Numerical Calibration of the Lasso

ENS Advanced Math, Non parametrics

Franck Picard, Fall 2020

Preliminaries

Groups of students are asked to send an `R markdown` report generated via `R studio` to franck.picard@ens-lyon.fr at the end of the tutorial. You will need `Rstudio`, LaTeX and packages for markdown:

```
library(knitr)
library(rmarkdown)
library(graphics) ##Necessary for plots
library(pracma)
library(glmnet)
library(rgl)
```

This report should answer the questions by commentaries and codes generating appropriate graphical outputs. A cheat sheet of the markdown syntax can be found here.

1 Regression model

This project aims at studying the empirical properties of the LASSO based on simulated data. The statistical framework is the linear regression model, such that

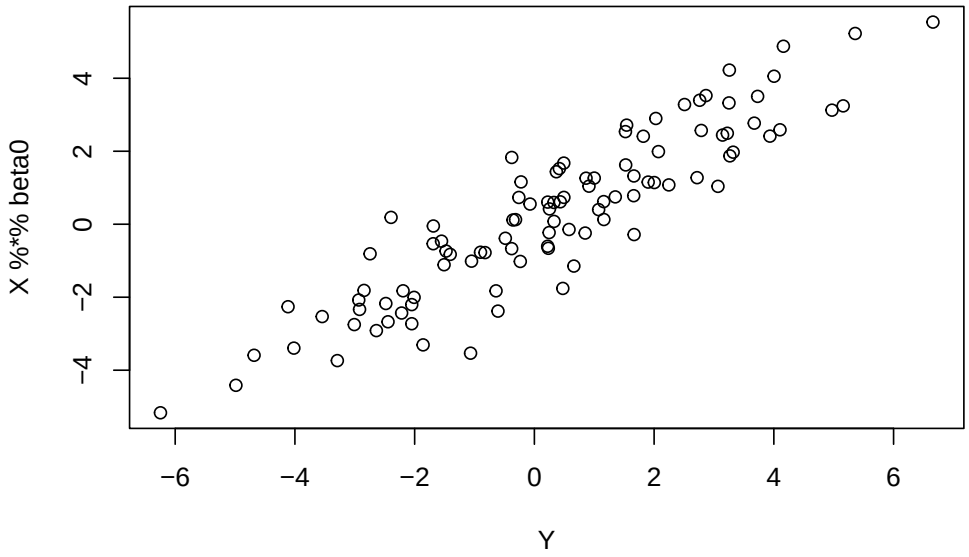
$$Y_i = x_i^T \beta^* + \varepsilon_i, \ \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

with Y a vector in $\mathbb{R}^n$, and β^* a vector in $\mathbb{R}^p$ with p_0 non-null elements. In the following $J_0 = \{j \in \{1, ..., p\}, \beta_j^* \neq 0\}$. For simplification, we will consider that there is only one distinct non null value in β^*: $\beta^* = \beta_0^* \times (1, ..., 1, 0, ...0)$.

2 Simulation of observations

```
n       = 100
p       = 10
p0      = 5
sigma   = 1
sigmaX  = 1
b0      = 1
beta0   = c(rep(b0,p0),rep(0,p-p0))
```

```r
X       = sapply(1:p, FUN=function(x){rnorm(n,0,sigmaX)})
Y       = X%*%beta0 + rnorm(n,0,sigma)

plot(Y,X%*%beta0)
```



3 Lasso and the `glmnet` R-package

In practice, model parameters are estimated using the `glmnet` R-package to compute the lasso estimator

$$\widehat{\beta}_\lambda = \min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_0 + \beta^T x_i) + \lambda \left[(1-\alpha)||\beta||_2^2/2 + \alpha||\beta||_1 \right],$$

with $\alpha = 1$ for the Lasso, $\alpha = 0$ for Ridge Regression and $\alpha \in]0,1[$ for Elastic Net. In a first step you are invited to check the online documentation of the package that is very complete. Then the purpose of calibration is to determine the value of the hyperparameter λ based on the observations.

```r
library(glmnet)
n       = 100
p       = 10
p0      = 5
sigma   = 1
sigmaX  = 1
b0      = 1
beta0   = c(rep(b0,p0),rep(0,p-p0))
X       = sapply(1:p, FUN=function(x){rnorm(n,0,sigmaX)})
```

```
Y       = X%*%beta0 + rnorm(n,0,sigma)

fit = glmnet(X, Y)
plot(fit,label=TRUE)
names(fit)
print(fit)
coef(fit)
```

4 Numerical Calibration in practice

Parameter λ is chosen by cross validation using `cv.glmnet` such that:

```
library(glmnet)
n       = 100
p       = 10
p0      = 5
sigma   = 1
sigmaX  = 1
b0      = 1
beta0   = c(rep(b0,p0),rep(0,p-p0))
X       = sapply(1:p, FUN=function(x){rnorm(n,0,sigmaX)})
Y       = X%*%beta0 + rnorm(n,0,sigma)

lambda.cv = cv.glmnet(X,Y, family = "gaussian",intercept=F)$lambda.1se
bh        = glmnet(X,Y,family = "gaussian",intercept=F,
                   lambda=lambda.cv)$beta
if ( sum(abs(bh))==0 ) {bh = rep(0,p)}
bh        = as.vector(bh)
```

The first part of your projet will be to re-implement the cross validation procedure, and to verify that your implementation is correct based on the `cv.glmnet` function that will be used to check your results. You will also implement the calibration of λ based on the AIC and on the BIC.

5 Simulations setting

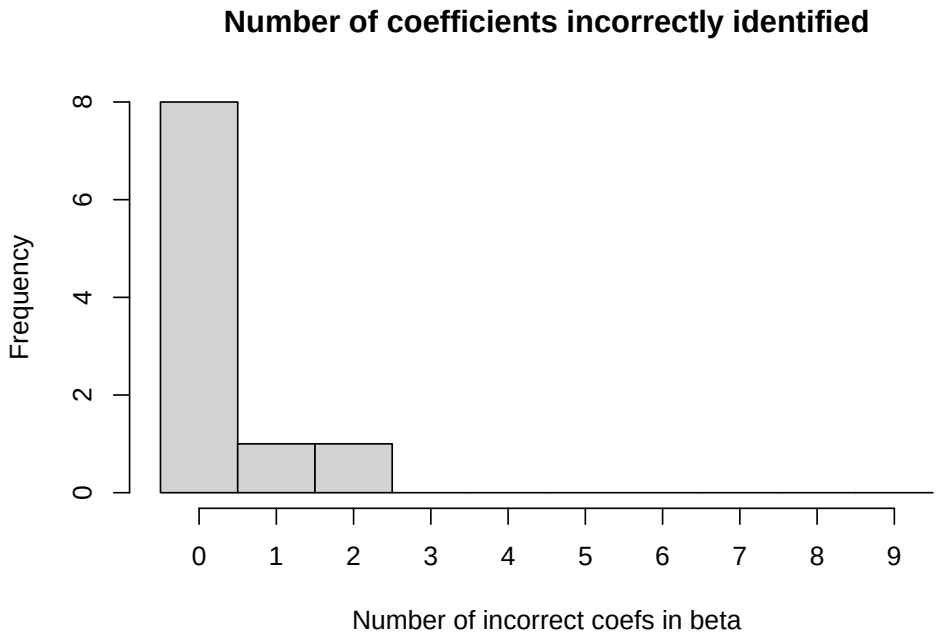
The performance of the lasso depends on different factors, and numerical simulations are used to study the impact of these factors on the capacity of the lasso to select the dimension of the model. Among those factors, we can identify n (number of observations), p (dimensionality), p_0 and β^* (strength of the signal), σ (strength of noise). Studying the impact of all factors would not be realistic, so we focus on:

- n/p will be chosen so that we study the "low-dimensional regime" as well as the "high-dimensional regime".
- p_0 will be fixed at 5
- $\beta_0^* = 1$
- σ will vary.

