University of Geneva GSEM
Model Selection in High Dimensions Spring 2018
Prof. Maria-Pia Victoria-Feser **Practical 4**

Correlation, Efron's q-class and ROC Curves

1 Simulations in a Correlated Environment

In this exercise we would like to investigate the changes induced by the presence of correlated predictors to the simulation setting outlined in *Practical 3*.

(a) Retrieve your codes done for Exercise 1 and 2 of *Practical 3*. You still need to generate from a MVN (multivariate normal) a matrix $\mathbf{X_{n*p}}$ with $n = 1000$ and $p = 5$ but now the scale matrix is of an autoregressive form $\mathbf{\Sigma} = [\sigma_{lm}]_{l,m=1,\ldots,p}$ with $\sigma_{lm} = \rho^{|l-m|}$.

(b) Reproduce the results of Exercise 1 and 2 for $\boldsymbol{\rho} = [0.2\ 0.5\ 0.7]$ ($\rho = 0$ corresponds to the identity case that you have already treated).

(c) In the specific simulation setting of *Practical 3*, we have found a better performance of AIC with respect to CV in light of three criteria: the proportion of times the correct model is selected (*Exact*), the proportion of times the selected model contains the correct one (*Correct*) and the average number of selected regressors (*Average* $\sharp$). Do you observe any changes in the results as we introduce correlation in the simulation setting? If so, why do you think it is the case?

2 Theoretical Part: q-class of Error Measures

Look at **paragraph 2.6.3** of the e-book and, with paper and pencil, show that:

(a) When $q(v) = v(1 - v)$ in **(2.3)** $\Rightarrow Q(u, v)$ is the quadratic error loss function.

(b) When $q(v) = \min\{v, (1 - v)\}$ in **(2.3)** $\Rightarrow Q(u, v)$ is the misclassification error loss function.

(c) When $q(v) = -2[v \log(v) + (1 - v) \log(1 - v)]$ in **(2.3)** $\Rightarrow Q(u, v)$ is the binomial deviance or twice the Kullback-Leibler divergence loss function.

Note that all the above derivations should be done in the binary case (i.e. logistic regression framework). Assume also that $q(0) = 0$ and $q(1) = 0$ to simplify the calculations.

3 ROC Curves

Look at the *Leukemia dataset* description on the e-book. In this exercise we will work on *data_leukemia_reduced* which contains a subset of 11 eleven predictors among the 3571 present in the *leukemia_small.csv*. These variables have been selected, because of their importance, by the binary lasso which is a shrinkage method that will be discussed later on in the course. After having load the dataset, perform the following steps:

(a) Fit the appropriate GLM for the situation at hand using all the available predictors

(b) Read the **ROC curve** section of the e-book. Then find the TPR (i.e. true positive rate), FPR (i.e. false positive rate), TNR (i.e. true negative rate), FNR (i.e. false negative rate) of the fitted values found at point (a) with a cut-off value $c = 0.5$.

(c) For a given cut-off grid of values, that you can choose as you wish, plot the ROC curve relative to the estimated model at point (a).

(d) Check the quality of your result at point (c) with the R package *pROC*.