A comparison of mean squared error using ridge regression

CAI Mingxuan

2021/5/26

We first simulate a case with relatively small n.

```
library(mvtnorm)
library(VCM)
library(data.table)

set.seed(10)
ntest <- 1000
n <- 1000
p <- 500

sb <- 0.5
se <- 1-sb


Xtest <- matrix(rnorm(ntest*p),ntest,p)
X <- matrix(rnorm(n*p),n,p)

Xstest <- scale(Xtest)/sqrt(p)
Xs <- scale(X)/sqrt(p)


beta <- rnorm(p,0,sqrt(sb))

y <- Xs%*%beta + rnorm(n,0,sqrt(se))
ytest <- Xstest%*%beta + rnorm(ntest,0,sqrt(se))

fit_cv <- cvperf.linReg(X,y)
```

```
## Info: Algorithm used: PXEM

## Info: Number of variables: 500

## Info: Sample size: 1000

## Info: Number of cv folds: 10

## start cv process......... total 10 validation sets
## 1 -th validation set...
## 2 -th validation set...
## 3 -th validation set...
## 4 -th validation set...
## 5 -th validation set...
## 6 -th validation set...
## 7 -th validation set...
```

```
## 8 -th validation set...
## 9 -th validation set...
## 10 -th validation set...
```

```
fit_ridge <- linRegPXEM(X,y)
```

```
## 2 -th iteration, lower bound =  -1312.221  ,diff= Inf ,sb2= 0.4892107 ,se2= 0.4892107
## 3 -th iteration, lower bound =  -1312.207  ,diff= 0.01406048 ,sb2= 0.4911115 ,se2= 0.4923357
## 4 -th iteration, lower bound =  -1312.205  ,diff= 0.002057076 ,sb2= 0.4902287 ,se2= 0.4933856
## 5 -th iteration, lower bound =  -1312.204  ,diff= 0.0006329769 ,sb2= 0.489528 ,se2= 0.4939059
## 6 -th iteration, lower bound =  -1312.204  ,diff= 0.0002026886 ,sb2= 0.4891109 ,se2= 0.494193
## 7 -th iteration, lower bound =  -1312.204  ,diff= 6.504093e-05 ,sb2= 0.4888726 ,se2= 0.4943549
## 8 -th iteration, lower bound =  -1312.204  ,diff= 2.088173e-05 ,sb2= 0.4887373 ,se2= 0.4944466
## 9 -th iteration, lower bound =  -1312.204  ,diff= 6.705982e-06 ,sb2= 0.4886607 ,se2= 0.4944986
## 10 -th iteration, lower bound =  -1312.204  ,diff= 2.153893e-06 ,sb2= 0.4886172 ,se2= 0.494528
## 11 -th iteration, lower bound =  -1312.204  ,diff= 6.918674e-07 ,sb2= 0.4885926 ,se2= 0.4945447
```

```
mse_cv <- fit_cv$cvm
mse_train <- mean((y-predict(fit_ridge,X))^2)
mse_test <- mean((ytest-predict(fit_ridge,Xtest))^2)

cat("Cross-validation error:",mse_cv)
```

```
## Cross-validation error: 0.7203834
```

```
cat("Training error:",mse_train)
```

```
## Training error: 0.3504182
```

```
cat("Testing error error:",mse_test)
```

```
## Testing error error: 0.7734115
```

Obviously, the cv error is close to the test error. In contrast, the training error is much smaller.

We than consider a case with large n.

```
set.seed(10)
ntest <- 1000
n <- 5000
p <- 500

sb <- 0.5
se <- 1-sb


Xtest <- matrix(rnorm(ntest*p),ntest,p)
X <- matrix(rnorm(n*p),n,p)

Xstest <- scale(Xtest)/sqrt(p)
Xs <- scale(X)/sqrt(p)


beta <- rnorm(p,0,sqrt(sb))

y <- Xs%*%beta + rnorm(n,0,sqrt(se))
ytest <- Xstest%*%beta + rnorm(ntest,0,sqrt(se))
```

```r
fit_cv <- cvperf.linReg(X,y)
```

```
## Info: Algorithm used: PXEM
## Info: Number of variables: 500
## Info: Sample size: 5000
## Info: Number of cv folds: 10
## start cv process......... total 10 validation sets
## 1 -th validation set...
## 2 -th validation set...
## 3 -th validation set...
## 4 -th validation set...
## 5 -th validation set...
## 6 -th validation set...
## 7 -th validation set...
## 8 -th validation set...
## 9 -th validation set...
## 10 -th validation set...
```

```r
fit_ridge <- linRegPXEM(X,y)
```

```
## 2 -th iteration, lower bound =  -5887.354  ,diff= Inf ,sb2= 0.4716599 ,se2= 0.4716599
## 3 -th iteration, lower bound =  -5885.722  ,diff= 1.631567 ,sb2= 0.463682 ,se2= 0.4878942
## 4 -th iteration, lower bound =  -5885.704  ,diff= 0.0184726 ,sb2= 0.4617289 ,se2= 0.4895746
## 5 -th iteration, lower bound =  -5885.704  ,diff= 0.0002287485 ,sb2= 0.4614969 ,se2= 0.4897603
## 6 -th iteration, lower bound =  -5885.704  ,diff= 2.849994e-06 ,sb2= 0.4614708 ,se2= 0.489781
## 7 -th iteration, lower bound =  -5885.704  ,diff= 3.552941e-08 ,sb2= 0.4614679 ,se2= 0.4897833
```

```r
mse_cv <- fit_cv$cvm
mse_train <- mean((y-predict(fit_ridge,X))^2)
mse_test <- mean((ytest-predict(fit_ridge,Xtest))^2)

cat("Cross-validation error:",mse_cv)
```

```
## Cross-validation error: 0.5397776
```

```r
cat("Training error:",mse_train)
```

```
## Training error: 0.4459211
```

```r
cat("Testing error error:",mse_test)
```

```
## Testing error error: 0.5615097
```

With large sample size, the cv error, traing error, and the test error all converge to the true residual variance.