As an indicator of performance, we will study only the dimension of the selected model, ie $\hat{s} = \sum_{j=1}^{p} 1_{\hat{\beta}_j \neq 0}$. In order to conduct your simulations, you will start from the following example code:

```
n      = 100
p      = 10
p0     = 5  #Number of nonzero coefficients
sigma  = 1
sigmaX = 1
b0     = 1
beta0  = c(rep(b0,p0),rep(0,p-p0))
B      = 10 #Number of simulations
res    = matrix(NA,ncol=5,nrow=B) #Stores n,p, sigma, number of nonzero
                  #coefficients recovered, number of nz incorrectly identified

betah  = matrix(NA, ncol = p, nrow = B)

# fixed design setting
X = sapply(1:p, FUN=function(x){rnorm(n,0,sigmaX)})

for (b in 1:B){
  Y     = X%*%beta0 + rnorm(n,0,sigma)
  # estimate betah using the calibrated lasso
  # betah =
  lambda.cv = cv.glmnet(X,Y, family = "gaussian",intercept=F)$lambda.1se
  temp = glmnet(X,Y,family = "gaussian",intercept=F, lambda=lambda.cv)$beta
  ##the value returned by glmnet(...)$beta is a sparse matrix
  #In a sparse matrix : @i (non zero indices); @x non zero values

  indices_identified = c(rep(0, p)) #Fill indices_identified with 1 at
                                    #indices in temp@i, set others to 0
  for (index in temp@i) {
    indices_identified[index] = 1
  }

  betah[b,]  = c(temp@i, rep(NA, p-length(temp@i))) #We store in betah[b,]
    #the nonzero indices and we fill with NA to get a vector of the proper size
  res[b,] = c(n, p, sigma, length(temp@i), sum(abs(indices_identified - beta0)))
}
res           = as.data.frame(res)
colnames(res) = c("n","p","sigma","nz", "nzcorrect")


hist1<-hist(res$nz, breaks = 1:10, axes = FALSE,
            xlab="Number of non-zero coefs in beta",
            main = "Number of coefficients identified")
axis(side=1,at=hist1$mids,labels=1:9) #So that breakpoints are centered
axis(side=2)          #Because of previous trick, we had to hide both axis,
```


```
                    #so we display again the Y-axis


hist2 <- hist(res$nzcorrect, breaks = 0:10, axes = FALSE,
              xlab="Number of incorrect coefs in beta",
              main = "Number of coefficients incorrectly identified")
axis(side=1,at=hist2$mids,labels=0:9)
axis(side=2)
```



5.1 AIC

In this section, we implement the Akaike Information Criterion (AIC):

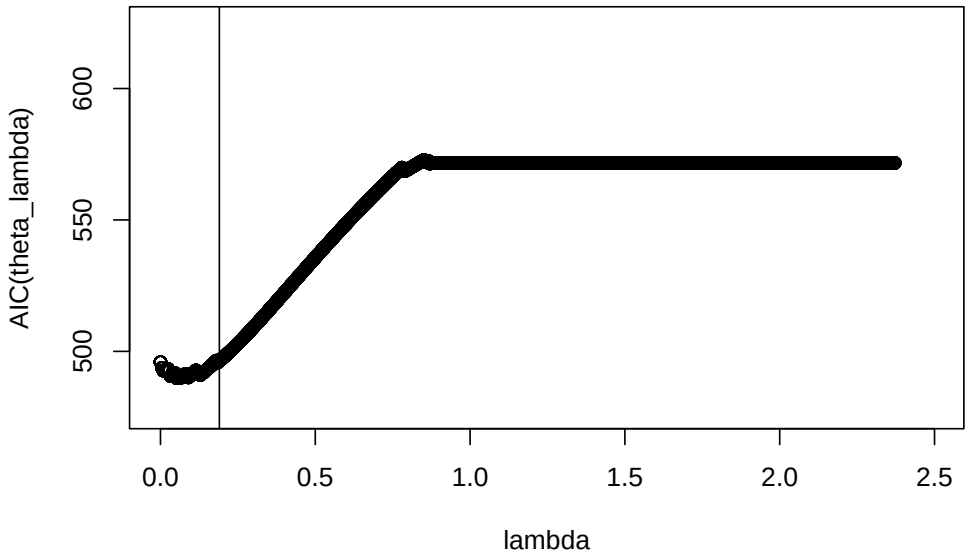
$$AIC = 2k - 2ln(\hat{L})$$

where $\hat{L}$ is the maximum value of the likelihood function for the model under consideration. The procedure consists of computing the Lasso with parameters $\lambda_1, \cdots, \lambda_N$.

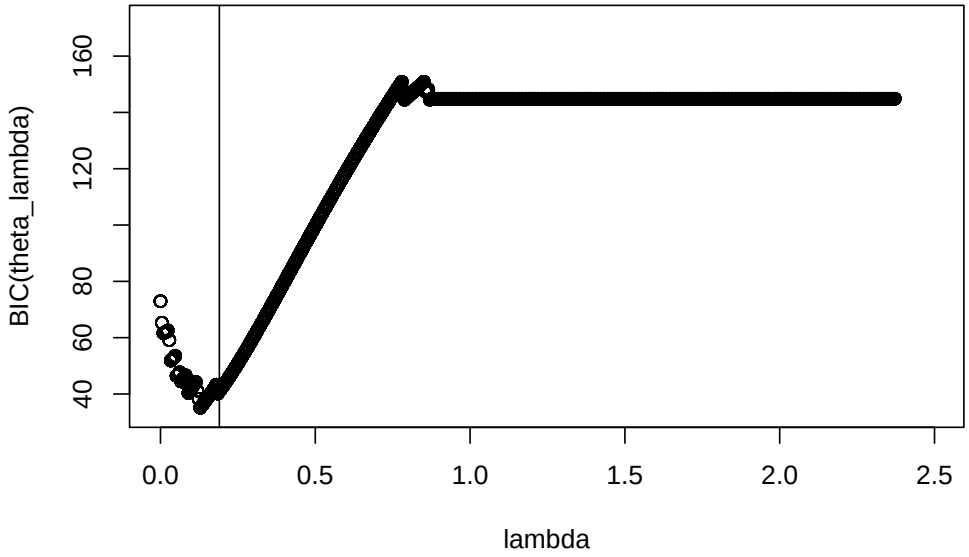
```
likelihood <- function(betal, Yl, sigma2 = 1.0){ ##Not used
  S<-0.0
  S <- S + t(Yl-betal)%*%(Yl-betal)
  n<- length(Yl)

  return (1/sqrt(2 * pi* sigma2))**(n/2)*exp(-S/(2*sigma2))
}



n      = 100
p      = 20
p0     = 5  #Number of nonzero coefficients
sigma  = 1
sigmaX = 1
b0     = 1
beta0  = c(rep(b0,p0),rep(0,p-p0))
betah  = matrix(NA, ncol = p, nrow = B)
X = sapply(1:p, FUN=function(x){rnorm(n,0,sigmaX)})
```

```
Y     = X%*%beta0 + rnorm(n,0,sigma)



N <- 500 #Number of lambda tested
lambda_step <- 0.005 #Difference between two consecutive lambda
lambda_init <- 0.0 #First value used by lambda

lambda <- lambda_init
results_frame <- data.frame(
  lambda = c(),
  betah = c(),
  k = c(),
  loglikelihood = c(),
  AIC = c(),
  BIC = c()
)

for (b in 1:N){#The increment is lambda
  temp = glmnet(X,Y,family = "gaussian",intercept=F, lambda=lambda)$beta
  ##the value returned by glmnet(...)$beta is a sparse matrix
  #In a sparse matrix : @i (non zero indices); @x non zero values

  betah <- c(rep(0,p))
  j<-0
  for(i in temp@i){
    betah[i+1] <- temp@x[j+1]
    j<-j+1
  }

  #llik = log(((1/sqrt(2 * pi))**(p/2))*exp(-t(Y-betah)%*%(Y-betah)/(2)))
  RSS = t(Y-X%*%betah)%*%(Y-X%*%betah) #Simpler expression for log likelihood in Gauss
  new_row_df <- data.frame(
    lambda = lambda,
    betah = betah,
    k = j,
    negloglikelihood = -log(RSS),
    AIC = 2*j + n*log(RSS),
    BIC = j*log(n) + n*log(RSS/n)
  )
  results_frame <- rbind(results_frame, new_row_df)

  lambda <- lambda+lambda_step

}
lambda_cv = cv.glmnet(X,Y, family = "gaussian",intercept=F)$lambda.1se  #We compute th
s <- c(rep(lambda_cv, 1000))  #And fill an axis with it
```

```
axis_lambda_cv <- xy.coords(x=s, y = 1:1000)  #that we will give to the plot

plot(x = results_frame$lambda, y = results_frame$AIC, type = "n", xlab = "lambda", yla
plot.xy(axis_lambda_cv, type = "l") #type = "l" to draw lines
plot.window(c(0.95*lambda_init, 1.05*lambda ), ylim = c(0.95*min(results_frame$AIC), 1
plot.xy( list(results_frame$lambda, results_frame$AIC, NULL, NULL), type="p")
```




```
plot(x = results_frame$lambda, y = results_frame$BIC, type = "n", xlab = "lambda", yla
plot.xy(axis_lambda_cv, type = "l") #type = "l" to draw lines
plot.window(c(0.95*lambda_init, 1.05*lambda ), ylim = c(0.95*min(results_frame$BIC), 1
plot.xy( list(results_frame$lambda, results_frame$BIC, NULL, NULL), type="p")
```



We then define the best value for λ with respect to AIC as:

$$\lambda_{AIC} = argmin_{\lambda_i} AIC.$$

Hence, for each i we need to compute the likelihood of the model with parameter β_{λ_i} and k as the number of nonzero coefficients of β_{λ_i}. The likelihood, with Gaussianity assumption and covariance matrix σI, is given by

$$\hat{L_{\lambda_i}} = \Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(Y_i - \beta)^t(Y_i - \beta)$$

. In our setting, we compute $S = \sum_i (Y_i - \beta_\lambda)^2$ and then

$$L_Y(\lambda) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp(-\frac{1}{2\sigma^2}S)$$

For each instance of $Y = X\theta^* + \epsilon$ for given parameters p/n(dimensionality)\$ of $k = s * (p/n)$ (sparsity), while the others are fixed; we obtain lambda_AIC, lambda_CV, lambda_BIC,.... the optimal calibration of lasso for a same problem and a given criterion.

