Simulaciones artículo Gregorutti

Cristian Camilo Hidalgo García

9 de septiembre de 2016

Comenzaremos simulando los casos presentados en el artículo:

Caso uno

Dos variables correlacionadas. El simple contexto donde (X_1, X_2, Y) $N_3(0, \Sigma)$ con $\Sigma = \begin{pmatrix} C & \tau \\ \tau^t & 1 \end{pmatrix}$

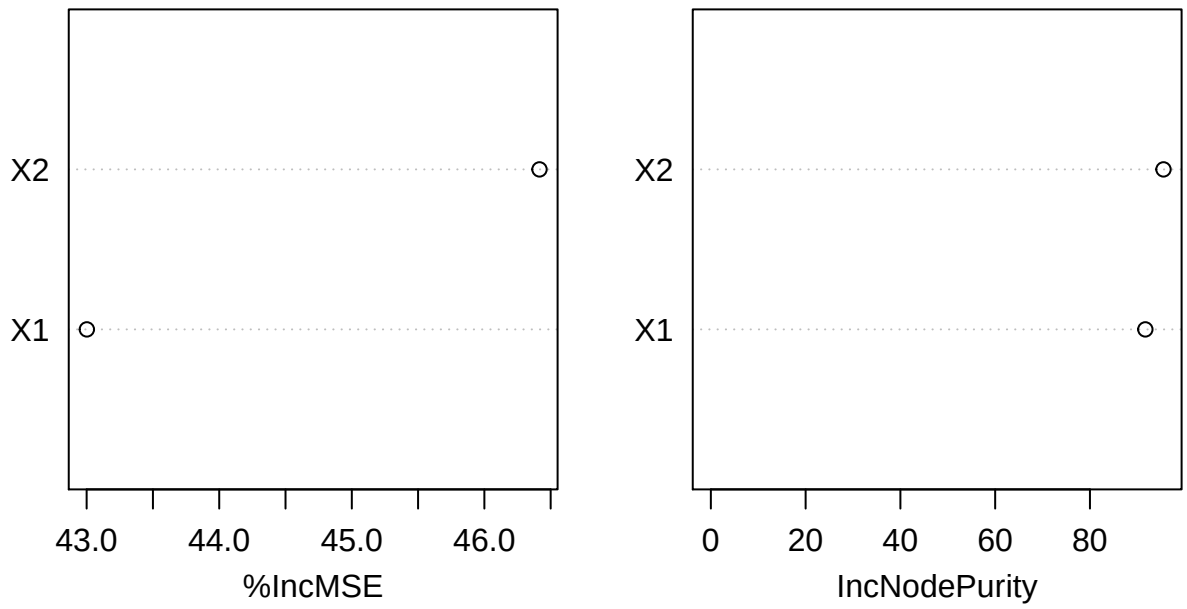
Con $C = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix}$ La matriz de varianzas covarianzas de X_1 y X_2, y $\tau = (\tau_0, \tau_0)$ que sería:

$$\begin{pmatrix} 1 & c & \tau_0 \\ c & 1 & \tau_0 \\ \tau_0 & \tau_0 & 1 \end{pmatrix}$$

Simulamos el escenario anterior:

```
set.seed(1)
#Simulacion de datos normales multivariados
# Simularemos varias distribuciones normales en tres dimensiones:
N=200
mu=c(0,0,0)
sigma <- matrix(c(1, 0.9,0.95,0.9,1,0.95,0.95,0.95,1),ncol=3,nrow=3)
## Simulacion de los datos correlacionados:
muestra1=rmvnorm(N,mean=mu,sigma=sigma,method = "eigen")
colnames(muestra1)=c("X1","X2","Y")

Bosque1=randomForest(Y~X1+X2, data=muestra1,ntree=800,importance=T)
varImpPlot(Bosque1,main="Bosque Caso 1")
```



Vemos que la importancia se da teoricamente como:

