell lung cancer. The 97 genes were selected with $P\text{-value} < 0.20$ under univariate Cox regression analyses as previously done in Emura et al. (2012) and Emura and Chen (2016). The intensity of gene expression was transformed to an ordinal level using the quantile, i.e. if the intensity of gene expression was $\leq 25\text{th}$, $> 25\text{th}$, $> 50\text{th}$, or $> 75\text{th}$ percentile, it was coded as 1, 2, 3, or 4, respectively (Chen et al. 2007).

Usage

```
data("Lung")
```

Format

A data frame with 125 observations on the following 100 variables.

`t.vec` survival times (time to either death or censoring) in months

`d.vec` censoring indicators, 1=death, 0=censoring

`train` TRUE=training set, FALSE=testing set, as defined in Chen et al. (2007)

`VHL` gene expression, coded as 1, 2, 3, or 4

`IHPK1` gene expression, coded as 1, 2, 3, or 4

`HMMR` gene expression, coded as 1, 2, 3, or 4

`CMKOR1` gene expression, coded as 1, 2, 3, or 4

`PLAU` gene expression, coded as 1, 2, 3, or 4

`IGF2` gene expression, coded as 1, 2, 3, or 4

`FGB` gene expression, coded as 1, 2, 3, or 4

`MYBL2` gene expression, coded as 1, 2, 3, or 4

`ODC1` gene expression, coded as 1, 2, 3, or 4

`MTHFD2` gene expression, coded as 1, 2, 3, or 4

`GLIPR1` gene expression, coded as 1, 2, 3, or 4

`EZH2` gene expression, coded as 1, 2, 3, or 4

`HCK` gene expression, coded as 1, 2, 3, or 4

`CCNC` gene expression, coded as 1, 2, 3, or 4

`XRCC1` gene expression, coded as 1, 2, 3, or 4

`CYP1B1` gene expression, coded as 1, 2, 3, or 4

`CDC25A` gene expression, coded as 1, 2, 3, or 4

`CD44` gene expression, coded as 1, 2, 3, or 4

LCK gene expression, coded as 1, 2, 3, or 4
MTHFS gene expression, coded as 1, 2, 3, or 4
PON3 gene expression, coded as 1, 2, 3, or 4
PTPN6 gene expression, coded as 1, 2, 3, or 4
KIDINS220 gene expression, coded as 1, 2, 3, or 4
KLHL22 gene expression, coded as 1, 2, 3, or 4
RBBP6 gene expression, coded as 1, 2, 3, or 4
GABARAPL2 gene expression, coded as 1, 2, 3, or 4
SEH1L gene expression, coded as 1, 2, 3, or 4
CITED2 gene expression, coded as 1, 2, 3, or 4
BARD1 gene expression, coded as 1, 2, 3, or 4
TLX1 gene expression, coded as 1, 2, 3, or 4
CRMP1 gene expression, coded as 1, 2, 3, or 4
CTNNA1 gene expression, coded as 1, 2, 3, or 4
ANXA5 gene expression, coded as 1, 2, 3, or 4
PTGS2 gene expression, coded as 1, 2, 3, or 4
SMC4L1 gene expression, coded as 1, 2, 3, or 4
LOC285086 gene expression, coded as 1, 2, 3, or 4
ATP11B gene expression, coded as 1, 2, 3, or 4
CDK10 gene expression, coded as 1, 2, 3, or 4
IRF4 gene expression, coded as 1, 2, 3, or 4
MYH11 gene expression, coded as 1, 2, 3, or 4
ME3 gene expression, coded as 1, 2, 3, or 4
CCT6A gene expression, coded as 1, 2, 3, or 4
SNCG gene expression, coded as 1, 2, 3, or 4
MAK3 gene expression, coded as 1, 2, 3, or 4
VCPIP1 gene expression, coded as 1, 2, 3, or 4
JMJD1A gene expression, coded as 1, 2, 3, or 4
STAT2 gene expression, coded as 1, 2, 3, or 4
DDX6 gene expression, coded as 1, 2, 3, or 4
ERBB3 gene expression, coded as 1, 2, 3, or 4
PAX2 gene expression, coded as 1, 2, 3, or 4
PCTK2 gene expression, coded as 1, 2, 3, or 4
NF1 gene expression, coded as 1, 2, 3, or 4
DLG2 gene expression, coded as 1, 2, 3, or 4
JMJD1A.1 gene expression, coded as 1, 2, 3, or 4
SUCLA2 gene expression, coded as 1, 2, 3, or 4

