High-Dimensional Metrics

Victor Chernozhukov, Christian Hansen, Martin Spindler

2015-09-27

Overview

function	methods	comment
rlasso	summary, print, predict, model.matrix	default and formula interfac
rlogisticlasso	summary, print, predict, model.matrix	default and forumla interfac
rlassoLM	print, summary, confint, plot	default and formula (TBD) i
rlassoIV	print, summary, confint	default interface
rlassoMany	print, summary, confint	default interface
rlassoATE, rlassoLATE, rlassoATET, rlassoLATET	print, summary, confint	

To Do:

- rename functions
- choice of lambda
- joint CIs
- significance for logistic lasso
- wrap function for rlassoIV, rlassoMany, rlassoLM
- formula interface for rlassoIV, rlassoMany: design
- plots for all functions?
- further testing
- documentation and vignettes

Example

rlasso

Generating data:

```
set.seed(2)
n <- 100
p <- 20
px <- 5
X <- matrix(rnorm(n*p), ncol=p)
Xnew <- matrix(rnorm(n*p), ncol=p)
beta <- c(rep(2,px), rep(0,p-px))
y <- X %*% beta + rnorm(n)
```

rlasso

```
l1 <- rlasso(X,y, intercept=TRUE, normalize=TRUE, post=TRUE ) # default interface
l2 <- rlasso(y~X) # formula interface
```

methods for rlasso

```
print(l1)
```

```
##
## Call:
## rlasso.default(x = X, y = y, post = TRUE, intercept = TRUE, normalize = TRUE)
##
## Coefficients:
##    V1     V2     V3     V4     V5     V6     V7     V8     V9    V10
## 2.079  1.934  2.076  2.034  1.792  0.000  0.000  0.000  0.000  0.000
##   V11    V12    V13    V14    V15    V16    V17    V18    V19    V20
## 0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000
```

```
summary(l1, all=FALSE) # show only coefficient which are not zero
```

```
##
## Call:
## rlasso.default(x = X, y = y, post = TRUE, intercept = TRUE, normalize = TRUE)
##
## Post-Lasso Estimation:  TRUE
##
## Total number of variables: 20
## Number of selected variables: 5
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15819 -0.64003  0.03704  0.74504  2.12253
##
##    Estimate
## V1    2.079
## V2    1.934
## V3    2.076
## V4    2.034
## V5    1.792
##
## Residual standard error: 1.016
```

```
head(predict(l1))
```

```
##            [,1]
## [1,] -0.2739370
## [2,] -6.8242088
## [3,]  8.5583774
## [4,] -8.7417950
## [5,] -0.8918198
## [6,]  6.1379824
```

```
head(predict(l1, newdata=Xnew))
```

```
##            [,1]
```

```
## [1,] -8.1342684
## [2,]  3.0357902
## [3,]  0.7342197
## [4,]  7.8477857
## [5,]  1.5141399
## [6,] -2.0305342
```

rlogisticlasso

Generating data:

```
set.seed(2)
n <- 100
p <- 20
px <- 5
X <- matrix(rnorm(n*p), ncol=p)
Xnew <- matrix(rnorm(n*p), ncol=p)
beta <- c(rep(2,px), rep(0,p-px))
ystar <-  X %*% beta + rnorm(n)
y <- as.integer(ystar >=0)
```

rlogisticlasso

```
l3 <- rlogisticlasso(X,y, intercept=TRUE, normalize=TRUE, post=TRUE ) # default interface
l4 <- rlogisticlasso(y~X) # formula interface
```

methods for rlogisticlasso

```
#print(l3)
summary(l3, all=FALSE) # show only coefficient which are not zero
```

```
##
## Call:
## rlogisticlasso.default(x = X, y = y, post = TRUE, intercept = TRUE,
##     normalize = TRUE)
##
## Post-Lasso Estimation:  TRUE
##
## Total number of variables: 20
## Number of selected variables: 2
##
##    Estimate
## V1    1.471
## V3    1.888
```

```
head(predict(l3))
```

```
##           [,1]
## [1,] 0.2939060
## [2,] 0.1451806
## [3,] 0.9995185
```

```
## [4,] 0.2026698
## [5,] 0.1127571
## [6,] 0.6900044
```

```
head(predict(l3, newdata=Xnew))
```

```
##            [,1]
## [1,] 0.02658591
## [2,] 0.79455253
## [3,] 0.31579760
## [4,] 0.53468016
## [5,] 0.26317655
## [6,] 0.11540817
```

