Compatible model selection

Jean-Michel Marin

4/02/2022

1 Gaussian example

We consider an n-sample from $F = \mathcal{N}(0,1)$.

1.1 Known variance - True model belongs to the parametric family

We consider the parametric family $Q_\theta = \mathcal{N}(\theta,1)$ where $\theta \in \mathbb{R}$.

In such a case, we have $\beta = \max_{\theta\in\mathbb{R}} \int \log(q(y|\theta))f(y)\mathrm{d}y = -\log(2\pi)/2 - 1/2$.

For the n-sample we have $\beta_n = -n/2\log(2\pi) - n/2$.

```
beta <- -1/2*log(2*pi)-1/2
N <- 1000 ; n <- 500 ; B <- 500 ; alpha <- 0.1
statboot <- rep(0,B) ; cover <- 0
for (i in 1:N)
{
  x <- rnorm(n)
  sn2 <- var(x)*(n-1)/n
  for (j in 1:B)
  {
    xstar <- sample(x,n,rep=TRUE)
    sn2star <- var(xstar)*(n-1)/n
    statboot[j] <- sn2/2-sn2star/2
  }
  interva <- -1/2*log(2*pi)-sn2/2-
    quantile(statboot,c(1-alpha/2,alpha/2),names=FALSE)
  cover <- cover + (beta >= interva[1] & beta <= interva[2])
}
cover/N
```

```
## [1] 0.903
```

1.2 Unknown variance - True model belongs to the parametric family

We consider the parametric family $Q_{(\mu,\sigma^2)} = \mathcal{N}(\mu,\sigma^2)$ where $(\mu,\sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$.

In such a case, we have $\beta = \max_{(\mu,\sigma^2)\in\in\mathbb{R}\times\mathbb{R}_+^*} \int \log(q(y|(\mu,\sigma^2)))f(y)\mathrm{d}y = -\log(2\pi)/2 - 1/2$.

For the n-sample, we have $\beta_n = -n/2\log(2\pi) - n/2$.

```
beta <- -1/2*log(2*pi)-1/2
N <- 1000 ; n <- 500 ; B <- 500 ; alpha <- 0.1
statboot <- rep(0,B) ; cover <- 0
for (i in 1:N)
{
  x <- rnorm(n)
  sn2 <- var(x)*(n-1)/n
  for (j in 1:B)
  {
    xstar <- sample(x,n,rep=TRUE)
    sn2star <- var(xstar)*(n-1)/n
    statboot[j] <- log(sn2)/2-log(sn2star)/2
  }
  interva <- -1/2*(log(2*pi)+1)-log(sn2)/2-
    quantile(statboot,c(1-alpha/2,alpha/2),names=FALSE)
  cover <- cover + (beta >= interva[1] & beta <= interva[2])
}
cover/N
```

```
## [1] 0.896
```

2 Regression example

We consider an n-sample from the Gaussian regression model

$$F = [y|x_1, x_2] = \mathcal{N}(1 + x_1 + x_2, 1)\,.$$

```
n <- 500
X <- cbind(rep(1,n),matrix(rnorm(n*2),n,2))
Xm <- solve(t(X)%*%X)
p <- dim(X)[2]
betan <- -n/2*log(2*pi)-n/2
```