MMP16 gene expression, coded as 1, 2, 3, or 4
AP3B2 gene expression, coded as 1, 2, 3, or 4
HGF gene expression, coded as 1, 2, 3, or 4
MAP2K3 gene expression, coded as 1, 2, 3, or 4
CPEB4 gene expression, coded as 1, 2, 3, or 4
ZNF264 gene expression, coded as 1, 2, 3, or 4
AXL gene expression, coded as 1, 2, 3, or 4
CDC23 gene expression, coded as 1, 2, 3, or 4
MAST3 gene expression, coded as 1, 2, 3, or 4
COX11 gene expression, coded as 1, 2, 3, or 4
PRKAG2 gene expression, coded as 1, 2, 3, or 4
MAN1B1 gene expression, coded as 1, 2, 3, or 4
F8 gene expression, coded as 1, 2, 3, or 4
RSU1 gene expression, coded as 1, 2, 3, or 4
MMD gene expression, coded as 1, 2, 3, or 4
AK5 gene expression, coded as 1, 2, 3, or 4
IDS gene expression, coded as 1, 2, 3, or 4
BNIP1 gene expression, coded as 1, 2, 3, or 4
ENG gene expression, coded as 1, 2, 3, or 4
PCDHGC3 gene expression, coded as 1, 2, 3, or 4
RALY gene expression, coded as 1, 2, 3, or 4
WDR33 gene expression, coded as 1, 2, 3, or 4
RNF4 gene expression, coded as 1, 2, 3, or 4
PRDX1 gene expression, coded as 1, 2, 3, or 4
FXN gene expression, coded as 1, 2, 3, or 4
PTPRU gene expression, coded as 1, 2, 3, or 4
FRAP1 gene expression, coded as 1, 2, 3, or 4
MMP7 gene expression, coded as 1, 2, 3, or 4
CST3 gene expression, coded as 1, 2, 3, or 4
TIMP2 gene expression, coded as 1, 2, 3, or 4
TAL1 gene expression, coded as 1, 2, 3, or 4
STAT1 gene expression, coded as 1, 2, 3, or 4
CCND1 gene expression, coded as 1, 2, 3, or 4
DUSP6 gene expression, coded as 1, 2, 3, or 4
SNRPF gene expression, coded as 1, 2, 3, or 4
MMP13 gene expression, coded as 1, 2, 3, or 4
NR2F6 gene expression, coded as 1, 2, 3, or 4

HOXA1 gene expression, coded as 1, 2, 3, or 4
 RIPK1 gene expression, coded as 1, 2, 3, or 4
 IL7R gene expression, coded as 1, 2, 3, or 4
 SEC13L1 gene expression, coded as 1, 2, 3, or 4
 RPL5 gene expression, coded as 1, 2, 3, or 4

Details

Survival data consisting of 125 patients.

Source

Chen HY, Yu SL, Chen CH, et al (2007). A Five-gene Signature and Clinical Outcome in Non-small-cell Lung Cancer, N Engl J Med 356: 11-20.

References

Chen HY, Yu SL, Chen CH, et al (2007). A Five-gene Signature and Clinical Outcome in Non-small-cell Lung Cancer, N Engl J Med 356: 11-20.

Emura T, Chen YH (2016). Gene Selection for Survival Data Under Dependent Censoring: a Copula-based Approach, Stat Methods Med Res 25(No.6): 2840-57

Examples

```
data(Lung)
Lung[1:3,] ## show the first 3 samples ##

## The five-gene signature in Chen et al. (2007) ##
temp=Lung[, "train"]==TRUE
t.vec=Lung[temp, "t.vec"]
d.vec=Lung[temp, "d.vec"]
coxph(Surv(t.vec, d.vec)~Lung[temp, "ERBB3"])
coxph(Surv(t.vec, d.vec)~Lung[temp, "LCK"])
coxph(Surv(t.vec, d.vec)~Lung[temp, "DUSP6"])
coxph(Surv(t.vec, d.vec)~Lung[temp, "STAT1"])
coxph(Surv(t.vec, d.vec)~Lung[temp, "MMD"])
```

PBC

Primary biliary cirrhosis (PBC) of the liver data

Description

A subset of primary biliary cirrhosis (PBC) of the liver data in the book "Counting Process & Survival Analysis" by Fleming & Harrington (1991). This subset is used in Tibshirani (1997).

