Package 'ITRSelect'

September 24, 2018

Type Package

Title Variable Selection for Optimal Individualized Dynamic Treatment Regime

Version 1.0-1

Date 2018-09-23

Author Chengchun Shi, Ailin Fan, Rui Song and Wenbin Lu

Maintainer Chengchun Shi <cshi4@ncsu.edu>

Description Sequential advantage selection (SAS, Fan, Lu and Song, 2016) <arXiv:1405.5239> and penalized A-learning (PAL, Shi, et al., 2018) methods are implement for selecting important variables involved in optimal individualized (dynamic) treatment regime in both single-stage or multi-stage studies.

Imports Formula, Rglpk, ncvreg, Matrix, svMisc

License GPL-2

NeedsCompilation yes

Repository CRAN

Date/Publication 2018-09-24 04:20:10 UTC

R topics documented:

ITRSelect-package *Variable Selection for Optimal Individualized Dynamic Treatment Regime*

Description

Sequential advantage selection (SAS) and penalized A-learning (PAL) methods are implement for selecting important variables involved in optimal individualized (dynamic) treatment regime in both single-stage or multi-stage studies.

Details

Package:	ITRSelect
Type:	Package
Version:	1.0
Date:	2018-04-19
License:	GPL-2

Author(s)

Chengchun Shi, Ailin Fan, Rui Song, Wenbin Lu

Maintainer: Chengchun Shi <cshi4@ncsu.edu>

References

Shi, C. and Fan, A. and Song, R. and Lu, W. (2018) High-Dimensional A-Learing for Optimal Dynamic Treatment Regimes. *Annals of Statistics,* **46:** 925-957.

Fan, A. and Lu, W. and Song, R. (2016) Sequential Advantage Selection for Optimal Treatment Regime. *Annals of Applied Statistics,* **10:** 32-53.

Shi, C. and Song, R. and Lu, W. (2018) Concordance and Value Information Criteria for Optimal Treatment Decision. *Under review*.

PAL *Penalized A-learning for optimal dynamic treatment regime*

Description

Selects important variables that are involved in the optimal treatment regime based on penalized A-learning estimating equation. This function can be applied to two-stage studies where treatments are sequentially assigned at two different time points.

Usage

```
PAL(formula, data, subset, na.action, IC = c("BIC", "CIC", "VIC"),
   lambda.list = exp(seq(-3.5, 2, 0.1)), refit = TRUE, control = PAL.control(...),
model = TRUE, y = TRUE, a1 = TRUE, x1 = TRUE, a2 = TRUE, x2 = TRUE, ...)

PAL.fit(y, x1, x2 = NULL, a1, a2 = NULL, IC = c("BIC", "CIC", "VIC"),
    lambda.list = exp(seq(-3.5, 2, 0.1)), refit = TRUE,
    control = PAL.control())
```

Arguments

`formula`	A symbolic description of the model to be fitted(of type y ~ x1 | a1 or y ~ x1 | a1 | x2 | a2. Details are given 'Details').
`data`	An optional list or environment containing variables in `formula`.
`subset, na.action`	Arguments controlling formula processing via `model.frame`.
`IC`	Information criterion used in determining the regularization parameter. See 'Details'.
`lambda.list`	A list of regularization parameter values. Default is `exp(seq(-3.5, 2, 0.1))`.
`refit`	After variable selection, should the coefficients be refitted using A-learning estimating equation? Default is TRUE.
`control`	A list of control argument via `PAL.control`.
`model`	A logical value indicating whether *model frame* should be included as a component of the return value.
`y, a1, x1, a2, x2`	For `PAL`: logical values indicating whether the response, the first and second treatments, the baseline and intermediate covariates should be included as a component of the return value. For `PAL.fit`: y is the response vector (the larger the better), a1 and a2 are the first and second treatments patients receive, x1 and x2 are the design matrices consisting of patients' baseline covariates and intermediate covariates.
`...`	Argument passed to `PAL.control`.

Details

Penalized A-learning is developed to select important variables involved in the optimal individualized treatment regime. An individualized treatment regime is a function that maps patients covariates to the space of available treatment options. The method can be applied to both single-stage and two-stage studies.