2.1 Known variance - True model belongs to the parametric family

We consider the parametric family $Q_{(\theta_0,\theta_1,\theta_2)} = \mathcal{N}(\theta_0 + \theta_1 x_1 + \theta_2 x_2, 1)$. In such a case, for the n-sample, we have $\beta_n = -n/2\log(2\pi) - n/2$.

```
N <- 1000 ; B <- 500 ; alpha <- 0.1
statboot <- rep(0,B) ; moy <- ampli <- cover <- 0
# tp <- txtProgressBar(min = 1, max = N, style = 3, char = "*")
for (i in 1:N)
{
```

```
  y <- 1*X[,1]+1*X[,2]+1*X[,3]+rnorm(n,0,1)
  yhat <- X%*%Xm%*%t(X)%*%y
  r <- y-yhat
  radj <- r/sqrt(1-p/n)
  for (j in 1:B)
    {
    radjstar <- sample(radj,n,rep=TRUE)
    ystar <- yhat+radjstar
    yhatstar <- X%*%Xm%*%t(X)%*%ystar
    statboot[j] <- sum(r^2)/2-sum((ystar-yhatstar)^2)/2
    }
  interva <- -n/2*log(2*pi)-sum(r^2)/2-
    quantile(statboot,c(1-alpha/2,alpha/2),names=FALSE)
  moy <- moy+sum(interva)/2
  ampli <- ampli+diff(interva)
  cover <- cover + (betan >= interva[1] & betan <= interva[2])
#   setTxtProgressBar(tp,i)
}
c(betan,moy/N,ampli/N,cover/N)
```

2.2 Unknown variance - True model belongs to the parametric family

We consider the parametric family $Q_{(\theta_0,\theta_1,\theta_2,\theta_3)} = \mathcal{N}(\theta_0+\theta_1 x_1+\theta_2 x_2, \theta_3)$ where $\theta_3>0$. In such a case, for the n-sample, we have

$$\beta_n = -n/2\log(2\pi) - n/2$$

```
N <- 1000 ; B <- 500 ; alpha <- 0.1
statboot <- rep(0,B) ; moy <- ampli <- cover <- 0
# tp <- txtProgressBar(min = 1, max = N, style = 3, char = "*")
for (i in 1:N)
{
  y <- 1*X[,1]+1*X[,2]+1*X[,3]+rnorm(n,0,1)
  yhat <- X%*%Xm%*%t(X)%*%y
  r <- y-yhat
  radj <- r/sqrt(1-p/n)
  for (j in 1:B)
    {
    radjstar <- sample(radj,n,rep=TRUE)
    ystar <- yhat+radjstar
    yhatstar <- X%*%Xm%*%t(X)%*%ystar
    statboot[j] <- n*log(mean(r^2))/2-n*log(mean((ystar-yhatstar)^2))/2
    }
  interva <- -n/2*(log(2*pi)+1)-n*log(mean(r^2))/2-
    quantile(statboot,c(1-alpha/2,alpha/2),names=FALSE)
  moy <- moy+sum(interva)/2
  ampli <- ampli+diff(interva)
```

```
  cover <- cover + (betan >= interva[1] & betan <= interva[2])
#   setTxtProgressBar(tp,i)
}
c(betan,moy/N,ampli/N,cover/N)
```

2.3 Unknown variance - True model bigger than the quasi true model

We consider the parametric family

$$Q_{(\theta_0,\theta_1)} = \mathcal{N}(\theta_0 + \theta_1 x_1, \theta_3)$$

```
N <- 1000 ; B <- 500 ; alpha <- 0.1
statboot <- rep(0,B) ; moy <- ampli <- cover <- 0
# tp <- txtProgressBar(min = 1, max = N, style = 3, char = "*")
for (i in 1:N)
{
  y <- 1*X[,1]+1*X[,2]+1*X[,3]+rnorm(n,0,1)
  yhat <- X%*%Xm%*%t(X)%*%y
  r <- y-yhat
  radj <- r/sqrt(1-p/n)
  for (j in 1:B)
    {
    radjstar <- sample(radj,n,rep=TRUE)
    ystar <- yhat+radjstar
    yhatstar <- X%*%Xm%*%t(X)%*%ystar
    statboot[j] <- sum(r^2)/2-sum((ystar-yhatstar)^2)/2
    }
  interva <- -n/2*log(2*pi)-sum(r^2)/2-
    quantile(statboot,c(1-alpha/2,alpha/2),names=FALSE)
  moy <- moy+sum(interva)/2
  ampli <- ampli+diff(interva)
  cover <- cover + (betan >= interva[1] & betan <= interva[2])
#   setTxtProgressBar(tp,i)
}
c(betan,moy/N,ampli/N,cover/N)
```