Usage

```
data(PBC)
```


Format

A data frame with 276 observations on the following 19 variables.

T Survival times (either time to death or censoring) in days
 d Censoring indicator, 1=death, 0=censoring
 trt Treatment indicator, 1=treatment by D-penicillamine, 0=placebo
 age Age in years (days divided by 365.25)
 sex Sex, 0=male, 1=female
 asc Presence of ascites, 0=no, 1=yes
 hep Presence of hepatomegaly, 0=no, 1=yes
 spi Presence of spiders, 0=no, 1=yes
 ede Presence of edema, 0=no edema, 0.5=edema resolved by therapy, 1=edema not resolved by therapy
 bil log(bilirubin, mg/dl)
 cho log(cholesterol, mg/dl)
 alb log(albumin, gm/dl)
 cop log(urine copper, mg/day)
 alk log(alkaline, U/liter)
 SGO log(SGOT, in U/ml)
 tri log(triglycerides, in mg/dl)
 pla log(platelet count, [the number of platelets per-cubic-milliliter of blood]/1000)
 pro log(prothrombin time, in seconds)
 gra Histologic stage of disease, graded 1, 2, 3, or 4

Details

Survival data consisting of 276 patients with 17 covariates. Among them, 111 patients died (d=1) while others were censored (d=0). The covariates consist of a treatment indicator (trt), age, sex, 5 categorical variables (ascites, hepatomegaly, spider, edema, and stage of disease) and 9 log-transformed continuous variables (bilirubin, cholesterol, albumin, urine copper, alkaline, SGOT, triglycerides, platelet count, and prothrombine).

Source

Fleming & Harrington (1991); Tibshirani (1997)

References

Tibshirani R (1997), The Lasso method for variable selection in the Cox model, *Statistics in Medicine*, 385-395.

Examples

```
data(PBC)
PBC[1:5,] ### profiles for the first 5 patients ###
# See also Appendix D.1 of Fleming & Harrington, Counting Process & Survival Analysis (1991) #
```

uni.selection	<i>Gene selection based on the univariate Cox regression</i>
---------------	--

Description

This function perform gene selection using the univariate Cox regression based on survival data with high-dimensional gene expressions (Matsui 2006; Emura and Chen 2016).

Usage

```
uni.selection(t.vec, d.vec, X.mat, P.value = 0.001, K = 5)
```

Arguments

t.vec	Vector of survival times (time to either death or censoring)
d.vec	Vector of censoring indicators, 1=death, 0=censoring
X.mat	n by p matrix of covariates, where n is the sample size and p is the number of covariates
P.value	Threshold
K	The number of cross-validation folds

Details

Predictive ability of the selected genes are evaluated through cross-validated log-likelihood (CVL) and c-index are computed.

Value

gene	Gene symbols
beta	Estimated regression coefficients
Z	Z-value for testing $H_0: \beta=0$ (Wald test)
P	P-value for testing $H_0: \beta=0$ (Wald test)
c_index	c-index
CVL	Cross-validated partial likelihood

Author(s)

Takeshi Emura

References

Matsui S (2006). Predicting Survival Outcomes Using Subsets of Significant Genes in Prognostic Marker Studies with Microarrays. *BMC Bioinformatics*: 7:156.

Emura T, Chen YH (2016). Gene Selection for Survival Data Under Dependent Censoring: a Copula-based Approach, *Stat Methods Med Res* 25(No.6): 2840-57

Examples

```
data(Lung)
t.vec=Lung$t.vec[Lung$train==TRUE]
d.vec=Lung$d.vec[Lung$train==TRUE]
X.mat=Lung[Lung$train==TRUE,-c(1,2,3)]
uni.selection(t.vec, d.vec, X.mat, P.value=0.05,K=5)
## the outputs reproduce Table 3 of Emura and Chen (2016) ##
```

X.pathway

*Generate a matrix of gene expressions in the presence of pathways***Description**

Generate a matrix of gene expressions in the presence of pathways (Scenario 2 of Emura et al. (2012)).

Usage

```
X.pathway(n,p,q1,q2)
```

Arguments

n	the number of individuals (sample size)
p	the number of genes
q1	the number of positive non-null genes
q2	the number of negative non-null genes

Details

n by p matrix of gene expressions are generated. Correlation between columns is introduced to reflect the presence of gene pathways. The distribution of each column is standardized to have mean=0 and SD=1. If two genes are correlated, the correlation is 0.5. Otherwise, the correlation is 0. Details are referred to p.4 of Emura et al. (2012). This deta generation scheme was used in the simulations of Emura et al. (2012), Emura and Chen (2016) and Emura et al. (2017).