We want to know given the dimensionality and perhaps the sparsity which criterion is optimal. For each instance of $Y = X\theta^*$ we may define

$$\lambda_{opt} = Argmin_\lambda ||X\theta^* - X\hat{\theta}_\lambda||_2$$

the "optimal" value of lambda, in the sense that given the solution θ^* (unknown in practice), we look for the lambda that would give the model with the smallest error (where the argmin could run through all lambdas or only on the one given by our criterions). Note that the choice of definition of λ_{opt} is arbitrary, replacing the

norm by an $\ell^1, \ell^\infty, \cdots$ can be justified and the argmin could ignore the dependence on X.

Idea : modelling λ_{opt} as a function of the dimensionality, the sparsity, the different criterions Naive idea: Define the norm in the argmin as some sup norm so that $hatlambda_{opt}$ is the criterion that gives the lambda of the criterion which is closest from sparsity estimate. Improved: A nearest-neighbour Improved (bis): Model $\hat{\lambda}_{opt}$ as some linear combination of $p/n, s, \lambda_{AIC}, \lambda_{BIC}, \ldots$. Improved (ter): Model as a polynomial/log linear of p/n and s (and $\lambda_{AIC}, \lambda_{Bic}, \cdots$)

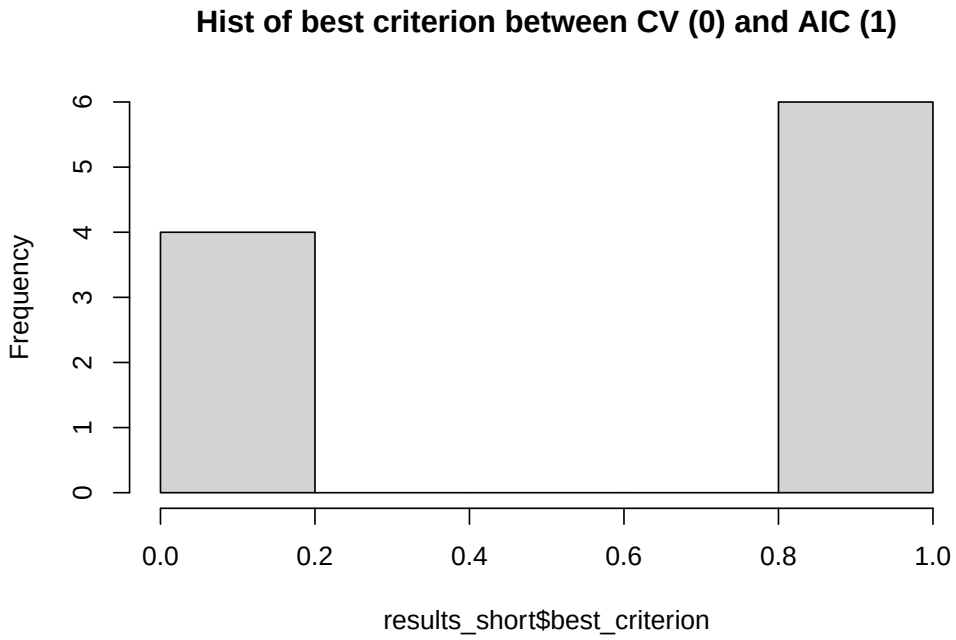
```
##We fix p, n and k
#We run M instances of Y = X\theta + eps
#Each time we obtain $\lambda_{AIC, CV, opt}$ and we create a list which contains for
#We plot the histogram to get an idea if AIC or CV is best in our setup
#TODO : same thing, but instead pick an index whenever it is the one that gives the $\
n      = 100
p      = 10
p0     = 5  #Number of nonzero coefficients
sigma  = 1
sigmaX = 1
b0     = 1
beta0  = c(rep(b0,p0),rep(0,p-p0))
betah  = matrix(NA, ncol = p, nrow = B)


B <- 20 #Number of lambda tested
lambda_step <- 0.1 #Difference between two consecutive lambda
lambda_init <- 0.0 #First value used by lambda


results_short <- data.frame(
  lambda_AIC =c(),
  lambda_CV = c(),
  lambda_opt = c(),
  best_criterion = c()
)

N<- 10 ## Number of instances of solving Y=X\theta + ups
for(t in 1:N){

  X = sapply(1:p, FUN=function(x){rnorm(n,0,sigmaX)})
  Y = X%*%beta0 + rnorm(n,0,sigma)

  lambda <- lambda_init
  results_frame <- data.frame(
    lambda = c(),
    betah = c(),
    k = c(),
```

```
  loglikelihood = c(),
  AIC = c(),
  BIC = c()
)

err_opt <- NA
lambda_opt <- 1

for (b in 1:B){#The increment is lambda
  temp = glmnet(X,Y,family = "gaussian",intercept=F, lambda=lambda)$beta
  ##the value returned by glmnet(...)$beta is a sparse matrix
  #In a sparse matrix : @i (non zero indices); @x non zero values

  betah <- c(rep(0,p))
  j<-0
  for(i in temp@i){
    betah[i+1] <- temp@x[j+1]
    j<-j+1
  }

  #llik = log(((1/sqrt(2 * pi))**(p/2))*exp(-t(Y-betah)%*%(Y-betah)/(2)))
  RSS = t(Y-X%*%betah)%*%(Y-X%*%betah) #Simpler expression for log likelihood in Ga
  new_row_df <- data.frame(
    lambda = lambda,
    betah = betah,
    k = j,
    negloglikelihood = -log(RSS),
    AIC = 2*j + n*log(RSS),
    BIC = j*log(n) + n*log(RSS/n)
  )
  results_frame <- rbind(results_frame, new_row_df)
  if(err_opt >= Norm(betah - beta0, p=2) || is.na(err_opt)){
    lambda_opt <- lambda   #Obtaining the lambda which gives the solution closest fr
    err_opt <- Norm(betah - beta0, p = 2)
  }
  lambda <- lambda+lambda_step
}
esc <-1
lambda_min_AIC <- results_frame$AIC[0]
min_lambda_AIC <- min(results_frame$AIC)
i<-1
while (esc != 0 || i < B) {
  if( min_lambda_AIC >= results_frame$AIC[i]){
    esc = 0
    lambda_min_AIC <-results_frame$lambda[i]  #Obtaining the optimal value of lambda
  }
  i<- i + 1
```

```
  }


  new_row_short_df <- data.frame(
    lambda_AIC = lambda_min_AIC,
    lambda_CV = cv.glmnet(X,Y, family = "gaussian",intercept=F)$lambda.1se,
    lambda_opt = lambda_opt,
    best_criterion = NA
  )
  if(abs(new_row_short_df$lambda_AIC - new_row_short_df$lambda_opt) > abs(new_row_shor
    new_row_short_df$best_criterion = 0 ##0 : best criterion is CV
  }
  else{
    new_row_short_df$best_criterion = 1  ##1: best criterion is AIC
  }
  results_short <- rbind(results_short, new_row_short_df)
  print(t)

}
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
```

```
hist(results_short$best_criterion, main = "Hist of best criterion between CV (0) and A
```




```
###We wrap the previous code in a function
#Then we compute for varying $n/p$ whether best is AIC, BIC or CV

GetLambdaComparison <- function(n = 100,
                                p      = 10,
                                p0     = 5,  #Number of nonzero coefficients
                                sigma  = 1,
                                sigmaX = 1,
                                b0     = 1,
                                B = 20, #Number of lambda tested
                                N = 10, #Number of instances of solving Y = X\theta +
                                lambda_step = 0.1, #Difference between two consecutive
                                lambda_init = 0.0 #First value used by lambda
                                ){

  if(p < p0 ){##We treat this (impossible) case so that we can obtain dataframes of eq
    results_short <- data.frame(
      lambda_AIC =NA,
      lambda_BIC = NA,
      lambda_CV = NA,
      lambda_opt = NA,
      best_criterion = NA
    )
    # results_short_row <- data.frame(
    #   lambda_AIC =NA,
    #   lambda_BIC = NA,
    #   lambda_CV = NA,
```

```
  #   lambda_opt = NA,
  #   best_criterion = NA
  # )
  # results_short <- rbind(results_short, results_short_row)

  return(results_short)
}

beta0  = c(rep(b0,p0),rep(0,p-p0))
betah  = matrix(NA, ncol = p, nrow = B)

results_short <- data.frame(
  lambda_AIC =c(),
  lambda_BIC = c(),
  lambda_CV = c(),
  lambda_opt = c(),
  best_criterion = c()
)

if(n<30){
  print("Warning, if n < 30, cv will start throwing warnings")
}

for(t in 1:N){ ##We do the simulation N times

  X = sapply(1:p, FUN=function(x){rnorm(n,0,sigmaX)})
  Y = X%*%beta0 + rnorm(n,0,sigma)

  lambda <- lambda_init
  results_frame <- data.frame(
    lambda = c(),
    betah = c(),
    k = c(),
    loglikelihood = c(),
    AIC = c(),
    BIC = c()
  )

  err_opt <- NA
  lambda_opt <- 1
  min_lambda_AIC <- NA
  lambda_AIC <- NA
  min_lambda_BIC <- NA
  lambda_BIC <- NA
  for (b in 1:B){#The increment is lambda
    temp = glmnet(X,Y,family = "gaussian",intercept=F, lambda=lambda)$beta
    ##the value returned by glmnet(...)$beta is a sparse matrix
```

```
    #In a sparse matrix : @i (non zero indices); @x non zero values

    betah <- c(rep(0,p))
    j<-1
    for(i in temp@i){
      betah[i+1] <- temp@x[j]
      j<-j+1
    }

    #llik = log(((1/sqrt(2 * pi))**(p/2))*exp(-t(Y-betah)%*%(Y-betah)/(2)))
    RSS = t(Y-X%*%betah)%*%(Y-X%*%betah) #Simpler expression for log likelihood in G
    new_row_df <- data.frame(
      lambda = lambda,
      betah = betah,
      k = j,
      negloglikelihood = -log(RSS),
      AIC = 2*j + n*log(RSS),
      BIC = j*log(n) + n*log(RSS/n)
    )
    results_frame <- rbind(results_frame, new_row_df)

    ##Before restarting to loop we check for the best value for lambda_opt
    if(err_opt >= Norm(betah - beta0, p=2) || is.na(err_opt)){
      lambda_opt <- lambda   #Obtaining the lambda which gives the solution closest
      err_opt <- Norm(betah - beta0, p = 2)
    }

    #And we also search for the best value for lambda under AIC
    if(min_lambda_AIC > new_row_df$AIC || is.na(min_lambda_AIC)){
      min_lambda_AIC <- new_row_df$AIC
      lambda_AIC <- lambda
    }
    #And we also search for the best value for lambda under BIC
    if(min_lambda_BIC > new_row_df$BIC || is.na(min_lambda_BIC)){
      min_lambda_BIC <- new_row_df$BIC
      lambda_BIC <- lambda
    }
    lambda <- lambda+lambda_step
  }##End of for loop for finding best lambda (depending on criterion)


  ##We create a frame in which we store the best values obtained
  new_row_short_df <- data.frame(
    lambda_AIC = lambda_AIC,
    lambda_BIC = lambda_BIC,
    lambda_CV = cv.glmnet(X,Y, family = "gaussian",intercept=F)$lambda.1se,
    lambda_opt = lambda_opt,
```

