Package ‘LatentClassJM’

September 15, 2022

Type Package

Title Joint latent-class modeling for multivariate longitudinal measurements and survival data

Version 0.1.0

Author Kin Yau (Alex) Wong <kin-yau.wong@polyu.edu.hk>

Maintainer Kin Yau (Alex) Wong <kin-yau.wong@polyu.edu.hk>

Description Perform sieve nonparametric maximum likelihood estimation for a semiparametric latent-class joint model using an EM algorithm

License GPL-3

Encoding UTF-8

LazyData true

RoxygenNote 7.1.1

Depends survival,
nnet,
statmod,
splines2

NeedsCompilation no

R topics documented:

`create.data` *Generate Dataset*

Description

Create a dataset based on the setting of the simulation studies in Wong et al. (2022)

Usage

```
create.data(n, seed = 1)
```

Arguments

`n`	Sample size
`seed`	Seed of the random generator (optional)

Value

A list of the following components:

- **Y** : An $(n \times J \times m)$ array of longitudinal outcome measurements, where n is the sample size, J is the number of longitudinal measurement types, and m is the maximum number of measurement times. It can contain `NA` values if the number of measurements for a subject is fewer than the maximum number of measurements. The (i, j, k)th element corresponds to the kth measurement of the jth type of longitudinal outcome for the ith subject
- **X** : An $(n \times J \times m \times p_X)$ array of covariates (excluding intercept) of the longitudinal outcome model, where n is the sample size, J is the number of longitudinal measurement types, m is the number of measurement times, and p_X is the number of covariates. The (i, j, k, l)th element corresponds to the lth covariate for the kth measurement of the jth type of longitudinal outcome for the ith subject
- **W** : An $(n \times p_W)$ matrix of covariates for the latent class regression model, where n is the sample size, and p_W is the number of covariate. The (i, l)th element corresponds to the lth covariate for the ith subject
- **Time** : An n-vector of observed event or censoring times
- **D** : An n-vector of event indicators
- **ni** : An $(n \times J)$ matrix of numbers of measurements for the longitudinal outcomes
- **Z** : An $(n \times p_Z)$ matrix of time-independent covariates for the survival model, where n is the sample size, and p_Z is the number of covariates. The (i, l)th element corresponds to the lth covariate for the ith subject

Based on the setting of the simulation studies in Wong et al. (2022), we fix $J = 2$, $m = 10$, $p_X = 3$, $p_W = 2$, and $p_Z = 2$.

References

Wong, K. Y., Zeng, D., & Lin, D. Y. (2022). Semiparametric latent-class models for multivariate longitudinal and survival data. The Annals of Statistics. 50 487–510.

Examples

```
dataset <- create.data(n=1000)
```

`LatentClassJM` *Sieve nonparametric maximum likelihood estimation for the semiparametric latent-class joint model*

Description

This function performs the (accelerated) EM algorithm to compute the sieve nonparametric maximum likelihood estimator. The algorithm starts with the standard EM algorithm. Once the difference between the log-likelihood values or the parameter values of consecutive iterations becomes smaller than a certain threshold, an accelerated EM algorithm (Vardhan and Roland 2008) will be adopted until convergence.

Usage

```
LatentClassJM(
  Y,
  X,
  W,
  Time,
  D,
  ni,
  Z,
  G,
  nknots,
  knots = NA,
  degree,
  covar = "ind",
  like.diff1 = FALSE,
  like.diff2 = TRUE,
  accelEM = TRUE,
  bound = 5,
  h = 10,
  epsilon = 0.001,
  epsilon2 = 1e-06,
  init.param = NULL,
  h2 = 10,
  cal.inf = FALSE,
  max.iter = 5000,
  seed = 1
)
```

Arguments

`Y`	An $(n \times J \times m)$ array of longitudinal outcome measurements, where n is the sample size, J is the number of longitudinal measurement types, and m is the maximum number of measurement times. It can contain `NA` values if the number of measurements for a subject is fewer than the maximum number of measurements. The (i, j, k)th element corresponds to the kth measurement of the jth type of longitudinal outcome for the ith subject
`X`	An $(n \times J \times m \times p_X)$ array of covariates (excluding intercept) of the longitudinal outcome model, where n is the sample size, J is the number of longitudinal measurement types, m is the number of measurement times, and p_X is the number of covariates. The (i, j, k, l)th element corresponds to the lth covariate for the kth measurement of the jth type of longitudinal outcome for the ith subject
`W`	An $(n \times p_W)$ matrix of covariates for the latent class regression model, where n is the sample size, and p_W is the number of covariate. The (i, l)th element corresponds to the lth covariate for the ith subject
`Time`	An n-vector of observed event or censoring times
`D`	An n-vector of event indicators
`ni`	An $(n \times J)$ matrix of numbers of measurements for the longitudinal outcomes
`Z`	An $(n \times p_Z)$ matrix of time-independent covariates for the survival model, where n is the sample size, and p_Z is the number of covariates. The (i, l)th element corresponds to the lth covariate for the ith subject

