Statistical Learning

Exercises: Multiple regression: model selection criteria, ridge regression and LASSO.

1. Prove that the trace of the symmetric real matrix is equal to the sum of its eigenvalues (Hint : use the spectral decomposition and the circular property of the trace).

2. Consider the real matrix X of the dimension $n \times p$.

   a) Prove that X'X is semipositive definite and that its eigenvalues are larger or equal to zero.

   b) Prove that when p>n than at least one eigenvalue of X'X is equal to zero (i.e. X'X is singular).

3. Your data contains 10 variables. You fit 10 regression models including the first variable, the first two variables, etc. The residual sums of squares for these 10 consecutive models are equal to (1731, 730, 49, 38.9, 32, 29, 28.5 27.8, 27.6, 26.6). The sample size is equal to 100. Which of these 10 models will be selected by AIC ? And which model will be selected by BIC or RIC? Assume that the standard deviation of the error term is known; $\sigma = 1$.

4. Assuming the orthogonal design ($X'X = I$) and $n = p = 10000$ calculate the expected number of false discoveries for AIC, BIC and RIC, when none of the variables is really important (i.e. $p_0 = p$).

5. When would you use AIC ? BIC ? RIC ?

6. Derive the formula for the bias, variance and mse of the ridge regression estimate under the orthogonal design (i.e when X'X=I). Compare to the respective values for the least square estimator.

7. For a given data set with 40 explanatory variables the residual sums of squares from the least squares method and the ridge regression are equal to : 4.5 and 11.6, respectively. For the ridge regression the trace of $X(X'X + \gamma I)^{-1}X'$ is equal to 32. Which of these two methods yields the better estimated prediction error.

8. Given X'X=I calculate the expected value of false discoveries and the power of LASSO.

9. Consider adaptive LASSO with $\lambda_i = w_i\lambda$.

   i) How can you calculate adaptive LASSO estimator using the numerical solver for LASSO (like *glmnet*).

   ii) In the orthogonal case (X'X=I) calculate the value of the adaptive LASSO estimator for the specific coordinate of the beta vector.

   iii) The ordinary least squares estimator of β_1 under the orthognal design (X'X=I) is equal to 3 and the LASSO estimator of this parameter is equal to 2. What is the value of the adaptive LASSO estimator of β_1 if we use the same value of λ and the weight for X_1 is $w_1 = 1/4$.

Project1 : James-Stein estimator and Prediction Error in Multiple Regression

1. The data set Lab3.Rdata contains the matrix xx with expressions of 300 genes for 210 individuals.

   a) Preprocess the data by standardizing each gene expression such that its mean is preserved but the standard deviation is equal to 1.

b) The average expression level over all genes is close to 10. Subtract 10 from all standardized gene expressions, such that the new standardized gene expressions oscillate around zero.

c) Use the first five individuals to estimate the vector of the means for the "standardized" gene expressions. Use the maximum likelihood estimator and both James-Stein estimators (shrunk towards zero and towards a common mean). Which σ^2 should you use when calculating the James-Stein estimators?

d) Verify the accuracy of these estimators by comparing them to the average gene expressions for the remaining 205 individuals.

i) Plot the scatter plots of the estimators versus the average gene expressions. You can use one graph and represent different estimates with different colors. Supplement the plot with the identity line $y = x$ (ideal location of your estimates).

ii) Calculate the squared error $||\hat{\mu} - \mu||^2$ for all three estimators. Here μ represents the vector of average gene expressions for the remaining 205 individuals.

iii) Comment on the results using the theory learned in class.

2. Generate the design matrix $X_{1000\times 950}$ such that its elements are iid random variables from $\mathcal{N}(0, \sigma = \sqrt{\frac{1}{1000}})$. Then generate the vector of the response variable according to the model $Y = X\beta + \epsilon$, where $\beta = (3, 3, 3, 3, 3, 0, \ldots, 0)^T$ and $\epsilon \sim \mathcal{N}(0, I)$.

i) 2 first variables

ii) 5 first variables

iii) 10 first variables

iv) 100 first variables

v) 500 first variables

vi) all 950 variables.

For each of the considered models:

a) Estimate β with the Least Squares method and calculate residual sum of squares and the conditional (given $\hat{\beta}$) expected value of the prediction error $PE = E||X(\beta - \hat{\beta}) + \epsilon^*||^2 = ||X(\beta - \hat{\beta})||^2 + n\sigma^2$. Here $\epsilon^* \sim \mathcal{N}(0, I)$ is a new noise vector, independent on the training sample.

b) Use the residual sum of squares to estimate PE assuming that σ is known and replacing σ with its regular unbiased estimator.

c) Estimate PE using leave-one-out cross-validation (do not perform analysis 1000 times but apply the formula for leave-one-out cross-validation error provided in class).

d) Which model is best in terms of PE? How well is PE estimated by the three estimators? Which model would you choose using these estimators?

e) Repeat the above calculations 30 times and for each of the considered models compare the boxplots of $\hat{PE} - PE$ for three estimates of PE, mentioned above. Comment on the results referring to the theory learned in class.

Project 2: Multiple regression - model selection and regularization

Generate the design matrix $X_{1000\times 950}$ such that its elements are iid random variables from $N\left(0, \sigma = \frac{1}{\sqrt{1000}}\right)$. Then generate the vector of the response variable according to the model

$$Y = X\beta + \epsilon \ ,$$

where $\beta_1 = \ldots = \beta_k = 6$, $\beta_{k+1} = \ldots = \beta_p = 0$ with $k = 20$, and $\epsilon \sim N(0, I)$.

Analyse this data using

- mBIC2 criterion from the *bigstep* package [You can also use RIC if you find it reasonably implemented in some other package]
- Ridge with the tuning parameter selected by cross-validation.
- LASSO with the tuning parameter selected by cross-validation (consider the results for *lambda.1se* and *lambda.min*).
- LASSO with the tuning parameter $\lambda = \Phi^{-1}\left(1 - \frac{0.1}{2p}\right)$ (In glmnet you need to divide this by n. Do you know why ?)
- SLOPE with the BH sequence of the tuning parameters $\lambda_i = \Phi^{-1}\left(1 - \frac{0.1i}{2p}\right)$ (again in the SLOPE package you need to divide this by n).

For each of these methods calculate the square estimation errors $||\hat{\beta} - \beta||^2$ and $||X(\hat{\beta} - \beta)||^2$. In case of LASSO and SLOPE consider also estimators obtained by performing the regular least squares fit within the selected model. For all methods apart from ridge calculate also the False Discovery Proportion and the True Positive Proportion (Power).

If you have time you may repeat it 10 times and compare the average values of squared errors and FDP and TPP. You may also play with different parameter setups (for example increasing k or changing the signal magnitude). You may also introduce the correlation between explanatory variables.

Malgorzata Bogdan