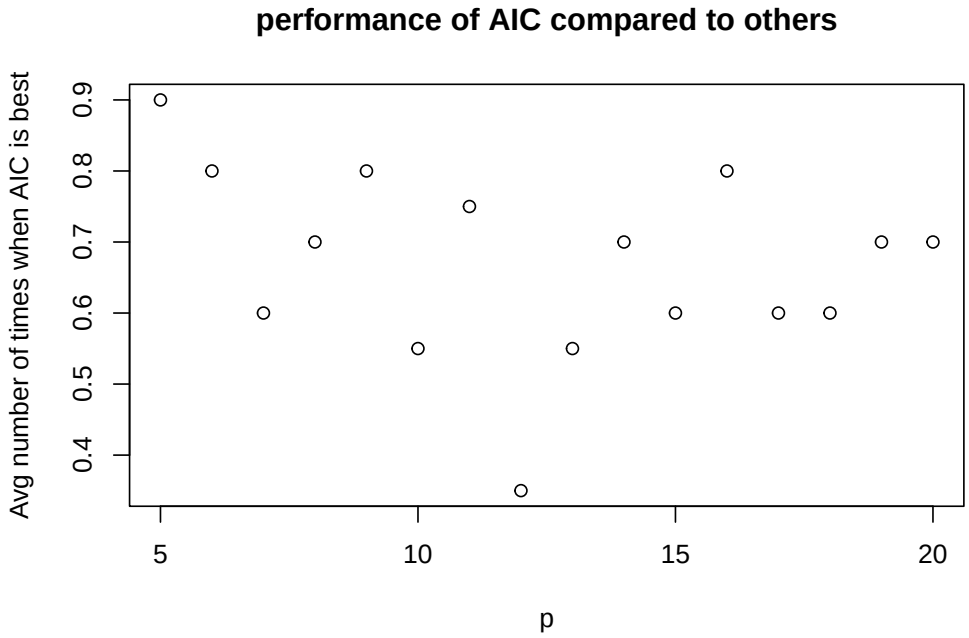
```
      best_criterion = NA
    )

    ##We check which criterion gives the closest lambda wrt lambda_opt
    if(min(abs(new_row_short_df$lambda_AIC - new_row_short_df$lambda_opt), abs(new_row
      new_row_short_df$best_criterion = 0 ##0 : best criterion is CV
    }
    else if(    abs(new_row_short_df$lambda_AIC - new_row_short_df$lambda_opt)
                  > abs(new_row_short_df$lambda_BIC - new_row_short_df$lambda_opt)
            ){
      new_row_short_df$best_criterion = 2  ##1: best criterion is BIC
    }
    else{
        new_row_short_df$best_criterion = 1  ##1: best criterion is AIC
    }


    results_short <- rbind(results_short, new_row_short_df)
  }##End of for loop for computing on different instance of Y = Xbeta + eps


  return(results_short)
}
```

```
n <- 30
prange<- 5:20
ratio_best_AIC = c(rep(0,length(prange)))
ratio_best_BIC = c(rep(0,length(prange)))
result_var_np = data.frame(
  p = c(),
  num_AIC_best = c(),   ##Number of times when AIC is the best
  num_CV_best = c(),   ##Number of times when CV is best
  num_BIC_best = c()
)
for(s in 1:length(prange)){
  results_lambda <- GetLambdaComparison(n=n, p = prange[s], B = 20, N= 20)
  new_row_result_var_np = data.frame(
    p = prange[s],
    num_AIC_best = sum(results_lambda$best_criterion == 1),
    num_CV_best = sum(results_lambda$best_criterion == 0),
    num_BIC_best = sum(results_lambda$best_criterion == 2)
  )

    ratio_best_AIC[s] <- new_row_result_var_np$num_AIC_best / (new_row_result_var_np$n
  ratio_best_BIC[s] <- new_row_result_var_np$num_BIC_best / (new_row_result_var_np$num

  print(s)
```

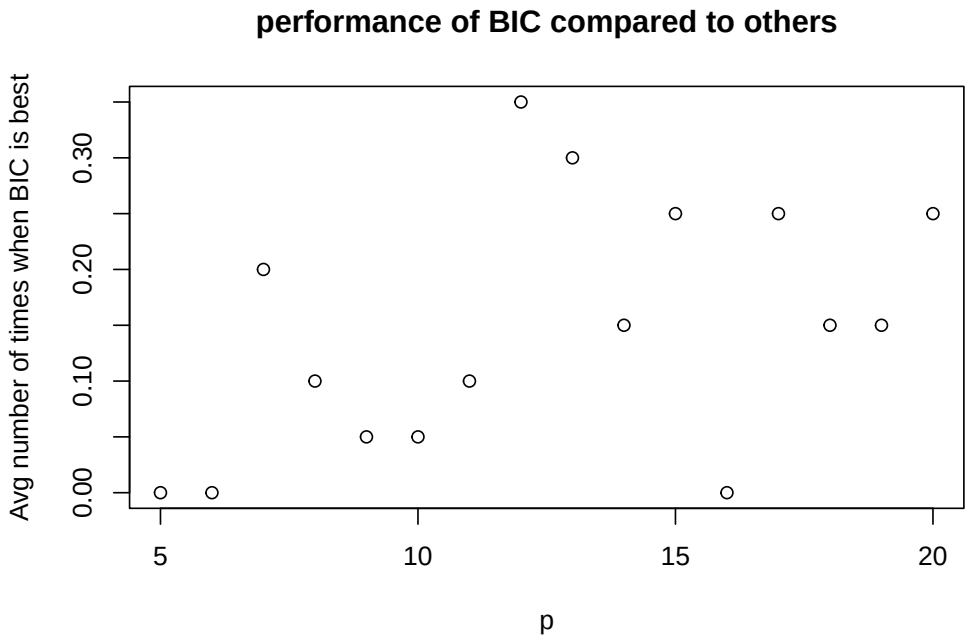
```
  result_var_np <- rbind(result_var_np, new_row_result_var_np)
}
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
```

```
plot(ratio_best_AIC, x=prange, main = "performance of AIC compared to others", ylab =
```




```
plot(ratio_best_BIC, x = prange, main = "performance of BIC compared to others", ylab
```