rlassoLM

Generating data:

```
set.seed(2)
n <- 100
p <- 20
px <- 5
X <- matrix(rnorm(n*p), ncol=p)
Xnew <- matrix(rnorm(n*p), ncol=p)
beta <- c(rep(2,px), rep(0,p-px))
y <- X %*% beta + rnorm(n)
```

rlassoLM

```
rLM <- rlassoLM(X,y, index=c(1,2,9)) # I: selection of variables for inference
```

```
## Warning in rlasso.default(x, d, ...): No variables selected!
```

```
## Warning in rlasso.default(x, d, ...): No variables selected!
```

```
## Warning in rlasso.default(x, d, ...): No variables selected!
```

methods

```
print(rLM)
```

```
##
## Call:
## rlassoLM.default(x = X, y = y, index = c(1, 2, 9))
##
## Coefficients:
##      V1       V2       V9
##  2.0786   1.9344  -0.1231
```

```r
summary(rLM)
```

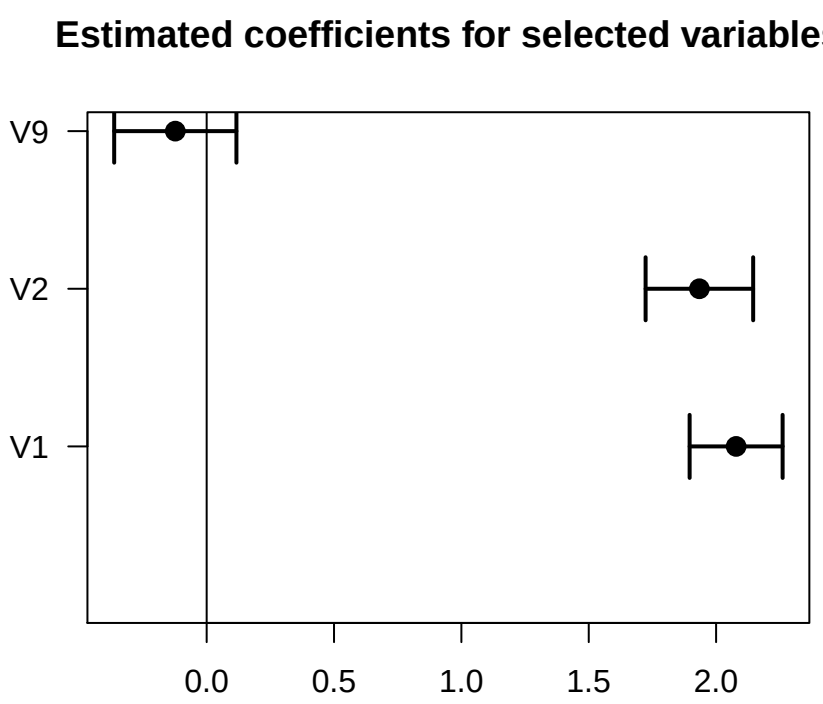
```
## [1] "Estimation of the effect of selected variables in a high-dimensional regression"
##      coeff.      se.  t-value p-value
## V1  2.07863  0.09196 22.60258   0.000
## V2  1.93439  0.10640 18.18037   0.000
## V9 -0.12315  0.12091 -1.01850   0.308
```

```r
confint(rLM)
```

```
##         2.5 %   97.5 %
## V1  1.8961064 2.261154
## V2  1.7232172 2.145566
## V9 -0.3631186 0.116826
```

```r
plot(rLM, main="Estimated coefficients for selected variables")
```

```
## [1] "Estimation of the effect of selected variables in a high-dimensional regression"
##      coeff.      se.  t-value p-value
## V1  2.07863  0.09196 22.60258   0.000
## V2  1.93439  0.10640 18.18037   0.000
## V9 -0.12315  0.12091 -1.01850   0.308
```




```
## NULL
```