TAKE HOME EXAM

PhD in Business Economics
Professor: Hedibert Freitas Lopes

Course: Bayesian Econometrics
Due date: June 12th, 2018.

Part I: Bayesian and non-Bayesian regularization in action

The file `wage2-wooldridge.txt`, from Blackburn and Neumark (1992)[1] contains information on monthly earnings, education, several demographic variables, and IQ scores for 935 men in 1980. One can argue that IQ accounts for omitted ability bias in a regression equation for log wage.

1. `wage:` monthly earnings
2. `hours::` average weekly hours
3. `iq:` IQ score
4. `kww:` knowledge of world work score
5. `educ:` years of education
6. `exper:` years of work experience
7. `tenure:` years with current employer
8. `age:` age in years
9. `married:` $= 1$ if married
10. `black:` $= 1$ if black
11. `south:` $= 1$ if live in south
12. `urban:` $= 1$ if live in SMSA
13. `sibs:` number of siblings
14. `brthord:` birth order
15. `meduc:` mother's education
16. `feduc:` father's education
17. `lwage:` natural log of wage

The variables `married`, `black`, `south` and `urban` are dummies. Below you can find a small R script to read and get the data ready for the regularised regressions. As you will see, the X matrix has $n = 935$ individuals and $p = 58$ potential covariates.

[1] Blackburn and Newmark (1992) Unobserved ability, efficiency wages and interindustry wage, *Quarterly Journal of Economics*, 107, 1421-36. See also Wooldridge (2012) *Introductory Econometrics: A Modern Approach* (5th edition) South-Western, Cengage Learning.

```
# R script
mydata = read.table("http://hedibert.org/wp-content/uploads/2014/02/wage2-wooldridge.txt",header=FALSE)

colnames(mydata) = c("wage","hours","iq","kww","educ","exper","tenure","age","married",
                     "black","south","urban","sibs","brthord","meduc","feduc","lwage")

y = mydata[,ncol(mydata)]
X = model.matrix(lwage ~ kww + hours*exper*tenure*age + sibs +
                 (iq+educ+married + black + south + urban)^3-1,data=mydata)

n = nrow(X)
p = ncol(X)
```

Your job: Your task is, assuming Gaussianity of the residuals, fit the model

$$y|X,\beta,\sigma^2 \sim N(X\beta,\sigma^2)$$

by maximum likelihood (OLS), penalized maximum likelihood (Lasso), Bayesian ridge, Bayesian lasso and Bayesian horseshoe. Compare your findings. First use the whole data, i.e. $n = 935$ observations. Then, redo your analysis by randomly selecting 50% of the data for fitting and the remainder for testing (via root mean square fit error).

Note: There are R packages/functions to fit all above models. Recall the `lm` function to fit ols regressions. For lasso and ridge regression you can use the package `glmnet`. For the Bayesian lasso, ridge and horseshoe you can use the R package `bayeslm`. You can find R scripts and more details in our set of slides on `http://hedibert.org/wp-content/uploads/2018/05/multiplelinearregression.pdf`

Part II: Reading, summarising and presenting a paper on regularisation

I have upload five regularisation-related papers to our course webpage. Each group (of two students) are supposed to write a three page summary of the paper and be ready to present it (with slides!) before the class on Tuesday, June 12th. Each group will have up to 15 minutes to present and both members of the group have to talk.

1. Mariana Orsini & Alexander Chow
   Tibshirani (1996) Regression shrinkage and selection via the Lasso

2. Fernando Deodato & Felipe Franca
   Park-Casella (2008) The Bayesian Lasso

3. Fernando Tassinari & Rafael Rocha
   Polson-Scott-Windle (2014) The Bayesian bridge

4. Raphael Gondo & Alysson Portella
   Kastner-Huber (2017) Sparse Bayesian vector autoregressions in huge dimensions

5. Leila Pereira & Rafael Pucci
   Kaufmann-Schumacher (2017) Identifying relevant/irrelevant variables in sparse factor models