We now plot, depending on n/p and k (the number of nonzero coefficient) the performance of AIC, BIC and CV.

```
##We wrap the previous code in a function to get, for fixed n and a fixed sparsity lev
#Then we will vary k between some values and then display the results
CompareCriterionVarNP <- function(
                                  n = 100,
                                  prange      = 5:200, #List of p tested, all p should b
                                  p0     = 5,  #Number of nonzero coefficients
                                  sigma  = 1,
                                  sigmaX = 1,
                                  b0     = 1,
                                  B = 30, #Number of lambda tested
                                  N = 10, #Number of instances of solving Y = X\theta +
                                  lambda_step = 0.05, #Difference between two consecutiv
                                  lambda_init = 0.0 #First value used by lambda

                                    ){
  # result_var_np = data.frame(
  #   p = c(rep(NA, max(p0 - prange[1], 0))),
  #   num_AIC_best = c(rep(NA, max(p0 -prange[1], 0))),   ##Number of times when AIC i
  #   num_CV_best = c(rep(NA, max(p0 -prange[1], 0))),   ##Number of times when CV is
  #   num_BIC_best = c(rep(NA, max(p0 -prange[1], 0)))
  #
  #   ## prange[1]  prange[2]
  #   #k1  k2  k3
  #   # k < prange[1] -> _best[k, ..., prange[1]] = NA
```

```
  # )
  result_var_np = data.frame(
    p = c(),
    num_AIC_best = c(),
    num_CV_best = c(),
    num_BIC_best = c()
  )
  for(p in prange ){
    results_lambda <- GetLambdaComparison(n=n, p = p, B = B, N= N, p0 = p0,sigma = sig
    new_row_result_var_np = data.frame(
      p = p,
      num_AIC_best = sum(results_lambda$best_criterion == 1),
      num_CV_best = sum(results_lambda$best_criterion == 0),
      num_BIC_best = sum(results_lambda$best_criterion == 2)
    )
    result_var_np <- rbind(result_var_np, new_row_result_var_np)

  }
  return(result_var_np)
}
```

```
n<-50
results_glob = data.frame(
  k = c(),
  comp_criterion = c()
)
krange = c(5, 10, 20, 30, 40, 50)
prange = c(5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100)
N = 30
for(k in krange){

  new_row = data.frame(
    k = k,
    comp_criterion = CompareCriterionVarNP(n = n, prange = prange, p0 = k,N=N)
  )

  results_glob <- rbind(results_glob, new_row)
  print(k)
}
```

```
## [1] 5
## [1] 10
## [1] 20
## [1] 30
## [1] 40
## [1] 50
```

```
##Here we do the plotting
print(length(results_glob$comp_criterion.num_AIC_best))
```

```
## [1] 66
```

```
x = prange
y = krange
print(length(x)*length(y))
```

```
## [1] 66
```

```
z = matrix(data = results_glob$comp_criterion.num_AIC_best, nrow = length(x), ncol = l
persp3d(x=x, y=y,z=z, ylab = "number of nonzero coefs", xlab = "p", col="skyblue", typ
z = matrix(data = results_glob$comp_criterion.num_BIC_best, nrow = length(y), ncol = l
persp3d(x=x, y=y,z=z, ylab = "number of nonzero coefs", xlab = "p", col="red", type =
z = matrix(data = results_glob$comp_criterion.num_CV_best, nrow = length(y), ncol = le
persp3d(x=x, y=y,z=z, ylab = "number of nonzero coefs", xlab = "p",  col="green", add
rgl.snapshot("comp_all_crits.png")
```

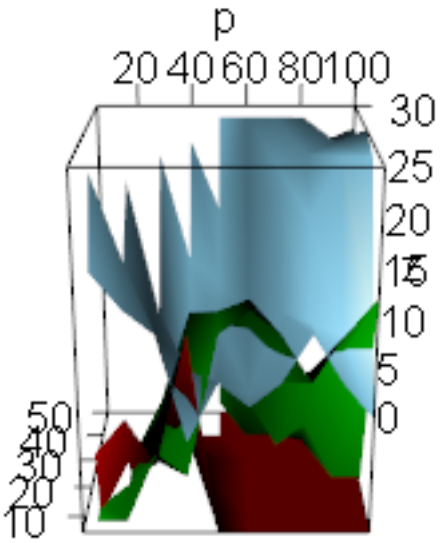


Figure 1: Comparison of which criterion gives the lambda closest to best lambda for varying dimensionality and sparsity

From this graph ($n = 50, p \in [5, 100], p_0 \in [5, 50]$) we have that AIC outperforms clearly both CV and BIC in the large p, large p_0 area (high-dimension, low sparsity). In the low-dimensionality, high sparsity area (my guess : $p_0 < p/2$) all criterions have a similar performance.

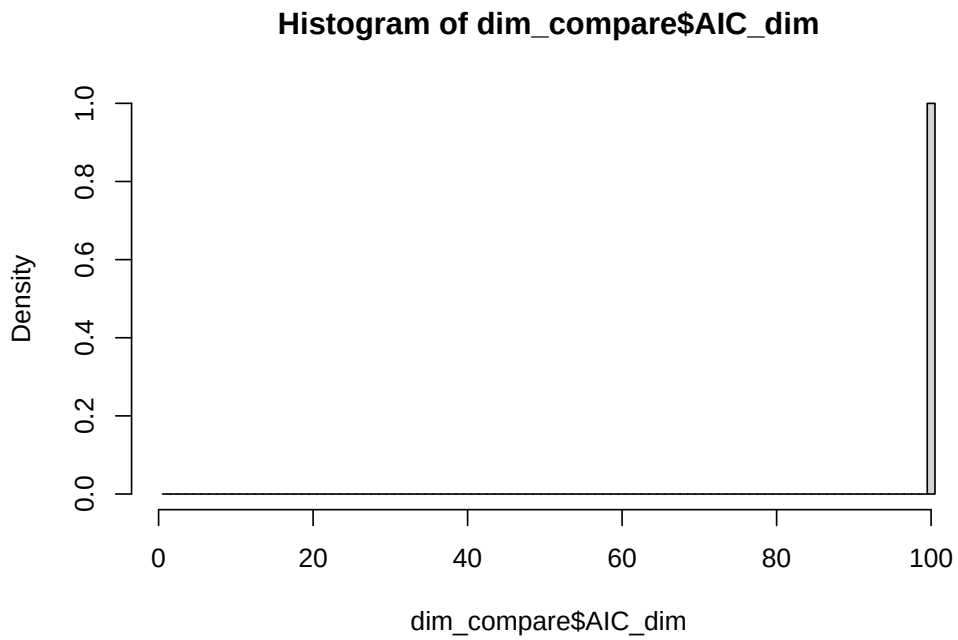
Now, instead of comparing among criterions which one gives the closest lambda to the best (knowing the true solution) lambda, we want to compare for each criterion which ones give the right model dimension.

```r
GetDimComparison <- function(n = 100,
                             p      = 10,
                             p0     = 5,  #Number of nonzero coefficients
                             sigma  = 1,
                             sigmaX = 1,
                             b0     = 1,
                             B = 20, #Number of lambda tested
                             N = 10, #Number of instances of solving Y = X\theta +
                             lambda_step = 0.1, #Difference between two consecutive
                             lambda_init = 0.0 #First value used by lambda
                             ){

  if(p < p0 ){##We treat this (impossible) case so that we can obtain dataframes of eq
    results_short <- data.frame(
      AIC_dim =NA,
      BIC_dim = NA,
      CV_dim = NA,
      true_dim = NA
    )
    return(results_short)
  }

  beta0  = c(rep(b0,p0),rep(0,p-p0))
  betah  = matrix(NA, ncol = p, nrow = B)

  results_dim <- data.frame(
    AIC_dim =c(),
    BIC_dim = c(),
    CV_dim = c(),
    true_dim = c()
  )



  for(t in 1:N){

    X = sapply(1:p, FUN=function(x){rnorm(n,0,sigmaX)})
    Y = X%*%beta0 + rnorm(n,0,sigma)

    lambda <- lambda_init
    results_frame <- data.frame(
      lambda = c(),
      betah = c(),
      k = c(),
      loglikelihood = c(),
      AIC = c(),
      BIC = c()
```

