

# Bayes factors based on g-priors for variable selection (R code)

Gonzalo Garcia-Donato and Mark F. Steel

4/27/2021

## OBICE study, part I

```
library(BayesVarSel)
```

```
## Loading required package: MASS
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: parallel
```

```
packageVersion("BayesVarSel")
```

```
## [1] '2.2.0'
```

*#Note: you can download this and more recent version in the github repository [comodin19/BayesVarSel](https://github.com/comodin19/BayesVarSel)*

Now we create the dataframe (obesity) with the candidate variables and removing items with Not Available cases (in some cases coded as 9 or similar) with the following script.

```
data(OBICE)
```

```
ActFisica02<- rep(0, dim(OBICE)[1])
```

```
ActFisica02[OBICE$ActFisica<=4]<- 0 #<2 times per week
```

```
ActFisica02[OBICE$ActFisica==5]<- 1 #>= 2 times per week
```

```
ActFisica02[OBICE$ActFisica==9]<- 9 #NA
```

```
obesity<- data.frame(BMI=OBICE$PesoActual/OBICE$TallaAct^2,  
  FaObese=OBICE$PadreObeso, MoObese=OBICE$MadreObesa,  
  WeightBorn=OBICE$PesoNac, HeightBorn=OBICE$TallaNac,  
  Meals5=OBICE$CincoComidas, Vegeta=OBICE$Verduras, Fruit=OBICE$Fruta, Age=OBICE$Edad,  
  Sex=OBICE$Sexo, Breastfeed=OBICE$LactMatern, AfternoonSnack=OBICE$Merienda,  
  Activ=ActFisica02, candies=OBICE$Chuches, HrsPCDay=OBICE$HorasPCDiaPond,  
  HrsTVDay=OBICE$HorasTVDiaPond, HrsSleep=OBICE$HoraSuenyo)
```

```
obesity<- obesity[-1031,]
```

```
remove<- obesity$Activ==9 | obesity$candies==9 | obesity$WeightBorn==9999 |  
  obesity$HeightBorn==999.0 | obesity$candies==9 | is.na(obesity$HrsPCDay) |  
  is.na(obesity$HrsTVDay) | obesity$HrsSleep==99
```

```
obesity<- obesity[-which(remove==T), ]  
dim(obesity)
```

```
## [1] 996 17
```

Now run the main instruction

```
system.time(
result<- Bvs(formula = BMI ~ . , data=obesity, null.model = BMI ~ Age)
)
```

```
## Info. . . .
## Most complex model has 17 covariates
## From those 2 are fixed and we should select from the remaining 15
## FaObese, MoObese, WeightBorn, HeightBorn, Meals5, Vegeta, Fruit, Sex, Breastfeed, AfternoonSnack, Ac
## The problem has a total of 32768 competing models
## Of these, the 10 most probable (a posteriori) are kept
## Working on the problem...please wait.

## user system elapsed
## 4.471 0.524 3.864
```

To obtain a summary of the result, run

```
summary(result)
```

```
##
## Inclusion Probabilities:
##           Incl.prob. HPM MPM
## FaObese           1 * *
## MoObese           1 * *
## WeightBorn       0.9918 * *
## HeightBorn       0.9849 * *
## Meals5           0.9975 * *
## Vegeta           0.3647
## Fruit            0.3258
## Sex              0.5437 *
## Breastfeed       0.4174
## AfternoonSnack   0.8263 * *
## Activ            0.8774 * *
## candies          0.9857 * *
## HrsPCDay         0.3848
## HrsTVDay         0.9999 * *
## HrsSleep         0.4207
## ---
## Code: HPM stands for Highest posterior Probability Model and
## MPM for Median Probability Model.
##
```

A relevant plot with the relations between covariates can be obtained with

```
plot(result, option="n")
#be sure to enlarge your window to accomodate this large plot
```