$I(X_j) = 2(\frac{\tau_0}{1+c})^2$

```
I_Xj=2*(0.95/(1+0.95))^2
I_Xj
```

```
## [1] 0.4746877
```

Podemos observar que la importancia se ha reducido, ya que al estar positivamente correlacionadas, al permutar las observaciones de una, el error no aumenta mucho ya que existe la presencia de la otra variable.

caso 2

En este caso tenemos dos variables correlacionadas y una variable independiente. Adicionamos al caso anterior una variable adicional X_3 que será independiente de X_1 y X_2

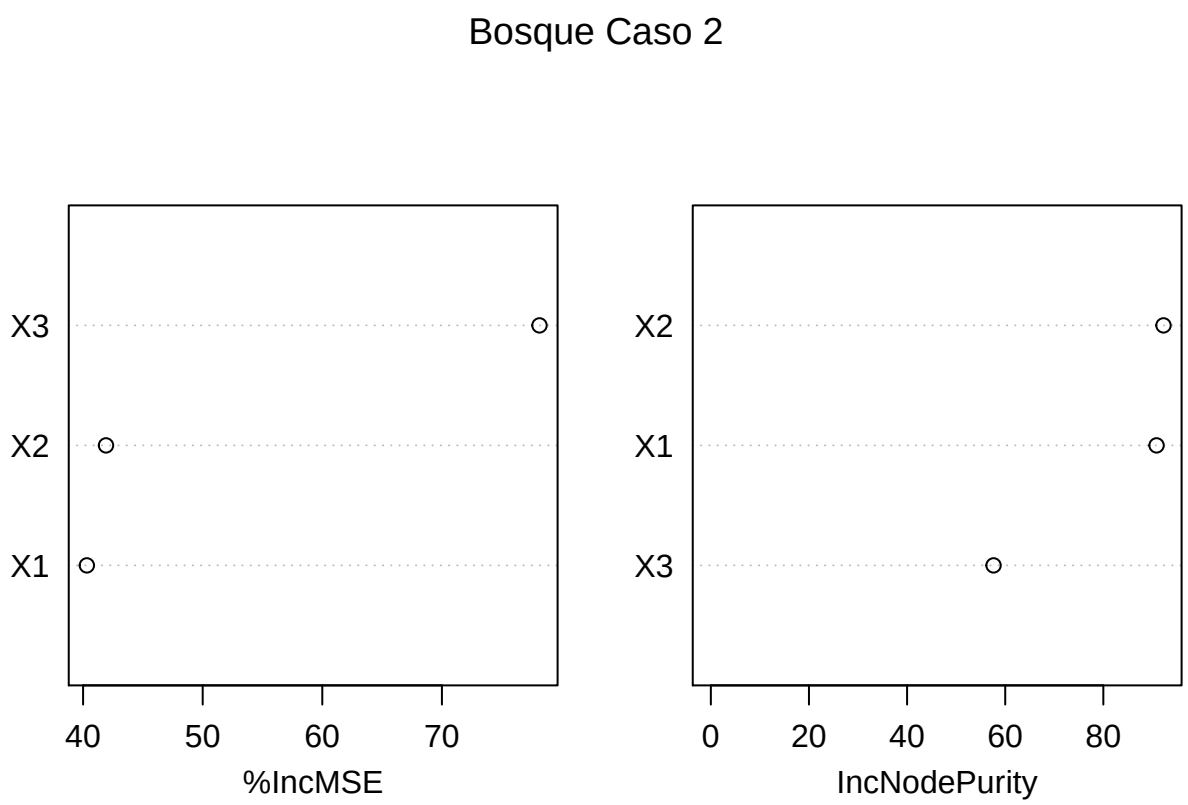
$C= \begin{pmatrix} 1 & c & 0 \\ c & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

y $\tau^t = (\tau_0, \tau_0, \tau_3)$. Podemos verificar facilmente que $C^{-1}(\tau_0, \tau_0, \tau_3)^t = (\frac{\tau_0}{1+c}, \frac{\tau_0}{1+c}, \tau_3)^t$