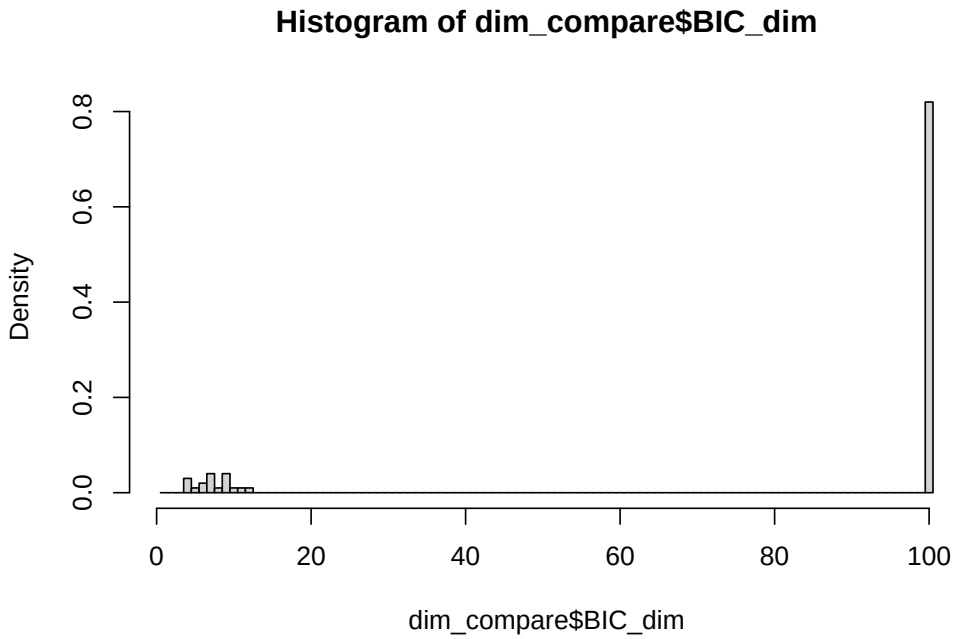
```
)

err_opt <- NA
lambda_opt <- 1
min_lambda_AIC <- NA
lambda_AIC <- NA
min_lambda_BIC <- NA
lambda_BIC <- NA
for (b in 1:B){#The increment is lambda
  temp = glmnet(X,Y,family = "gaussian",intercept=F, lambda=lambda)$beta
  ##the value returned by glmnet(...)$beta is a sparse matrix
  #In a sparse matrix : @i (non zero indices); @x non zero values

  betah <- c(rep(0,p))
  j<-1
  for(i in temp@i){
    betah[i+1] <- temp@x[j]
    j<-j+1
  }

  #llik = log(((1/sqrt(2 * pi))**(p/2))*exp(-t(Y-betah)%*%(Y-betah)/(2)))
  RSS = t(Y-X%*%betah)%*%(Y-X%*%betah) #Simpler expression for log likelihood in G
  new_row_df <- data.frame(
    lambda = lambda,
    betah = betah,
    k = j,
    negloglikelihood = -log(RSS),
    AIC = 2*j + n*log(RSS),
    BIC = j*log(n) + n*log(RSS/n)
  )
  results_frame <- rbind(results_frame, new_row_df)


  #We search for the best value for lambda under AIC
  if(min_lambda_AIC > new_row_df$AIC || is.na(min_lambda_AIC)){
    min_lambda_AIC <- new_row_df$AIC
    lambda_AIC <- lambda
  }
  #And we also search for the best value for lambda under BIC
  if(min_lambda_BIC > new_row_df$BIC || is.na(min_lambda_BIC)){
    min_lambda_BIC <- new_row_df$BIC
    lambda_BIC <- lambda
  }
  lambda <- lambda+lambda_step
}
lambda_CV = cv.glmnet(X,Y, family = "gaussian",intercept=F)$lambda.1se
```

```
    #We compute for each optimal lambda the corresponding solution beta
    Beta_AIC = glmnet(X, Y, family = "gaussian", intercept=F, lambda = lambda_AIC)$bet
    Beta_BIC = glmnet(X, Y, family = "gaussian", intercept=F, lambda = lambda_BIC)$bet
    Beta_CV = glmnet(X, Y, family = "gaussian", intercept=F, lambda = lambda_CV)$beta

    #For each beta we compute its dimension
    new_row_results_dim <- data.frame(
      AIC_dim = sum(Beta_AIC != 0),
      BIC_dim = sum(Beta_BIC != 0),
      CV_dim = sum(Beta_CV != 0),
      true_dim = p0
    )
    results_dim <- rbind(results_dim, new_row_results_dim)
  }
  return(results_dim)
}


dim_compare = GetDimComparison(N = 100, p = 100, p0 = 4)
```

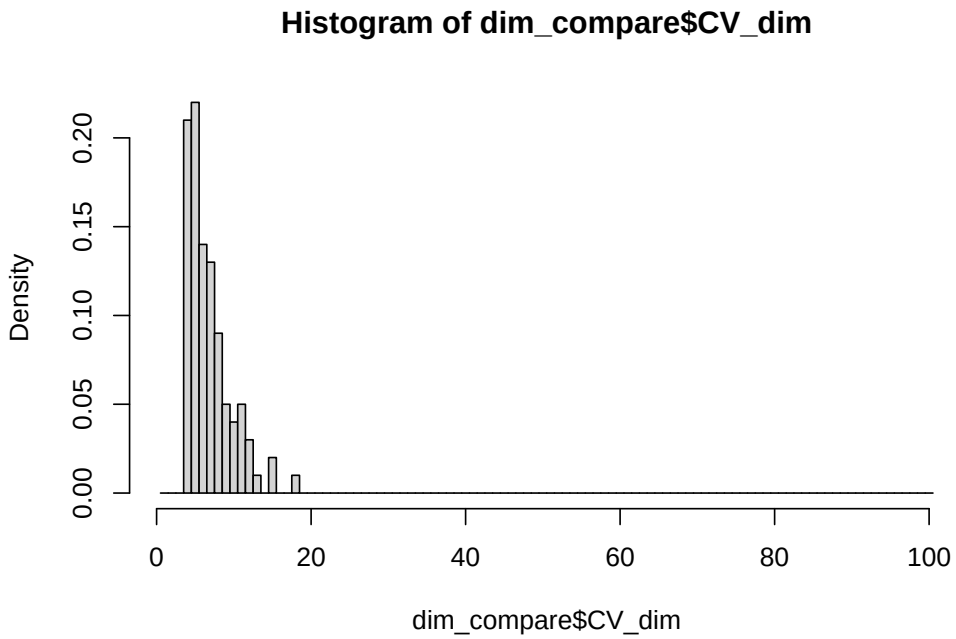
```
hist(dim_compare$AIC_dim, breaks = 0:100 +0.5, freq = FALSE)
```




```
hist(dim_compare$BIC_dim, breaks =0:100+0.5, freq = FALSE)
```




```
hist(dim_compare$CV_dim, breaks =0:100 +0.5, freq = FALSE)
```



We obtain from previous plots (with e.g. N=100, p=10 or p =100) that CV is the best criterion to obtain the right model dimension. Also BIC performs better than AIC for dimension identification, however in a high dimensional setting, the performance of BIC is quite low. In order to back this claim, we do as before and display for varying n/p the dimension of the models identified by the different

criterions.