`G`	Number of latent classes
`nknots`	Number of interior knots for the B-spline basis functions
`knots`	An optional vector of interior knot positions. If not supplied, then the interior knots will be selected based on quantiles of the observed event times
`degree`	The degree of the B-spline basis functions
`covar`	Covariance structure for Y. For covar = `ind`, repeated longitudinal measurements are independent conditional on the random effect b and latent class C ($\sigma_{gj2} = 0$); for covar = `exchange`, repeated longitudinal measurements have an exchangeable covariance matrix conditional on the random effect b and latent class C ($\sigma_{gj2} \neq 0$); σ_{gj2} is defined in Details below. Default is `ind`
`like.diff1`	Logical; If `TRUE`, then convergence of the standard EM algorithm is based on the difference between log-likelihood values of consecutive iterations; otherwise, convergence is based on the maximum difference between parameter values; Default is `FALSE`
`like.diff2`	Logical; If `TRUE`, then convergence of the accelerated EM algorithm is based on the difference between log-likelihood values of consecutive iterations; otherwise, convergence is based on the maximum difference between parameter values; Default is `TRUE`
`accelEM`	Logical; The iteration begins with standard EM algorithm. If `TRUE`, then the accelerated EM algorithm will be adopted after the end of the standard EM algorithm; otherwise, the program terminates after the standard EM algorithm
`bound`	The upper bound of the absolute value of the parameter estimates
`h`	The number of abscissas for the Gauss-Hermite quadrature in the E-step
`epsilon`	Threshold for convergence of the standard EM algorithm
`epsilon2`	Threshold for convergence of the accelerated EM algorithm
`init.param`	A named list of user-input initial values of model parameters, including `alpha`, `beta`, `sigma2`, `xi`, `eta`, `gamma` and `haz` • `alpha` is a matrix of $(G \times p_W)$ regression parameters for the multinomial regression; the last row must be zero • `beta` is an array of $(G \times J \times p_X)$ regression parameters • `sigma2` is the variance of the error terms of the longitudinal measurements. If covar = `exchange`, then `sigma2` is a $(G \times J \times 2)$ array. The $(g, j, 1)$th element is σ_{gj1}, and the $(g, j, 2)$th element is σ_{gj2}. If covar = `ind`, then `sigma2` is a $(G \times J)$ matrix. The (g, j)th element is σ_{gj1}. In this case, σ_{gj2} is fixed to be zero • `xi` is a G-vector of class-specific variances of the latent variable • `eta` is a G-vector of class-specific regression parameters of the random effect in the survival model • `gamma` is a $(G \times (p_Z + q))$ matrix of class-specific regression parameters, consisting of 2 parts. The first p_Z columns correspond to regression parameters of the covariates `Z`, and the last q columns correspond to regression parameters of the spline functions, where $q =$ `nknots`+`degree`+1; that is, the gth row of `gamma` is $(\boldsymbol{\gamma}_g^T, \boldsymbol{a}_g^T)$. See Details below • `haz` is a vector of jumps of the first class-specific cumulative hazard function. The jumps should correspond to the ordered unique observed event times
`h2`	The number of abscissas for the Gauss-Hermite quadrature in the calculation of the log-likelihood

`cal.inf`	Logical; if `TRUE`, then the information matrix will be calculated
`max.iter`	Maximum number of iterations
`seed`	Seed used for parameter initialization; default is 1

Details

In this function, we consider a special case of the model introduced in Wong et al. (2022). We consider a model with G latent classes. Let C denote the latent class membership, with $C = g$ if a subject belongs to the gth latent class ($g = 1, ..., G$). We fit a multinomial logistic regression model for C:

$$P(C = g \mid \boldsymbol{W}) = \frac{e^{\boldsymbol{\alpha}_g^T \boldsymbol{W}}}{\sum_{l=1}^{G} e^{\boldsymbol{\alpha}_l^T \boldsymbol{W}}},$$

where $\boldsymbol{W}$ is a vector of time-independent covariates that include the constant 1 and $\boldsymbol{\alpha}_g$ is the vector of class-specific regression parameters with $\boldsymbol{\alpha}_G = 0$. Each latent class is characterized by class-specific trajectories of multivariate longitudinal outcomes and a class-specific risk of the event of interest. The longitudinal outcomes and the event time are assumed to be conditionally independent given the latent class membership and a multivariate random effect.