Simulando el escenario anterior, suponiendo que $\tau_0 > \tau_3$, observamos:

```r
set.seed(2)
#Simulacion de datos normales multivariados
# Simularemos varias distribuciones normales en tres dimensiones:
N=200
mu=c(0,0,0,0)
C=matrix(c(1, 0.94,0.01,0.94,1,0.01,0.01,0.01,1),ncol=3,nrow=3)
tau=c(0.9,0.9,0.65)
vary=1
sigma<-rbind(cbind(C,tau),c(tau,1))
sigma=Matrix::nearPD(sigma)
## Simulacion de los datos correlacionados:
muestra1=MASS::mvrnorm(n=200, mu=mu, Sigma=sigma$mat)
colnames(muestra1)=c("X1","X2","X3","Y")

Bosque1=randomForest(Y~X1+X2+X3, data=muestra1,ntree=800,importance=T)
varImpPlot(Bosque1,main="Bosque Caso 2")
```



De lo anterior, se concluye que a pesar de que hay una mayor correlación entre X_1 y X_2 y la variable dependiente Y, que esta última con $X3$, el bosque aleatorio le asigna mayor importancia a X_3.

Vemos que la importancia se da teoricamente como:

$I(X_j) = 2(\frac{\tau_0}{1+c})^2$

```r
Xj=c(0.95533045,0.95533045,0.6033489*(1+0.95533045))
I_Xj=2*(Xj/(1+0.95533045))^2
I_Xj
```

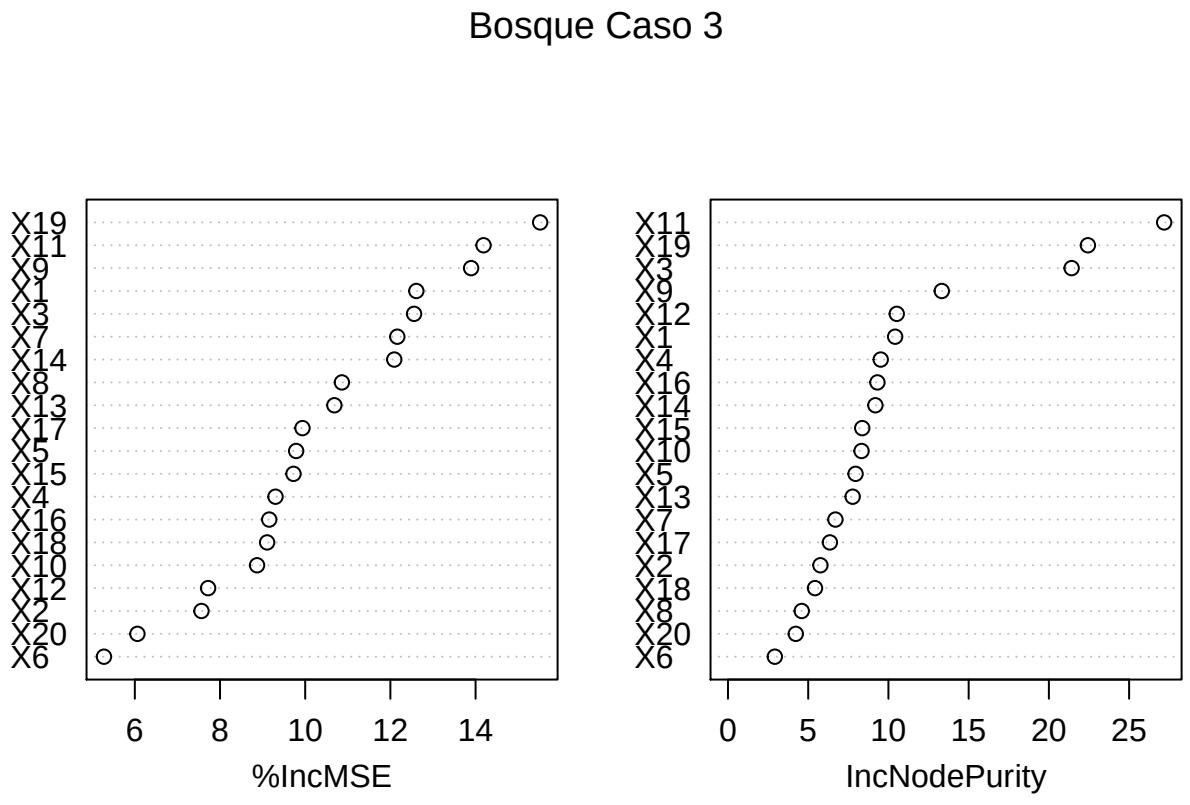
```
## [1] 0.4774159 0.4774159 0.7280598
```

