Package ‘HDJM’

September 2, 2023

Type         Package
Title        Penalized High-Dimensional Joint Model
Version      0.1.0
Author       Jiehuan Sun [aut, cre]
Maintainer   Jiehuan Sun <jiehuan.sun@gmail.com>
Description  Joint models have been widely used to study the associations between longitudinal biomarkers and a survival outcome. However, existing joint models only consider one or a few longitudinal biomarkers and cannot deal with high-dimensional longitudinal biomarkers. This package can be used to fit our recently developed penalized joint model that can handle high-dimensional longitudinal biomarkers. Specifically, an adaptive lasso penalty is imposed on the parameters for the effects of the longitudinal biomarkers on the survival outcome, which allows for variable selection. Also, our algorithm is computationally efficient, which is based on the Gaussian variational approximation method.

Depends      R (>= 3.6.0)
Imports      Rcpp (>= 1.0.0), survival(>= 3.2), statmod(>= 1.4)
LinkingTo    Rcpp, RcppArmadillo, RcppEnsmallen
License      GPL-2
Encoding     UTF-8
LazyData     true
RoxygenNote  7.2.1
NeedsCompilation yes
Repository    CRAN
Date/Publication 2023-09-02 08:00:02 UTC

R topics documented:

<table>
<thead>
<tr>
<th>control_list</th>
<th>HDJM_fit</th>
<th>LongData</th>
<th>SurvData</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
|              |          |          |          | 2
|              |          |          |          | 2
|              |          |          |          | 4
|              |          |          |          | 5
Index

control_list

Description

This list contains a list of parameters specifying the joint model.

Details

- ID_name the variable name for the patient ID in both longitudinal data and survival data.
- item_name the variable name for the longitudinal outcomes in the longitudinal data.
- value_name the variable name for the longitudinal measurements in the longitudinal data.
- time_name the variable name for the measurement timepoints in the longitudinal data.
- fix_cov a set of variables names indicating the covariates of fixed-effects in the longitudinal submodel. If NULL, not baseline covariates are included.
- random_cov a set of variables names indicating the covariates of random-effects in the longitudinal submodel. If NULL, not baseline covariates are included.
- FUN a function specifying the time-related basis functions in the longitudinal submodel.
- ran_time_ind a vector of integers specifying which time-related basis functions are also included with random-effects in the longitudinal submodel.
- surv_time_name the variable name for the survival time in the survival data.
- surv_status_name the variable name for the censoring indicator in the survival data.
- surv_cov a set of variables names specifying the baseline covariates in the survival submodel.
- n_points an integer indicating the number of nodes being used in the Gaussian quadrature.

Author(s)

Jiehuan Sun <jiehuan.sun@gmail.com>

HDJM_fit

The function to fit penalized HDJM.

Description

The function is used to fit the penalized HDJM with adaptive lasso penalty.
**Usage**

```r
HDJM_fit(
    LongData = NULL,
    SurvData = NULL,
    marker.name = NULL,
    control_list = NULL,
    nlam = 50,
    ridge = 0,
    pmax = 10,
    min_ratio = 0.01,
    maxiter = 100,
    eps = 1e-04,
    UseSurvN = FALSE
)
```

**Arguments**

- **LongData**: a data frame containing the longitudinal data (see `LongData`).
- **SurvData**: a data frame containing the survival data (see `SurvData`).
- **marker.name**: a vector indicating which set of longitudinal biomarkers to be analyzed. If `NULL`, all biomarkers in LongData will be used.
- **control_list**: a list of parameters specifying the joint model (see `control_list`).
- **nlam**: number of tuning parameters.
- **ridge**: ridge penalty.
- **pmax**: the maximum of biomarkers being selected. The algorithm will stop early if the maximum has been reached.
- **min_ratio**: the ratio between the largest possible penalty and the smallest penalty to tune.
- **maxiter**: the maximum number of iterations.
- **eps**: threshold for convergence.
- **UseSurvN**: a logical variable indicating whether the effective sample size (i.e., the number of events) should be used in calculating BIC.

**Value**

return a list with the following objects.

- **marker.name**: the names for biomarkers being analyzed.
- **alpha**: the estimates for the effects of biomarkers in the survival submodel.
- **weib**: the estimates for the Weibull baseline hazard in the survival submodel.
- **gamma**: the estimates for the effects of baseline covariates in the survival submodel.
- **beta**: the estimates for the fixed-effects in the longitudinal submodel.
- **sig2**: the estimates for the noise variances in the longitudinal submodel.
- **Sigma**: the estimates for the covariance matrices of the random effects in the longitudinal submodel.
References

Jiehuan Sun and Sanjib Basu. "Penalized Joint Models of High-Dimensional Longitudinal Biomarkers and A Survival Outcome".

Examples

data(HDJMdata)

flex_time_fun <- function(x=NULL){
  xx = matrix(x, ncol = 1)
  colnames(xx) = c("year_1")
  xx
}

ran_time_ind = 1 ## random time-trend effects
control_list = list(
  ID_name = "ID", item_name = "item",
  value_name = "value", time_name = "years",
  fix_cov = NULL, random_cov = NULL,
  FUN = flex_time_fun, ran_time_ind=ran_time_ind,
  surv_time_name = "ftime", surv_status_name = "fstat",
  surv_cov = "x", n_points = 5
)

## takes about one minute.
res = HDJM_fit(LongData=LongData, SurvData=SurvData,
               control_list=control_list)

LongData

Simulated Longitudinal Data

Description

This dataset contains longitudinal outcomes.

Usage

data(HDJMdata)

Format

A data frame with 48700 rows and 4 variables

Details

- ID patient ID
- item types of longitudinal outcome
- years measurement timepoints
- value measurements
SurvData 5

Author(s)
Jiehuan Sun <jiehuan.sun@gmail.com>

---

SurvData  Simulated Survival Data

Description
This dataset contains survival outcome.

Usage
data(HDJMdata)

Format
A data frame with 100 rows and 4 variables

Details
- ID patient ID
- fstat censoring indicator
- ftime survival time
- x baseline covariates

Author(s)
Jiehuan Sun <jiehuan.sun@gmail.com>
Index

* data
  control_list, 2
  LongData, 4
  SurvData, 5

control_list, 2, 3
HDJM_fit, 2
LongData, 3, 4
SurvData, 3, 5