Suppose that there are J types of longitudinal outcomes, and the jth type is measured at N_j time points. For $j = 1, \ldots, J$ and $k = 1, \ldots, N_j$, let Y_{jk} denote the kth measurement of the jth longitudinal outcome and $\boldsymbol{X}_{jk}$ denote corresponding covariates, which include the constant 1. We assume:

$$Y_{jk} \mid_{C=g} = \boldsymbol{\beta}_g^T \boldsymbol{X}_{jk} + b + \epsilon_{jk}$$

for $g = 1, \ldots, G$, where $\boldsymbol{X}_{jk}$ is a vector of covariates that include the constant 1, $\boldsymbol{\beta}_g$ is a vector of class-specific regression parameters, and b is a normal random effect with mean 0 and variance ξ_g. The error terms $(\epsilon_{j1}, \ldots, \epsilon_{jN_j})$ are dependent zero-mean normal random variables with variance $\sigma_{gj1} + \sigma_{gj2}$ and pairwise covariance σ_{gj2}.

Let T denote the event time of interest. We assume a proportional hazards model:

$$\lambda(t \mid \boldsymbol{Z}, \boldsymbol{b}, C = g) = \lambda_g(t) e^{\boldsymbol{\gamma}_g^T \boldsymbol{Z} + \eta_g b}$$

where $\boldsymbol{Z}$ is a vector of time-independent covariates, $\lambda_g(.)$ is an arbitrary class-specific baseline hazard function, and $\boldsymbol{\gamma}_g$ and η_g are class-specific regression parameters.

We use a sieve nonparametric maximum likelihood estimation method to estimate the model parameters. In particular, we let $\lambda = \lambda_1$ and $\psi_g = \log(\lambda_g/\lambda_1)$ for $g = 1, \ldots, G$. We approximate ψ_g by $\sum_{j=1}^{q} a_{gj} B_j$, where $B_1, \ldots, B_q$ are B-spline functions. Then, we can write the survival model as

$$\lambda(t \mid \boldsymbol{Z}, \boldsymbol{b}, C = g) = \lambda(t) e^{\boldsymbol{\gamma}_g^T \boldsymbol{Z} + \boldsymbol{a}_g^T \boldsymbol{B}(t) + \eta_g b}$$

where $\boldsymbol{a}_g = (a_{g1}, \ldots, a_{gq})^T$ and $\boldsymbol{B}(t) = (B_1(t), \ldots, B_q(t))^T$.

Value

A list of the following components:

- **alpha** : A matrix of $(G \times p_W)$ regression parameters for the multinomial regression. The gth row is the parameter vector for the gth latent class; the last row must be zero
- **beta** : An array of $(G \times J \times p_X)$ regression parameters. The (g, j)th row is the lth parameter vector for the gth latent class at jth measurement type
- **sigma2** : The variance of the error terms of the longitudinal measurements. If covar = `exchange`, then `sigma2` is a $(G \times J \times 2)$ array. The $(g, j, 1)$th element is σ_{gj1}, and the $(g, j, 2)$th element is σ_{gj2}. If covar = `ind`, then `sigma2` is a $(G \times J)$ matrix. The (g, j)th element is σ_{gj1}; in this case, σ_{gj2} is fixed to be zero

- **xi** : A G-vector of class-specific variances of the latent variable
- **gamma** : A $(G \times (p_Z + q))$ matrix of class-specific regression parameters, consisting of 2 parts. The first p_Z columns correspond to regression parameters of the covariates `Z`, and the last q columns correspond to regression parameters of the spline functions, where $q =$ `nknots`$+$`degree`$+1$; that is, the gth row of `gamma` is $(\boldsymbol{\gamma}_g^T, \boldsymbol{a}_g^T)$
- **eta** : A G-vector of class-specific regression parameters of the random effect in the survival model
- **Tt** : A vector of ordered unique observed event times
- **Haz** : A $(t \times q)$ matrix of all estimated class-specific cumulative hazard function values at `Tt`, where t is the length of `Tt`
- **Bmat** : A $(t \times q)$ matrix of B-spline basis function values at `Tt`
- **post.prob** : Subject-specific posterior group probabilities
- **gridb** : An $(n \times G \times h)$ array of grid for the adaptive Gauss-Hermite quadrature. The (n, g)th row corresponds to the grid for the ith subject under the gth latent class
- **weightb** : An $(n \times G \times h)$ array of weight for the adaptive Gauss-Hermite quadrature. The (n, g)th row corresponds to the weight for the ith subject of the gth latent class
- **Information** : Information matrix; `NA` when `cal.inf` $=$ `FALSE`
- **loglike** : The log-likelihood value

Author(s)

Kin Yau (Alex) Wong <kin-yau.wong@polyu.edu.hk>

References

Varadhan, R. & Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. Scandinavian Journal of Statistics. 35 335–353.

Wong, K. Y., Zeng, D., & Lin, D. Y. (2022). Semiparametric latent-class models for multivariate longitudinal and survival data. The Annals of Statistics. 50 487–510.

See Also

`survival`

Examples

```
dataset <- create.data(n=1000)
result <- LatentClassJM(Y=dataset$Y,X=dataset$X,W=dataset$W,Time=dataset$Time,D=dataset$D,ni=dataset$ni,
Z=dataset$Z,G=4,nknots=2,degree=1,cal.inf=TRUE,init.param=NULL,bound=10,h=20,h2=20,covar="exchange")
```

Index