PAL applied the Dantzig selector on the A-learning estimating equation for variable selection. The regularization parameter in the Dantzig selector is chosen according to the information criterion. Specifically, we provide a Bayesian information criterion (BIC), a concordance information criterion (CIC) and a value information criterion (VIC). For illustration of these information criteria, consider a single-stage study. Assume the data is summarized as $(Y_i, A_i, X_i), i = 1, ..., n$ where Y_i is the response of the i-th patient, A_i denotes the treatment that patient receives and X_i is the

corresponding baseline covariates. Let $\hat{\pi}_i$ and $\hat{h}_i$ denote the estimated propensity score and baseline mean of the i-th patient. For any linear treatment regime $I(x^T\beta > c)$, BIC is defined as

$$BIC = -n\log\left(\sum_{i=1}^{n}(A_i - \hat{\pi}_i)^2(Y_i - \hat{h}_i - A_i c - A_i X_i^T\beta)^2\right) - \|\beta\|_0\kappa_B,$$

where $\kappa_B = \{\log(n) + \log(p+1)\}/$`kappa` and `kappa` is the model complexity penalty used in the function `PAL.control`. VIC is defined as

$$VIC = \sum_{i=1}^{n}\left(\frac{A_i d_i}{\hat{\pi}_i} + \frac{(1-A_i)(1-d_i)}{1-\hat{\pi}_i}\right)\{Y_i - \hat{h}_i - A_i(X_i^T\beta + c)\} + \{\hat{h}_i + \max(X_i^T\beta + c, 0)\} - \|\beta\|_0\kappa_V,$$

where $d_i = I(X_i^T\beta > -c)$ and $\kappa_V = n^{1/3}\log^{2/3}(p)\log(\log(n))/$`kappa`. CIC is defined as

$$CIC = \sum_{i\neq j}\frac{1}{n}\left(\frac{(A_i - \hat{\pi}_i)\{Y_i - \hat{h}_i\}A_j}{\hat{\pi}_i(1-\hat{\pi}_i)\hat{\pi}_j} - \frac{(A_j - \hat{\pi}_j)\{Y_j - \hat{h}_j\}A_i}{\hat{\pi}_j(1-\hat{\pi}_j)\hat{\pi}_i}\right) I(X_i^T\beta > X_j^T\beta) - \|\beta\|_0\kappa_C,$$

where $\kappa_C = \log(p)\log_{10}(n)\log(\log_{10}(n))/$`kappa`.

Under certain conditions, it can be shown that CIC and VIC is consistent as long as either the estimated propensity score or the estimated baseline is consistent.

For single-stage study, the formula should specified as y ~ x1 | a1 where y is the reponse vector (y should be specified in such a way that a larger value of y indicates better clinical outcomes), x1 is patient's baseline covariates and a1 is the treatment that patient receives.

For two-stage study, the formula should be specified as y ~ x1 | a1 | x2 | a2 where y is the response vector, a1 and a2 the vectors of patients' first and second treatments, x1 and x2 are the design matrices consisting of patients' baseline covariates and intermediate covariates.

`PAL` standardizes the covariates and includes an intercept in the estimated individualized treatment regime by default. For single-stage study, the estimated treamtent regime is given by $I(\texttt{x1}^T\texttt{beta1.est} > 0)$. For two-stage study, the estimated regime is given by `a1` $= I(x1^T$`beta1.est` $> 0)$ and `a2` $= I(\texttt{x}^T\texttt{beta2.est} > 0)$ where `x=c(x1, a1, x2)`.

Value

`beta2.est`	Estimated coefficients in the second decision rule.
`beta1.est`	Estimated coefficients in the first decision rule.
`pi2.est`	Estimated propensity score at the second stage.
`pi1.est`	Estimated propensity score at the first stage.
`h2.est`	Estimated baseline function at the second stage.
`h1.est`	Estimated baseline function at the first stage.
`alpha2.est`	Regression coefficients in the estimated propensity score at the second stage.
`alpha1.est`	Regression coefficients in the estimated propensity score at the first stage.
`theta2.est`	Regression coefficients in the estimated baseline function at the second stage.
`theta1.est`	Regression coefficients in the estimated baseline function at the first stage.
`model`	The full model frame (if `model = TRUE`).

y	Response vector (if y = TRUE).
x1	Baseline covariates (if x1 = TRUE).
a1	A vector of first treatment (if a1 = TRUE).
x2	Intermediate covariates (if x2 = TRUE).
a2	A vector of second treatment (if a2 = TRUE).