Caso 3

En este caso consideraremos p variables correlacionadas, elegiremos p=20, donde:

```
set.seed(3)
p=20
C=matrix(rep(0.88:0.88,400),ncol=p,nrow=p)
diag(C)=1
tau=c(rep(0.85:0.85,20))
vary=1
mu=c(rep(0:0,21))
sigma<-rbind(cbind(C,tau),c(tau,1))
sigma=Matrix::nearPD(sigma)
## Simulacion de los datos correlacionados:
muestra1=MASS::mvrnorm(n=200, mu=mu, Sigma=sigma$mat)
colnames(muestra1)=c("X1","X2","X3","X4","X5",
"X6","X7","X8","X9","X10","X11","X12","X13","X14","X15",
"X16","X17","X18","X19","X20","Y")

Bosque1=randomForest(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+
                       X10+X11+X12+X13+X14+X15+X16+X17+X18+X19+X20, data=muestra1,ntree=800,importance="
varImpPlot(Bosque1,main="Bosque Caso 3")
```



Observamos que la importancia de la variable, si se corre la muestra aleatoria extraída de la normal multivariada, se debe a la aleatoriedad, pues siempre elegirá una variable diferente.

La importancia de las variables viene dada por:

```
I_Xj=2*(0.85/(1-0.88+p*0.88))^2
I_Xj
```

```
## [1] 0.004601934
```

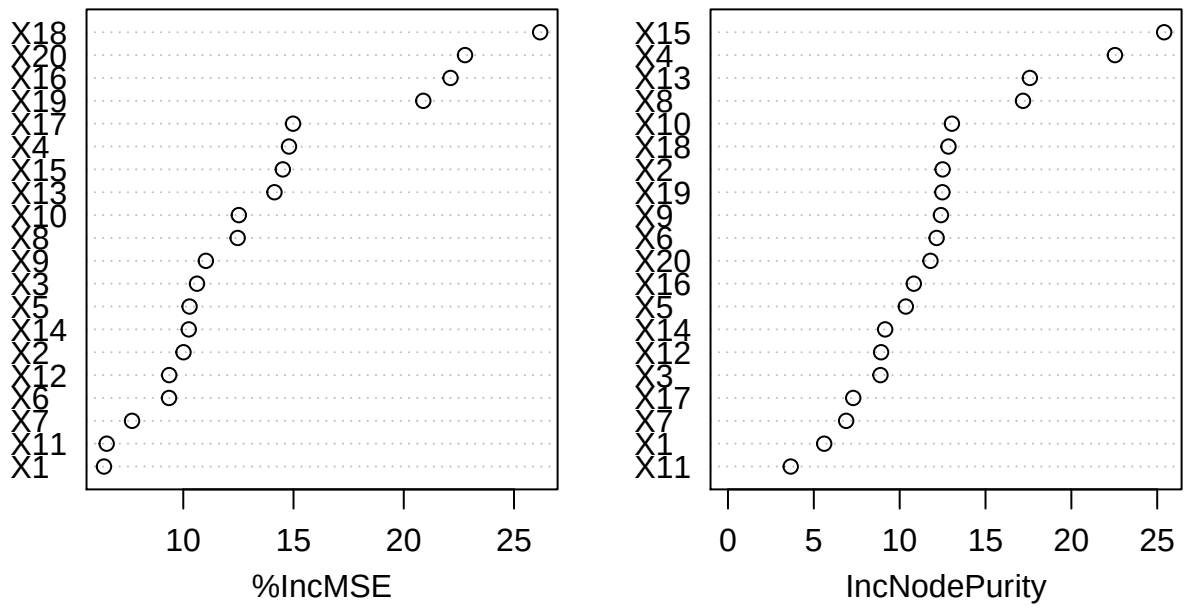
Que de hecho, viene siendo una generalización del caso uno.

