Some topics in the analysis of Large Data Sets

Multiple regression and principle components analysis

1. Let $n = 500$ and $p \in \{50, 100, 200, 450\}$. For each of these combinations generate one matrix X such that its rows are iid random vectors from $N(0, \frac{1}{n}\Sigma)$, where

   a) $\Sigma = I$

   b) Σ is in a compound symmetry form with $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.5$ for $i \neq j$.

   For each of the above combinations of p and Σ generate the response vector Y according to the formula

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

   where the vector β contains 50 nonzero elements of value 3.5 randomly placed among p positions and $\epsilon \sim N(0, I)$.

   i) Estimate β using the least squares method and the ridge regression with the coefficient $\gamma = 0.15$.

   ii) Calculate $||X(\hat{\beta} - \beta)||^2$ and the unbiased estimate of $E||X(\hat{\beta} - \beta)||^2$ (check the formula for SURE).

   Repeat the simulations 200 times (randomly generate ϵ, keep X and β fixed). For each scenario

   a) Plot $||X(\hat{\beta} - \beta)||^2$ versus the respective prediction by SURE. Calculate the correlation coefficient between these two values as well as the average value of the distance between $||X(\hat{\beta} - \beta)||^2$ and SURE.

   b) Compare boxplots of $||X(\hat{\beta} - \beta)||^2$ for ridge and least squares regressio.

2. Consider the setup with $n = p = 200$. Generate one matrix X such that its rows are iid random vectors from $N(0, \frac{1}{n}\Sigma)$, where

   a) $\Sigma = I$

   b) Σ is in a compound symmetry form with $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.7$ for $i \neq j$.

   For each of these two setups generate the response vector Y according to the formula

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

   where the vector β contains 50 nonzero elements of value 3.5 randomly placed among p positions and $\epsilon \sim N(0, I)$. For each of the above setups identify the "best" model using AIC criterion and

a) Fast forward strategy in *bigstep* package

b) Default multiple forward strategy in *bigstep* package

c) Estimate the regression coefficients using the ridge regression and LASSO with the tuning parameters selected by crossvalidation (*cv.glmnet* function in *glmnet* package) and form the sequence of nested models according to the absolute values of estimated regression coefficients. Then apply AIC on this sequence.

Compare time of these three approaches and the values of AIC for selected models (make sure you are using the same formula for AIC).

Repeat the above 100 times and report average time and the average value of AIC in selected models. Summarize critically.

3. Use the setup of Problem 2 with $k = 10$ and $k = 40$ of nonzero effects. Consider both the correlated and uncorrelated case. Identify the best model using

a) mBIC2 with the default search option in *bigstep*

b) mBIC2 based on the sequence of nested model created based on cross-validated LASSO

c) Model free knockoffs based on cross-validated LASSO and ridge regression at FDR level $= 0.2$.

d) SLOPE with the "gaussian" sequence at FDR level 0.2.

e) Adaptive SLOPE at FDR level=0.2

For each of these approaches calculate

a) False Discovery Proportion

b) True Positive Proportion

c) $SSE(\hat{\beta}) = ||\hat{\beta} - \beta||^2$.

When using mBIC2 estimate β by the least squares regression in the selected model.

For LASSO and ridge regression, consider the full estimate obtained by cross-validation with knockoffs (reduced to the first p coordinates) as well as the "hard thresholded" estimate obtained by assigning zeros to the coefficients not selected by knockoffs.

For SLOPE use the regular SLOPE estimator as well as the estimate obtained by least squares regression in the selected model.

Consider one exemplary realization for each of the considered scenarios and compare the paths of SLOPE and LASSO solutions (use SLOPE package with the length of path equal to 50).

4. Generate the data set according to the following procedure.

    1. Generate the matrix $F_{200\times 2}$ of hidden latent variables such that each of its elements comes from $N(0, sd = \frac{1}{\sqrt{n}})$.
    2. Generate the matrix $X_{200\times 50}$ of observed features as $X = FC + \epsilon$ where $C_{2\times 50}$ contains elements generated from $N(0, sd = 3)$ and $\epsilon \sim N(0, 1)$.
    3. Generate $Y = F\gamma + \epsilon$, where $\gamma = (3, 3)$.

    Perform the following analyses.

    1. Estimate $E(Y)$ using X. Use the least squares regression, ridge regression and LASSO with cross-validation and adaptive SLOPE with FDR level $q = 0.2$.
    2. Calculate $SSE = ||F\gamma - X\hat{\beta}||^2$ for all these methods. Pick the method which is the most accurate.
    3. Perform the Principal Components Analysis of the matrix X .
    4. Plot the values of scores for two first PCs for all observations. Use the red color to mark the observations for which $Y > 0$ and the blue color to mark the observations for which $Y < 0$.
    5. Predict $E(Y)$ using the multiple regression of Y on the matrix $\hat{F}_{200\times 2}$ with the first and the second principal components and calculate $SSE = ||F\gamma - \hat{F}\hat{\gamma}||^2$. Compare to the values obtained in point 2.
    6. Estimate $E(Y)$ by regression of Y on the full matrix of principal components $\hat{F}_{n\times 50}$. Use the least squares regression, ridge regression and LASSO with cross-validation, SLOPE and adaptive SLOPE at the FDR level $q = 0.1$. After using SLOPE correct your estimates by performing the least squares regression on selected PCs. (Do not do it for adaptive SLOPE). Compare $SSE = ||F\gamma - \hat{F}\hat{\gamma}||^2$ for all these methods. Summarize critically.

Malgorzata Bogdan