Author(s)

Chengchun Shi and Ailin Fan

References

Shi, C. and Fan, A. and Song, R. and Lu, W. (2018) High-Dimensional A-Learing for Optimal Dynamic Treatment Regimes. *Annals of Statistics,* **46:** 925-957.

Shi, C. and Song, R. and Lu, W. (2018) Concordance and Value Information Criteria for Optimal Treatment Decision. *Under review*.

See Also

PAL.control

Examples

```
## single-stage study
set.seed(12345)
n <- 200
p <- 1000
X <- matrix(rnorm(n*p), nrow=n, ncol=p)
A <- rbinom(n, 1, 0.5)
CX <- (X[,1] + X[,2])
h <- 1 + X[,1] * X[,3]
Y <- h + A*CX + 0.5*rnorm(n)
result <- PAL(Y~X|A)

## two-stage study
set.seed(12345*2)
n <- 200
p <- 1000
X1 <- matrix(rnorm(n*p), nrow=n, ncol=p)
A1 <- rbinom(n, 1, 0.5)
X2 <- X1[,1] + A1 + 0.5*rnorm(n)
A2 <- rbinom(n, 1, 0.5)
Y <- A2*(A1 + X2) + A1*X1[,1] + 0.5*rnorm(n)
result <- PAL(Y~X1|A1|X2|A2)


## single-stage study
set.seed(12345)
n <- 50
```

```
p <- 20
X <- matrix(rnorm(n*p), nrow=n, ncol=p)
A <- rbinom(n, 1, 0.5)
CX <- (X[,1] + X[,2])
h <- 1 + X[,1] * X[,3]
Y <- h + A*CX + 0.5*rnorm(n)
result <- PAL(Y~X|A)

## two-stage study
set.seed(12345*2)
n <- 50
p <- 20
X1 <- matrix(rnorm(n*p), nrow=n, ncol=p)
A1 <- rbinom(n, 1, 0.5)
X2 <- X1[,1] + A1 + 0.5*rnorm(n)
A2 <- rbinom(n, 1, 0.5)
Y <- A2*(A1 + X2) + A1*X1[,1] + 0.5*rnorm(n)
result <- PAL(Y~X1|A1|X2|A2)
```

`PAL.control` *Control parameters for penalized A-learning*

Description

Parameters that control fitting of penalized A-learning.

Usage

```
PAL.control(pi1.est = NULL, pi2.est = NULL, h1.est = NULL, h2.est = NULL, kappa = NULL,
  penalty = 'SCAD')
```

Arguments

`pi1.est`	Estimated propentisy score at the first stage. By default, a penalized logistic regression model is fitted to estimate the propensity score.
`pi2.est`	Estimated propentisy score at the second stage. By default, a penalized logistic regression model is fitted to estimate the propensity score.
`h1.est`	Estimated baseline function at the first stage. By default, a penalized linear regression model is fitted to estimate the baseline function.
`h2.est`	Estimated baseline function at the second stage. By default, a penalized linear regression model is fitted to estimate the baseline function.
`kappa`	The model complexity penalty used in the information criteria. By default, $kappa = 1$ if BIC or CIC is used and $kappa = 4$ if VIC is used.
`penalty`	The penalty to be applied to the propensity score and baseline model. Either "MCP", "SCAD" (the default), or "lasso".

Value

A list with the arguments specified.

See Also

PAL, PAL.fit

Examples

```
set.seed(12345)
n <- 200
p <- 200
X <- matrix(rnorm(n*p), nrow=n, ncol=p)
A <- rbinom(n, 1, 0.5)
CX <- (X[,1] + X[,2])
h <- 1 + X[,1] * X[,3]
Y <- h + A*CX + 0.5*rnorm(n)
result <- SAS(Y~X|A, pi1.est=0.5)
```

SAS *Sequential advantage selection for optimal dynamic treatment regime*

Description

Select variables that are qualitatively interacted with the treatment based on a modified S-score method and a BIC-type criterion. This function can be applied to two-stage studies where treatments are sequentially assigned at two different time points.