Value

X	n by p matrix of gene expressions
---	-----------------------------------

Author(s)

Takeshi Emura & Yi-Hau Chen

References

Emura T, Chen YH, Chen HY (2012). Survival Prediction Based on Compound Covariate under Cox Proportional Hazard Models. PLoS ONE 7(10): e47627. doi:10.1371/journal.pone.0047627

Emura T, Chen YH (2016). Gene Selection for Survival Data Under Dependent Censoring: a Copula-based Approach, Stat Methods Med Res 25(No.6): 2840-57

Emura T, Nakatochi M, Matsui S, Michimae H, Rondeau V (2017) Personalized dynamic prediction of death according to tumour progression and high-dimensional genetic factors: meta-analysis with a joint model, Stat Methods Med Res, doi:10.1177/0962280216688032

Examples

```
X.mat=X.pathway(n=200,p=100,q1=10,q2=10)
round( colMeans(X.mat),3 ) ## mean ~ 0 ##
round( apply(X.mat, MARGIN=2, FUN=sd),3) ## SD ~ 1 ##
```

X.tag	<i>Generate a matrix of gene expressions in the presence of tag genes</i>
-------	---

Description

Generate a matrix of gene expressions in the presence of tag genes (Scenario 1 of Emura et al. (2012)).

Usage

```
X.tag(n, p, q, s = 1)
```

Arguments

- | | |
|---|--|
| n | the number of individuals (sample size) |
| p | the number of genes |
| q | the number of non-null genes |
| s | the number of null genes correlated with a non-null gene (tag) |

Details

n by p matrix of gene expressions are generated. Correlation between columns is introduced to reflect the presence of tag genes. The distribution of each column is standardized to have mean=0 and SD=1. If two genes are correlated, the correlation is 0.5. Otherwise, the correlation is 0. Details are referred to p.4 of Emura et al. (2012). This deta generation scheme was used in the simulations of Emura et al. (2012) and Emura and Chen (2016).

Value

- | | |
|---|-----------------------------------|
| X | n by p matrix of gene expressions |
|---|-----------------------------------|

Author(s)

Takeshi Emura & Yi-Hau Chen

References

Emura T, Chen YH, Chen HY (2012). Survival Prediction Based on Compound Covariate under Cox Proportional Hazard Models. PLoS ONE 7(10): e47627. doi:10.1371/journal.pone.0047627

Emura T, Chen YH (2016). Gene Selection for Survival Data Under Dependent Censoring: a Copula-based Approach, Stat Methods Med Res 25(No.6): 2840-57.

Examples

```
X.mat=X.tag(n=200,p=100,q=10,s=4)
round( colMeans(X.mat),3 ) ## mean ~ 0 ##
round( apply(X.mat, MARGIN=2, FUN=sd),3) ## SD ~ 1 ##
```


Index

*Topic **\textasciitildekwd1**

compound.reg, [5](#)

*Topic **\textasciitildekwd2**

CG.Clayton, [3](#)

cindex.CV, [4](#)

compound.reg, [5](#)

uni.selection, [16](#)

X.pathway, [17](#)

X.tag, [18](#)

*Topic **c-index**

cindex.CV, [4](#)

uni.selection, [16](#)

*Topic **cross-validated partial likelihood**

uni.selection, [16](#)

*Topic **cross-validation**

cindex.CV, [4](#)

*Topic **datasets**

Lung, [11](#)

PBC, [14](#)

*Topic **dependent censoring**

CG.Clayton, [3](#)

dependCox.reg, [7](#)

dependCox.reg.CV, [9](#)

*Topic **gene expression**

Lung, [11](#)

X.pathway, [17](#)

X.tag, [18](#)

*Topic **package**

compound.Cox-package, [2](#)

*Topic **univariate Cox regression**

dependCox.reg, [7](#)

dependCox.reg.CV, [9](#)

uni.selection, [16](#)

dependCox.reg, [7](#)

dependCox.reg.CV, [9](#)

Lung, [11](#)

PBC, [14](#)

uni.selection, [16](#)

X.pathway, [17](#)

X.tag, [18](#)

CG.Clayton, [3](#)

cindex.CV, [4](#)

compound.Cox (compound.Cox-package), [2](#)

compound.Cox-package, [2](#)

compound.reg, [5](#)