What follows is how the sensitivity study was performed

```
#Results with different priors and SB prior for model space:
result.robust.SB<- Bvs(formula = BMI ~ . , data=obesity, null.model = BMI ~ Age)
result.gn.SB<- Bvs(formula = BMI ~ . , data=obesity, null.model = BMI ~ Age,
prior.betas="gZellner")
result.JZS.SB<- Bvs(formula = BMI ~ . , data=obesity, null.model = BMI ~ Age,
prior.betas="ZellnerSiow")
result.liang.SB<- Bvs(formula = BMI ~ . , data=obesity, null.model = BMI ~ Age,
prior.betas="Liangetal")
```

```

result.intrinsic.SB<- Bvs(formula = BMI ~ . , data=obesity, null.model = BMI ~ Age,
prior.betas="intrinsic")
result.FLS.SB<- Bvs(formula = BMI ~ . , data=obesity, null.model = BMI ~ Age,
prior.betas="FLS")
#
#Results with different priors and Constant prior for model space:
result.robust.C<- Bvs(formula = BMI ~ . , data=obesity, null.model = BMI ~ Age,
prior.models="Constant")
result.gn.C<- Bvs(formula = BMI ~ . , data=obesity, null.model = BMI ~ Age,
prior.betas="gZellner", prior.models="Constant")
result.JZS.C<- Bvs(formula = BMI ~ . , data=obesity, null.model = BMI ~ Age,
prior.betas="ZellnerSiow", prior.models="Constant")
result.liang.C<- Bvs(formula = BMI ~ . , data=obesity, null.model = BMI ~ Age,
prior.betas="Liangetal", prior.models="Constant")
result.intrinsic.C<- Bvs(formula = BMI ~ . , data=obesity, null.model = BMI ~ Age,
prior.betas="intrinsic", prior.models="Constant")
result.FLS.C<- Bvs(formula = BMI ~ . , data=obesity, null.model = BMI ~ Age,
prior.betas="FLS", prior.models="Constant")

results<- matrix(0, nr=15, nc=12)
colnames(results)<- c("Robust-SB", "ZS-SB", "liang-SB", "intrinsic-SB", "FLS-SB", "g=n-SB",
"Robust-C", "ZS-C", "liang-C", "intrinsic-C", "FLS-C", "g=n-C")
rownames(results)<- names(result.robust.SB$inclprob)
results[,1]<- result.robust.SB$inclprob
results[,2]<- result.JZS.SB$inclprob
results[,3]<- result.liang.SB$inclprob
results[,4]<- result.intrinsic.SB$inclprob
results[,5]<- result.FLS.SB$inclprob
results[,6]<- result.gn.SB$inclprob

results[,7]<- result.robust.C$inclprob
results[,8]<- result.JZS.C$inclprob
results[,9]<- result.liang.C$inclprob
results[,10]<- result.intrinsic.C$inclprob
results[,11]<- result.FLS.C$inclprob
results[,12]<- result.gn.C$inclprob

round(results, 2)

```

## riboflavin dataset

The dataset is distributed with the hdi R package

```

library(hdi) #this is the package that contains the data

## Loading required package: scalreg
## Loading required package: lars
## Loaded lars 1.2
packageVersion("hdi")

## [1] '0.1.7'

```

```
data(riboflavin)
dim(riboflavin)
```

```
## [1] 71 2
```

```
dataribo<- as.data.frame(cbind(y=riboflavin$y, riboflavin$x))
```

The results given in the paper are obtained with a fix initial seed and three different chains as follows (each run takes approximately 2.3 hours)

```
set.seed(16091956)
```

```
init.model<- rep(0, 4088)
init.model[sample(1:4088, 2, rep=F)]<- 1
resultRB.chain1<- GibbsBvs(formula=y~., data=dataribo, time.test=F, init.model=init.model,
  n.iter=5e4)
```

```
set.seed(2*16091956)
init.model<- rep(0, 4088)
init.model[sample(1:4088, 2, rep=F)]<- 1
resultRB.chain2<- GibbsBvs(formula=y~., data=dataribo, time.test=F, init.model=init.model,
  n.iter=5e4)
```

```
set.seed(3*16091956)
init.model<- rep(0, 4088)
init.model[sample(1:4088, 2, rep=F)]<- 1
resultRB.chain3<- GibbsBvs(formula=y~., data=dataribo[,sample(1:4089)], time.test=F,
  init.model=init.model, n.iter=5e4)
```