Usage

```
SAS(formula, data, subset, na.action, step,
model = TRUE, y = TRUE, a1 = TRUE, x1 = TRUE, a2 = TRUE, x2 = TRUE, ...)

SAS.fit(y, x1, x2 = NULL, a1, a2 = NULL, step)
```

Arguments

formula	A symbolic description of the model to be fitted(of type y ~ x1 | a1 or y ~ x1 | a1 | x2 | a2. Details are given 'Details').
data	An optional list or environment containing variables in formula.
subset, na.action	Arguments controlling formula processing via model.frame.
step	SAS uses a forward selection procedure. The maximum size of the model is specified by step. By default, it is equal to $n/\log(n)$ where n is the sample size.

`model` A logical value indicating whether *model frame* should be included as a component of the return value.

`y, a1, x1, a2, x2`
For `SAS`: logical values indicating whether the response, the first and second treatments, the baseline and intermediate covariates should be included as a component of the return value.

For `SAS.fit`: `y` is the response vector (the larger the better), a1 and a2 are the first and second treatments patients receive, x1 and x2 are the design matrices consisting of patients' baseline covariates and intermediate covariates.

`...` Currently not used.

Details

For single-stage study, the formula should specified as y ~ x1 | a1 where y is the reponse vector (y should be specified in such a way that a larger value of y indicates better clinical outcomes), x1 is patient's baseline covariates and a1 is the treatment that patient receives.

For two-stage study, the formula should be specified as y ~ x1 | a1 | x2 | a2 where y is the response vector, a1 and a2 the vectors of patients' first and second treatments, x1 and x2 are the design matrices consisting of patients' baseline covariates and intermediate covariates.

The function returns linear dynamic treatment regimes. For single-stage study, the estimated treamtent regime for future patients is given by $I(\texttt{x1}^T\texttt{beta1.est} > 0)$. For two-stage study, the estimated regime is given by $\texttt{a1} = I(x1^T\texttt{beta1.est} > 0)$ and $\texttt{a2} = I(\texttt{x}^T\texttt{beta2.est} > 0)$ where `x=c(x1, a1, x2)`.

Value

`beta2.est`	Estimated coefficients in the second decision rule.
`beta1.est`	Estimated coefficients in the first decision rule.
`model`	The full model frame (if `model = TRUE`).
`y`	Response vector (if `y = TRUE`).
`x1`	Baseline covariates (if `x1 = TRUE`).
`a1`	A vector of first treatment (if `a1 = TRUE`).
`x2`	Intermediate covariates (if `x2 = TRUE`).
`a2`	A vector of second treatment (if `a2 = TRUE`).

Author(s)

Ailin Fan and Chengchun Shi

References

Fan, A. and Lu, W. and Song, R. (2016) Sequential Advantage Selection for Optimal Treatment Regime. *Annals of Applied Statistics,* **10:** 32-53.

Examples

```
## single-stage study
set.seed(12345)
n <- 200
p <- 200
X <- matrix(rnorm(n*p), nrow=n, ncol=p)
A <- rbinom(n, 1, 0.5)
CX <- (X[,1] + X[,2])
h <- 1 + X[,1] * X[,3]
Y <- h + A*CX + 0.5*rnorm(n)
result <- SAS(Y~X|A)

## two-stage study
set.seed(12345*2)
n <- 200
p <- 200
X1 <- matrix(rnorm(n*p), nrow=n, ncol=p)
A1 <- rbinom(n, 1, 0.5)
X2 <- X1[,1] + A1 + 0.5*rnorm(n)
A2 <- rbinom(n, 1, 0.5)
Y <- A2*(A1 + X2) + A1*X1[,1] + 0.5*rnorm(n)
result <- SAS(Y~X1|A1|X2|A2)


## single-stage study
set.seed(12345)
n <- 50
p <- 20
X <- matrix(rnorm(n*p), nrow=n, ncol=p)
A <- rbinom(n, 1, 0.5)
CX <- (X[,1] + X[,2])
h <- 1 + X[,1] * X[,3]
Y <- h + A*CX + 0.5*rnorm(n)
result <- SAS(Y~X|A)

## two-stage study
set.seed(12345*2)
n <- 50
p <- 20
X1 <- matrix(rnorm(n*p), nrow=n, ncol=p)
A1 <- rbinom(n, 1, 0.5)
X2 <- X1[,1] + A1 + 0.5*rnorm(n)
A2 <- rbinom(n, 1, 0.5)
Y <- A2*(A1 + X2) + A1*X1[,1] + 0.5*rnorm(n)
result <- SAS(Y~X1|A1|X2|A2)
```

SSTARD.onestage *Simulated single-stage dataset from the STAR*D study*

Description

This data is constructed based on the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study. Due to a data confidentiality agreement, we are not able to provide the original dataset. Here, we generate simulated data that are similar to the original dataset.