We now plot, depending on n/p and k (the number of nonzero coefficient) the performance (avg model dimension) of AIC, BIC and CV.

```
##We wrap the previous code in a function to get, for fixed n and a fixed sparsity lev
#Then we will vary k between some values and then display the results
CompareDimCriterionVarNP <- function(
                                  n = 100,
                                  prange      = 5:200, #List of p tested, all p should b
                                  p0     = 5,  #Number of nonzero coefficients
                                  sigma  = 1,
                                  sigmaX = 1,
                                  b0     = 1,
                                  B = 30, #Number of lambda tested
                                  N = 10, #Number of instances of solving Y = X\theta +
                                  lambda_step = 0.05, #Difference between two consecutiv
                                  lambda_init = 0.0 #First value used by lambda
                                    ){
  # result_dim_var_np = data.frame(
  #   p = c(rep(prange[1], max(prange[1]-1, 0))),
  #   dim_AIC = c(rep(NA, prange[1]-1)),   ##Average dimension of AIC
  #   dim_BIC = c(rep(NA, prange[1]-1)),   ##of BIC
  #   dim_CV = c(rep(NA, prange[1]-1)),    ##of CV
  #   dim_true = c(rep(NA, prange[1] - 1))
  # )
  result_dim_var_np = data.frame(
    p = c(),
    dim_AIC = c(),
    dim_BIC = c(),
    dim_CV = c(),
    dim_true = c()
  )
  for(s in 1:length(prange)){
    results_dim <- GetDimComparison(n=n, p = prange[s], B = B, N= N, p0 = p0, sigma =
    new_row_result_dim_var_np = data.frame(
      p = prange[s],
      dim_AIC = sum(results_dim$AIC_dim)/N,
      dim_BIC = sum(results_dim$BIC_dim)/N,
      dim_CV = sum(results_dim$CV_dim)/N,
      dim_true = p0
    )
    result_dim_var_np <- rbind(result_dim_var_np, new_row_result_dim_var_np)

    #print(s)
  }
  return(result_dim_var_np)
}
```

```
results_dim_glob = data.frame(
  k = c(),
  comp_criterion = c()
)


n = 50
krange = c(1, 2, 3, 4, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50)
prange = c(5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95,

for(k in krange){
  new_row_dim = data.frame(
    k = k,
    comp_criterion = CompareDimCriterionVarNP(n = n, prange = prange, p0 = k)
  )
  results_dim_glob <- rbind(results_dim_glob, new_row_dim)
  print(k)
}
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 10
## [1] 15
## [1] 20
## [1] 25
## [1] 30
## [1] 35
## [1] 40
## [1] 45
## [1] 50
```

And we plot the results

```
##Here we do the plotting
print(length(results_dim_glob$comp_criterion.dim_CV))
```

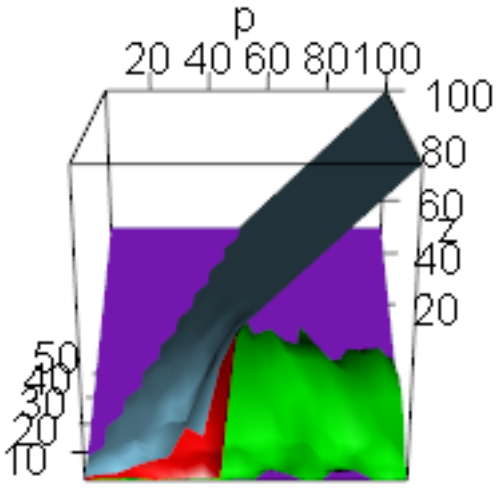
```
## [1] 280
```

```
x = krange
y = prange

print(length(x)*length(y))
```

```
## [1] 280
```

```
z = matrix(data = results_dim_glob$comp_criterion.dim_AIC, nrow = length(y), ncol = le
persp3d(x=y, y=x,z=z, ylab = "number of nonzero coefs", xlab = "p", col="skyblue", typ
```

```
z = matrix(data = results_dim_glob$comp_criterion.dim_BIC, nrow = length(y), ncol = le
persp3d(x=y, y=x,z=z, ylab = "number of nonzero coefs", xlab = "p", col="red", type =
z = matrix(data = results_dim_glob$comp_criterion.dim_CV, nrow = length(y), ncol = len
persp3d(x=y, y=x,z=z, ylab = "number of nonzero coefs", xlab = "p",  col="green", add
z = matrix(data = results_dim_glob$comp_criterion.dim_true, nrow = length(y), ncol = l
persp3d(x=y, y=x,z=z, ylab = "number of nonzero coefs", xlab = "p",  col="purple", add
rgl.snapshot("comp_all_crits_dim.png")
```



We can observe that when $p = p0$, then all criterions perform equally. This probably comes from (TODO : check and add details) the fact when the sparsity of the input vector is constant and $p = p0$, hence AIC and BIC can be proved equivalent. Also, in the Gaussian case this certainly amounts to solving a least square to find the CV solution which is the same as for AIC or BIC.

We can also take a closer look at what happens in the sparse low dimensional area:

```
results_dim_glob = data.frame(
  k = c(),
  comp_criterion = c()
)


n = 50
krange = 1:25
prange = 2*(2:25)

for(k in krange){
  new_row_dim = data.frame(
    k = k,
```

```
    comp_criterion = CompareDimCriterionVarNP(n = n, prange = prange, p0 = k)
  )
  results_dim_glob <- rbind(results_dim_glob, new_row_dim)
  print(k)
}
```

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
## [1] 17
## [1] 18
## [1] 19
## [1] 20
## [1] 21
## [1] 22
## [1] 23
## [1] 24
## [1] 25
```

```
##Here we do the plotting
print(length(results_dim_glob$comp_criterion.dim_CV))
```

```
## [1] 600
```

```
x = krange
y = prange

print(length(x)*length(y))
```

```
## [1] 600
```

```
z = matrix(data = results_dim_glob$comp_criterion.dim_AIC, nrow = length(y), ncol = le
persp3d(x=y, y=x,z=z, ylab = "number of nonzero coefs", xlab = "p", col="skyblue", typ
z = matrix(data = results_dim_glob$comp_criterion.dim_BIC, nrow = length(y), ncol = le
persp3d(x=y, y=x,z=z, ylab = "number of nonzero coefs", xlab = "p", col="red", type =
z = matrix(data = results_dim_glob$comp_criterion.dim_CV, nrow = length(y), ncol = len
```

```
persp3d(x=y, y=x,z=z, ylab = "number of nonzero coefs", xlab = "p",  col="green", add
z = matrix(data = results_dim_glob$comp_criterion.dim_true, nrow = length(y), ncol = l
persp3d(x=y, y=x,z=z, ylab = "number of nonzero coefs", xlab = "p",  col="purple", add
rgl.snapshot("comp_all_crits_low_dim_sparse.png")
```

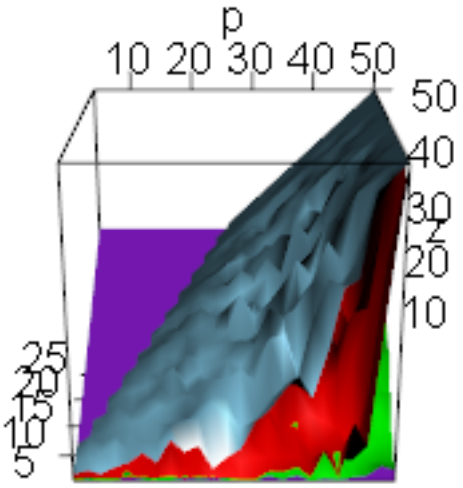


Figure 2: Comparison of the dimension identified depending on dimensionality and sparsity for sparse low dimension

#Influence of noise

We are now left with having to measure the influence of the noise. In order to do that we will plot graphs of dimensionality identified (or distance w.r.t true λ) where the parameters are $(\sigma, n/p)$ or (σ, k). #Dimensionality We compute dimension of the estimated model w.r.t to criterions for varying *sigma* and k

```
CompareDimCriterionVarSigma <- function(
                                n = 100,
                                p = 10,
                                sigmarange      = (0:20)/10, #List of sigma tested
                                p0      = 5,  #Number of nonzero coefficients
                                sigma  = 1,
                                sigmaX = 1,
                                b0     = 1,
                                B = 30, #Number of lambda tested
                                N = 10, #Number of instances of solving Y = X\theta +
                                lambda_step = 0.05, #Difference between two consecutiv
                                lambda_init = 0.0 #First value used by lambda
                                  ){
```

```
  result_dim_var_sigma = data.frame(
    sigma = c(),#c(rep(prange[1], max(prange[1]-1, 0))),
    dim_AIC = c(), #c(rep(NA, prange[1]-1)),   ##Average dimesnsion of AIC
    dim_BIC = c(), #c(rep(NA, prange[1]-1)),   ##of BIC
    dim_CV = c(),#c(rep(NA, prange[1]-1)),    ##of CV
    dim_true = c()#c(rep(NA, prange[1] - 1))
  )
  for(s in sigmarange){
    results_dim_sigma <- GetDimComparison(n=n, p = p, sigma = s, B = B, N= N, p0 = p0,
    new_row_result_dim_var_sigma = data.frame(
      sigma = s,
      dim_AIC = sum(results_dim_sigma$AIC_dim)/N,
      dim_BIC = sum(results_dim_sigma$BIC_dim)/N,
      dim_CV = sum(results_dim_sigma$CV_dim)/N,
      dim_true = p0
    )
    result_dim_var_sigma <- rbind(result_dim_var_sigma, new_row_result_dim_var_sigma)

    #print(s)
  }
  return(result_dim_var_sigma)
}
```

```
n<-50
p<-100
results_dim_glob_sigma = data.frame(
  k = c(),
  comp_criterion = c()
)
sigmarange = c(0.0, 0.5, 1.0, 1.50, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0)
krange = c(5, 10, 20, 30, 40, 50)

for(k in krange){
  new_row_dim_sigma = data.frame(
    k = k,
    comp_criterion = CompareDimCriterionVarSigma(n = n, p = p, p0 = k, sigmarange = si
  )
  results_dim_glob_sigma <- rbind(results_dim_glob_sigma, new_row_dim_sigma)
  print(k)
}
```