Caso 4

En este caso consideraremos un grupo (p) de variables correlacionadas y un grupo (q) de variables incorrelacionadas.

```
set.seed(4)
#Simulacion de datos normales multivariados
# Simularemos varias distribuciones normales en tres dimensiones:
N=200
p=20
mu=c(rep(0:0,20))
C=matrix(rep(0.92:0.92,400),ncol=p,nrow=p)
C[16:20,]=0
C[,16:20]=0
diag(C)=1
tau=c(rep(0.87:0.87,15),rep(0.48:0.48,5))
vary=1
mu=c(rep(0:0,21))
sigma<-rbind(cbind(C,tau),c(tau,1))
sigma=Matrix::nearPD(sigma)
## Simulacion de los datos correlacionados:
muestra1=MASS::mvrnorm(n=200, mu=mu, Sigma=sigma$mat)
colnames(muestra1)=c("X1","X2","X3","X4","X5","X6","X7",
"X8","X9","X10","X11","X12","X13","X14","X15","X16","X17","X18","X19","X20","Y")

Bosque1=randomForest(Y~X1+X2+X3+X4+X5+X6+X7+X8+X9+X10+
                       X11+X12+X13+X14+X15+X16+X17+X18+X19+X20, data=muestra1,ntree=800,importance=T)
varImpPlot(Bosque1,main="Bosque Caso 4")
```



En este último resultado, observamos que se seleccionan las variables menos correlacionadas con la variable dependiente como importantes ya que se resta importancia a las demás por la presencia de correlación entre ellas.

Teóricamente, la correlación para las variables regresoras dependientes se da como:

```
I_Xj=2*(0.87/(1-0.92+p*0.92))^2
I_Xj
```

```
## [1] 0.004432661
```