Usage

```
data("SSTARD.onestage")
```

Format

A list with 319 observations.

`Y` Patients' responses. The larger the better.

`A` Treatments patients receive.

`X` A 319*305 matrix consisting of patients' baseline covariates.

References

Fava, M., Rush, A. J., Trivedi, M. H., Nierenberg, A. A., Thase, M. E., Sack- eim, H. A., Quitkin, F. M., Wisniewski, S., Lavori, P. W., Rosenbaum, J. F. et al. (2003). Background and rationale for the Sequenced Treatment Alternatives to Relieve Depression (STAR* D) study. *Psychiatric Clinics of North America,* **26:** 457-494.

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sack- eim, H. A., Thase, M. E., Nierenberg, A. A., Quitkin, F. M., Kashner, T. M. et al. (2004). Sequenced treatment alternatives to relieve depression (STAR* D): rationale and design. *Controlled clinical trials,* **25:** 119-142.

Examples

```
data(SSTARD.onestage)
```

`SSTARD.twostage` *Simulated two-stage dataset from the STAR*D study*

Description

This data is constructed based on the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study. Due to a data confidentiality agreement, we are not able to provide the original dataset. Here, we generate simulated data that are similar to the original dataset.

Usage

```
data("SSTARD.twostage")
```

Format

A list with 73 observations.

`Y` Patients' responses. The larger the better.

`A2` Treatments patients receive at the second decision point.

`X2` The intermediate covariates collected between two decision points.

`A1` Treatments patients receive at the first decision point.

`X1` Patients' baseline covariates.

References

Fava, M., Rush, A. J., Trivedi, M. H., Nierenberg, A. A., Thase, M. E., Sack- eim, H. A., Quitkin, F. M., Wisniewski, S., Lavori, P. W., Rosenbaum, J. F. et al. (2003). Background and rationale for the Sequenced Treatment Alternatives to Relieve Depression (STAR* D) study. *Psychiatric Clinics of North America,* **26:** 457-494.

Rush, A. J., Fava, M., Wisniewski, S. R., Lavori, P. W., Trivedi, M. H., Sack- eim, H. A., Thase, M. E., Nierenberg, A. A., Quitkin, F. M., Kashner, T. M. et al. (2004). Sequenced treatment alternatives to relieve depression (STAR* D): rationale and design. *Controlled clinical trials,* **25:** 119-142.

Examples

```
data(SSTARD.twostage)
```

TR *Individualized treatment regime based on PAL or SAS.*

Description

Recommend individualized treatment regime for future patients, based on the penalized A-learning method, or sequential advantage selection method.

Usage

```
TR(object, x1, a1 = NULL, x2 = NULL, stage = 1)
```

Arguments

`object`	Fitted object of class "PAL" or "SAS".
`x1`	A matrix consisting of future patients baseline covariates.
`a1`	A vector consisting of future patients first treatments. Not needed if `stage = 2`.
`x2`	A matrix consisting of future patients intermediate covariates. Not needed if `stage = 2`.
`stage`	Outputs the first-stage decision rule for future patients if `stage = 1`. Otherwise, outputs the second-stage decision rule for future patients.

Value

A vector of individualized treatments tailored for future patients.

Author(s)

Chengchun Shi

See Also

PAL, SAS

Examples

```
## load simulated STARD data
data(SSTARD.twostage)
## estimate individualized treatment regime using SAS
result <- SAS(Y~X1|A1|X2|A2, data=SSTARD.twostage)
## make the recommendation
TR(result, x1=SSTARD.twostage$X1, a1=SSTARD.twostage$A1, x2=SSTARD.twostage$X2, stage=2)
```

Index