Exploring the results:

```
which(resultRB.chain1$inclprob>0.5)
#YOAB_at
# 2564
which(resultRB.chain1$HPMbin==1)
#ARGB_at YFII_at YHEA_at YLXQ_at YOAB_at YXLE_at
# 69 1849 2035 2459 2564 4004
resultRB.chain1$inclprob[which(resultRB.chain1$inclprob>0.1)]
#ARGB_at ARGF_at CARB_at YFII_at YHDZ_at YISU_at YOAB_at YXLD_at YXLE_at
#0.13912 0.41146 0.20738 0.17516 0.10916 0.16770 0.97034 0.39972 0.42798
sum(resultRB.chain1$inclprob<.005)
#3948
```

```
which(resultRB.chain2$inclprob>0.5)
#YOAB_at
# 2564
which(resultRB.chain2$HPMbin==1)
#ARGB_at YFII_at YHEA_at YLXQ_at YOAB_at YXLE_at
# 69 1849 2035 2459 2564 4004
resultRB.chain2$inclprob[which(resultRB.chain2$inclprob>0.1)]
#ARGB_at ARGF_at CARB_at YFII_at YHDZ_at YISU_at YOAB_at YXLD_at YXLE_at
#0.17072 0.43930 0.13498 0.19312 0.11830 0.14204 0.96748 0.39922 0.39408
```

```
which(resultRB.chain3$inclprob>0.5)
#YOAB_at
# 2834
```

```

which(resultRB.chain3$HPMbin==1)
#YFII_at YHEA_at ARGB_at YXLE_at YOAB_at YLXQ_at
# 154 482 618 2372 2834 2904

resultRB.chain3$inclprob[which(resultRB.chain3$inclprob>0.1)]
#YFII_at ARGB_at YXLD_at YXLE_at ARGF_at CARB_at YOAB_at YISU_at YHDZ_at
#0.21230 0.14284 0.38002 0.43656 0.41930 0.18520 0.96680 0.14508 0.11524

```

Regarding singular models, with the following study we see that these models barely accumulate any posterior mass.

```

#we haven't sampled any of these models:
sum(resultRB.chain1$postprobdim[72:4089])
#0

#Is the correction for the part of the model space
#k>n needed?
corrected.inclprob<- pltltn(resultRB.chain1)
#Info: Use this function only for problems with p>>n (not just p>n)
#Estimate of the posterior probability of the
# model space with singular models is: 0

```

To estimate the posterior probabilities of the best models found, we must run

```

p<- 4088
#Dimensions of sampled models:
modeldims<- rowSums(resultRB.chain1$modelslogBF[,-(p+1)])
#posterior probabilities (log scale and except for constant)
ker.postprob<- resultRB.chain1$modelslogBF[, (p+1)]-lchoose(p, modeldims)
#ranked<- sort(ker.postprob, decreasing=T)

#the five largest log-posterior probs and the model with that prob
fivebest.pp<- rep(0, 5); fivebest<- rep(0,5)
fivebest.pp[1]<- max(ker.postprob); fivebest[1]<- min(which(ker.postprob==fivebest.pp[1]))
aux<- ker.postprob
for (j in 2:5){
  aux[aux==max(aux, na.rm=T)]<- NA
  fivebest.pp[j]<- max(aux, na.rm=T)
  fivebest[j]<- min(which(ker.postprob==fivebest.pp[j]))
}

ker.postprob[fivebest]
#[1] 27.33218 25.79967 25.57203 25.23315 25.09847
exp(ker.postprob[fivebest[1]]-ker.postprob[fivebest])
#[1] 1.000000 4.629780 5.813260 8.158234 9.334383
#To inspect which are the five best models:
which(resultRB.chain1$modelslogBF[fivebest[1], 1:p]==1)
#ARGB_at YFII_at YHEA_at YLXQ_at YOAB_at YXLE_at
# 69 1849 2035 2459 2564 4004
which(resultRB.chain1$modelslogBF[fivebest[2], 1:p]==1)
#ARGF_at YHDZ_at YOAB_at YXLD_at
# 73 2034 2564 4003
which(resultRB.chain1$modelslogBF[fivebest[3], 1:p]==1)
#DNAJ_at YOAB_at YRHE_at YXLE_at
# 315 2564 3173 4004