Y las independientes:

```
I_Xj=2*(0.48)^2
I_Xj
```

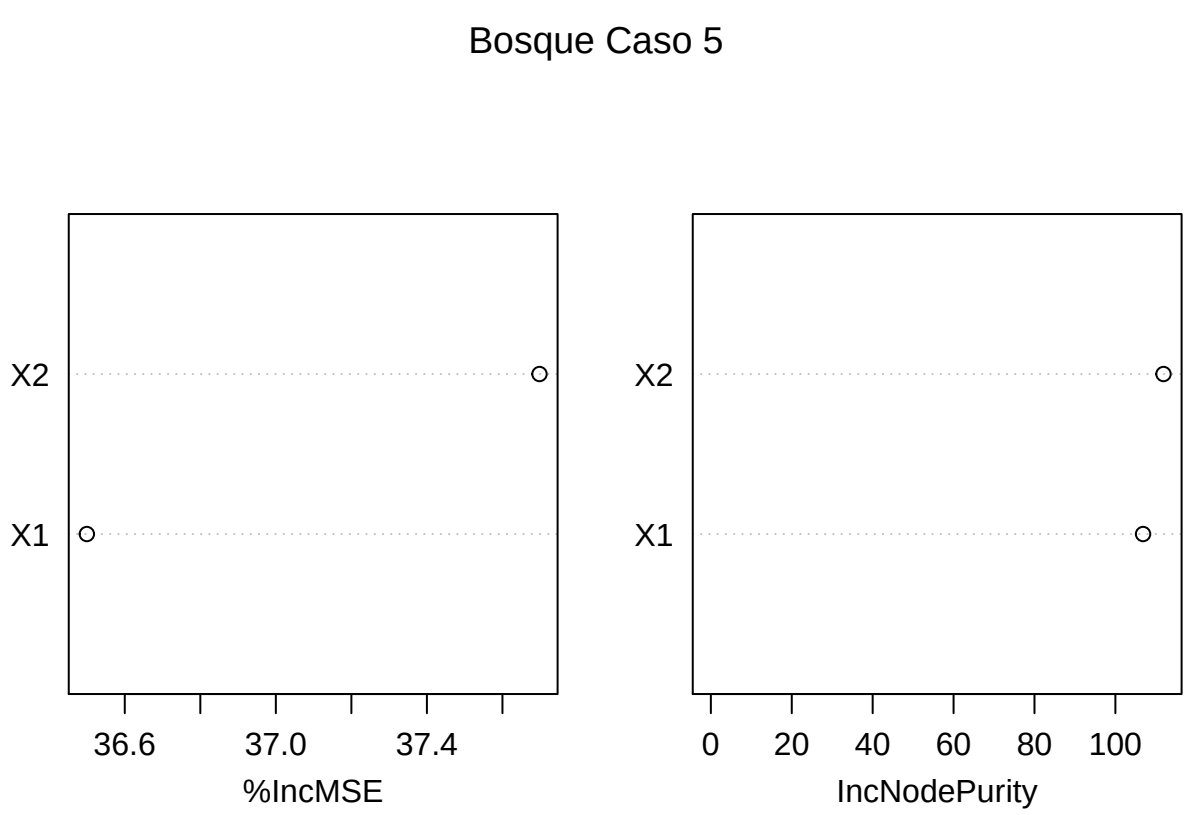
```
## [1] 0.4608
```

Caso 5

En este caso, mencionaremos la anti-correlación, anteriormenta sólo consideramos los casos donde la correlación era positiva entre los predictores. Consideremos la correlación entre X_1 y X_2 igual a $-\rho$, asumiendo varianzas unitarias, la importancia del incremento en la permutación que mide el *random forest* cuando ρ tiende a -1 induce a una alta predicción del error, ya que las dos explican la variable dependiente en sentidos contrarios por lo que se necesitarían las dos en el modelo. Así que la importancia de la variable será alta.

```r
set.seed(5)
N=200
mu=c(0,0,0)
sigma <- matrix(c(1, -1,0.95,-1,1,-0.95,0.95,-0.95,1),ncol=3,nrow=3)
## Simulacion de los datos correlacionados:
muestra1=rmvnorm(N,mean=mu,sigma=sigma,method = "eigen")
colnames(muestra1)=c("X1","X2","Y")

Bosque1=randomForest(Y~X1+X2, data=muestra1,ntree=800,importance=T)
varImpPlot(Bosque1,main="Bosque Caso 5")
```



Experimentos

Experimento 1: Pequeñas correlaciones entre algunos predictores

Generamos un grupo de 3 variables relevantes apra explicar Y, 3 moderadas y 3 debilmente relacionadas con y, también otro grupo de variables irrelevantes.

```r
Yj=c(1,2,3)
SimYj=rnorm(800,Yj,1)
X1=0.7*rnorm(800,Yj[1],1)+0.3*rnorm(800)
X2=0.7*rnorm(800,Yj[2],1)+0.3*rnorm(800)
X3=0.7*rnorm(800,Yj[3],1)+0.3*rnorm(800)
cor(SimYj,X1)
```

```
## [1] -0.003204928
```

```
cor(SimYj,X2)
```

```
## [1] -0.09209919
```

```
cor(X1,X3)
```

```
## [1] 0.0494445
```

¿?