```
## [1] 5
## [1] 10
## [1] 20
## [1] 30
## [1] 40
## [1] 50
```

And we plot the results

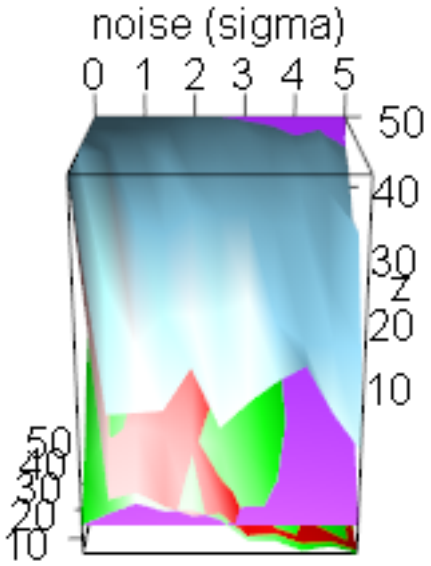
```
print(length(results_dim_glob_sigma$comp_criterion.dim_CV))
```

```
## [1] 66
```

```
y = krange
x = sigmarange
print(length(x)*length(y))
```

```
## [1] 66
```

```
z = matrix(data = results_dim_glob_sigma$comp_criterion.dim_AIC, nrow = length(y), ncc
persp3d(x=x, y=y,z=z, ylab = "number of nonzero coefs", xlab = "noise (sigma)", col="s
z = matrix(data = results_dim_glob_sigma$comp_criterion.dim_BIC, nrow = length(y), ncc
persp3d(x=x, y=y,z=z, ylab = "number of nonzero coefs", xlab = "noise (sigma)", col="r
z = matrix(data = results_dim_glob_sigma$comp_criterion.dim_CV, nrow = length(y), ncol
persp3d(x=x, y=y,z=z, ylab = "number of nonzero coefs", xlab = "noise (sigma)",  col="
z = matrix(data = results_dim_glob_sigma$comp_criterion.dim_true, nrow = length(y), nc
persp3d(x=x, y=y,z=z, ylab = "number of nonzero coefs", xlab = "noise (sigma)",  col="
rgl.snapshot("comp_all_crits_dim_var_sigma_high_dim.png")
```



We can make several observations from the plot in a low-dimension setting (for n = 50, p = 100, sigmarange = [0,5]) : 1) All criterions overfit (include too many parameters) the data. CV is the best among three as it consistently identifies a dimension between 2 and 3 times the true dimension. 2) There is a cliff given by some $\sigma_i = \phi_i(p_0)$ AIC and BIC perform nearly no dimension reduction (BIC makes dimension reduction soon) which is for small noise (!) 3) In a situation without noise CV is the best, when noise is quite large BIC performs equivalently to CV (but is faster)

```
n<-100
p<-50
results_dim_glob_sigma = data.frame(
  k = c(),
  comp_criterion = c()
)
sigmarange = c(0.0, 0.5, 1.0, 1.50, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0)
krange = c(5, 10, 20, 30, 40, 50)

for(k in krange){
  new_row_dim_sigma = data.frame(
    k = k,
    comp_criterion = CompareDimCriterionVarSigma(n = n, p = p, p0 = k, sigmarange = si
  )
  results_dim_glob_sigma <- rbind(results_dim_glob_sigma, new_row_dim_sigma)
  print(k)
}
```

```
## [1] 5
## [1] 10
## [1] 20
## [1] 30
## [1] 40
## [1] 50
```

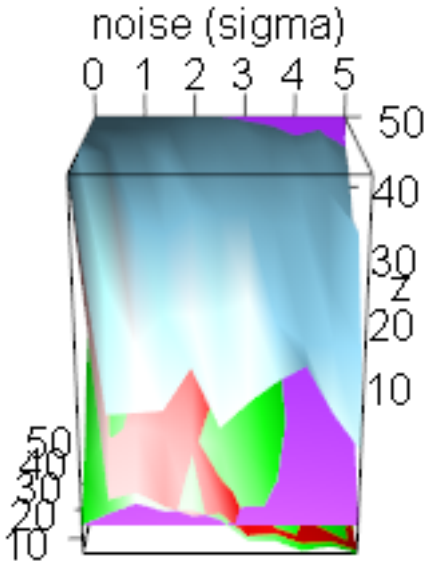
```
print(length(results_dim_glob_sigma$comp_criterion.dim_CV))
```

```
## [1] 66
```

```
y = krange
x = sigmarange
print(length(x)*length(y))
```

```
## [1] 66
```

```
z = matrix(data = results_dim_glob_sigma$comp_criterion.dim_AIC, nrow = length(y), nco
persp3d(x=x, y=y,z=z, ylab = "number of nonzero coefs", xlab = "noise (sigma)", col="s
z = matrix(data = results_dim_glob_sigma$comp_criterion.dim_BIC, nrow = length(y), nco
persp3d(x=x, y=y,z=z, ylab = "number of nonzero coefs", xlab = "noise (sigma)", col="r
z = matrix(data = results_dim_glob_sigma$comp_criterion.dim_CV, nrow = length(y), ncol
persp3d(x=x, y=y,z=z, ylab = "number of nonzero coefs", xlab = "noise (sigma)",  col="
z = matrix(data = results_dim_glob_sigma$comp_criterion.dim_true, nrow = length(y), nc
persp3d(x=x, y=y,z=z, ylab = "number of nonzero coefs", xlab = "noise (sigma)",  col="
rgl.snapshot("comp_all_crits_dim_var_sigma_low_dim.png")
```



We can make several observations from the plot (for n = 50, p = 100, sigmarange = [0,5]) :

#Closeness to $lambda_{opt}$

```
n<-50
results_glob_sigma_np = data.frame(
  sigma = c(),
  comp_criterion = c()
)
sigmarange = c(0.0, 0.25, 0.5, 0.75, 1.0, 1.25, 1.50, 1.75, 2.0, 2.5)
prange = c(5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100)
for(sigma in sigmarange){
  new_row_sigma_np = data.frame(
    sigma = sigma,
    comp_criterion = CompareCriterionVarNP(n = n, prange = prange, p0 = 5, sigma = sig
  )
  results_glob_sigma_np <- rbind(results_glob_sigma_np, new_row_sigma_np)
  #print(sigma)
}
```

```
print(length(results_glob_sigma_np$comp_criterion.num_CV_best))
```

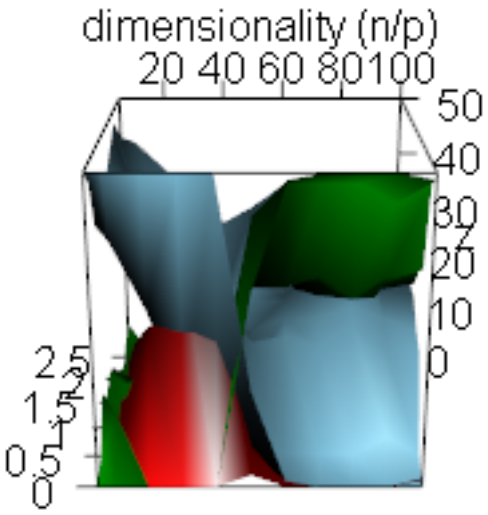
```
## [1] 110
```

```
x = prange
y = sigmarange
print(length(x)*length(y))
```

```
## [1] 110
```

```
z = matrix(data = results_glob_sigma_np$comp_criterion.num_AIC_best, nrow = length(y),
persp3d(x=x, y=y,z=z, xlab = "dimensionality (n/p)", ylab = "noise (sigma)", col="skyb
```

```
z = matrix(data = results_glob_sigma_np$comp_criterion.num_BIC_best, nrow = length(y),
persp3d(x=x, y=y,z=z, xlab = "dimensionality (n/p)", ylab = "noise (sigma)", col="red"
z = matrix(data = results_glob_sigma_np$comp_criterion.num_CV_best, nrow = length(y),
persp3d(x=x, y=y,z=z, xlab = "dimensionality (p)", ylab = "noise (sigma)",  col="green
rgl.snapshot("comp_all_crits_var_sigma_p.png")
```



At the moment, this graph makes not much sense. It might be better to plot something like ℓ^2 distance between $\beta_{\lambda_{crit}}$ and β^*.

6 Overview of the project

You will first implement the cross validation and the BIC to calibrate the Lasso numerically in the regression setting. Then you will study the statistical properties of the lasso in terms of model selection in different simulation setting (high signal / high noise, low dimensional / high dimensional). You will compete CV, the AIC, and the BIC in order to determine if one procedure is more accurate than the other.

7 Reference

Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, Springer, 2009