```

```

which(resultRB.chain1$modelslogBF[fivebest[4], 1:p]==1)
#ARGB_at YFII_at YHEA_at YLXQ_at YOAB_at YXLD_at
#      69      1849      2035      2459      2564      4003
which(resultRB.chain1$modelslogBF[fivebest[5], 1:p]==1)
#DNAJ_at YOAB_at YRHE_at YXLD_at
#      315      2564      3173      4003

#The estimation of the posterior probabilities (assuming zero for singular models)
#based on the estimation of the normalizing constant C
#which, for p>n, is only for regular models.
n<- 71
postprob<- resultRB.chain1$modelslogBF[, (p+1)]-lchoose(p, modeldims)-log(p+1)-log(resultRB.chain1$C)
exp(postprob[fivebest])
#[1] 0.010523719 0.002273049 0.001810295 0.001289951 0.001127415

```

## OBICE study, part II

This part of the analysis was performed with the R package glmBfp. For reasons that are unknown to the authors, at the time of writing this report, the package was not directly available through CRAN. The version used here is available through the archive web <https://cran.r-project.org/src/contrib/Archive/glmBfp/>

```

library(glmBfp)
packageVersion("glmBfp")

```

```
## [1] '0.0.60'
```

Include in the dataset the dependent variable with the classification of a child in the case group (obese) or in the control group (not obese)

```

y<- OBICE$Caso01
y<- y[-1031]
y<- y[-which(remove==T)]
length(y)

```

```
## [1] 996
```

```
#996
```

```
obesity$y<- y
```

Now run the main command:

```

formula.full<- y ~ uc(FaObese) + uc(MoObese) + uc(WeightBorn) + uc(HeightBorn) + uc(Meals5) +
uc(Vegeta) + uc(Fruit) + uc(Age) + uc(Sex) + uc(Breastfeed) + uc(AfternoonSnack) +
uc(Activ) + uc(candies) + uc(HrsPCDay) + uc(HrsTVDay) + uc(HrsSleep)

system.time(
res<- glmBayesMfp(formula=formula.full, family=binomial("logit"), fixedg=dim(obesity)[1],
priorSpecs = list(gPrior = HypergPrior(), modelPrior = "dependent"), data=obesity, method="
)

```

```
## Warning: no fractional polynomial terms in formula
```

```
## Starting with computation of every model...
```

```
## 0%-----100%
```

```
## -----
```

```
## Actual number of possible models: 65536
```

```
## Number of non-identifiable models: 0
## Number of saved possible models: 65536

## user system elapsed
## 30.615 0.261 30.925
```

Finally, analyze the results:

```
table.res<- as.data.frame(res)
#the first five best models:
table.res[1:5,]
```

```
## posterior logMargLik logPrior Activ AfternoonSnack Age Breastfeed candies
## 1 0.34756325 -543.9550 -11.21527 FALSE FALSE FALSE FALSE TRUE
## 2 0.06736415 -544.9897 -11.82141 TRUE FALSE FALSE FALSE TRUE
## 3 0.05876660 -546.6078 -10.33981 FALSE FALSE FALSE FALSE TRUE
## 4 0.04182980 -545.4662 -11.82141 FALSE FALSE TRUE FALSE TRUE
## 5 0.03557570 -547.1097 -10.33981 FALSE FALSE FALSE FALSE TRUE
## FaObese Fruit HeightBorn HrsPCDay HrsSleep HrsTVDay Meals5 MoObese Sex
## 1 TRUE FALSE FALSE FALSE FALSE TRUE TRUE TRUE FALSE
## 2 TRUE FALSE FALSE FALSE FALSE TRUE TRUE TRUE FALSE
## 3 TRUE FALSE FALSE FALSE FALSE FALSE TRUE TRUE FALSE
## 4 TRUE FALSE FALSE FALSE FALSE TRUE TRUE TRUE FALSE
## 5 TRUE FALSE FALSE FALSE FALSE TRUE FALSE TRUE FALSE
## Vegeta WeightBorn
## 1 FALSE FALSE
## 2 FALSE FALSE
## 3 FALSE FALSE
## 4 FALSE FALSE
## 5 FALSE FALSE
```

```
#the inclusion posterior probabilities:
round(inclusionProbs(res), 2)
```

```
## Activ AfternoonSnack Age Breastfeed candies
## 0.18 0.16 0.12 0.03 1.00
## FaObese Fruit HeightBorn HrsPCDay HrsSleep
## 1.00 0.07 0.06 0.03 0.03
## HrsTVDay Meals5 MoObese Sex Vegeta
## 0.86 0.87 1.00 0.03 0.03
## WeightBorn
## 0.11
```