

Question 1.1:

Consider a high dimensional problem with n observations $\mathbf{X}_1, \dots, \mathbf{X}_n$, with each observation being p -dimensional on p variables as $\mathbf{X}_i = (X_i^{(1)}, \dots, X_i^{(p)})$, $i = 1, \dots, n$. Suppose that we are interested in investigating the behaviour of all the pairwise distances between observations using the Euclidean distance with L_2 -norm.

- (a) In R, simulate n i.i.d. observations \mathbf{X}_i from this problem with each variable $X_i^{(k)}$ generated from the uniform distribution on $[0, 1]$ with $n = 100$ and $p = 2$, using the R function `runif`. Save these into an $n \times p$ matrix of data called \mathbf{X} with n observations in rows and p variables in columns.
- (b) Calculate the pairwise distances between all the n observations based on the Euclidean distance, using the R function `dist` applied to the data matrix \mathbf{X} .
- (c) Plot a histogram of the pairwise distances, using the R function `hist`.
- (d) Repeat the above steps for $p = 10$, $p = 100$ and $p = 1000$ to show that the distances between observations tend to diverge with p .
- (e) Repeat the above analysis with the modified Euclidean distance, which is the Euclidean distance scaled by $p^{-1/2}$. Is there any improvement?

Question 1.2:

This question aims to examine the performance of the ordinary least squares (OLS) estimator in regression analysis when the dimension p gets larger. It is difficult to give a direct picture of a linear regression with dimension p larger than 2, however this can be illustrated using the following formulation. Suppose that the p -dimensional covariates \mathbf{X}_i are given by $\mathbf{X}_i = h(i/n)$, where $h : [0, 1] \rightarrow \mathbb{R}^p$ is a vector-valued function defined as $h(t) = [3 \sin(-\pi j t)]_{j=1, \dots, p}$. Consider the regression model

$$Y_i = \sum_{j=1}^p 3\beta_j \sin(-\pi j i/n) + \varepsilon_i = \mathbf{X}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

with $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)$. Note that we can equivalently rewrite this regression model as $Y_i = f_\beta(i/n) + \varepsilon_i$ where $f_\beta(t_i) = \sum_{j=1}^p 3\beta_j \sin(-\pi j t_i)$ (note that $t_i = i/n$).

- (a) In R, simulate a random sample from regression model (1) with $n = 100$ and $p = 2$. To do this, first simulate random errors $\varepsilon_1, \dots, \varepsilon_n$ from the standard normal distribution $N(0, \sigma^2 = 1)$ and also simulate the parameters β_j from the normal distribution $N(0, j^{-4})$, both using the R function `rnorm`. Then, calculate the simulated response values Y_i using the formula in regression model (1).
- (b) Fit the regression model (1) with no intercept to the simulated data in Part (a), using the R function `lm`. Extract the OLS estimates of the coefficients β_j from this fitted model and denote them by $\hat{\beta}_j$.
- (c) Using R, calculate the estimated response values \hat{Y}_i using the OLS estimates $\hat{\beta}_j$ as follows

$$\hat{Y}_i = \sum_{j=1}^p 3\hat{\beta}_j \sin(-\pi ji/n), \quad i = 1, \dots, n.$$

- (d) Using R, calculate the true response values Y_i^{true} (noise free or without random errors) using the true parameter values β_j simulated in Part (a) as follows

$$Y_i^{\text{true}} = \sum_{j=1}^p 3\beta_j \sin(-\pi ji/n), \quad i = 1, \dots, n.$$

- (e) Plot the observed or simulated response values Y_i in Part (a) versus $t_i = i/n$ where $i = 1, \dots, n$, using the R function `plot`.
- (f) Add the true regression curve (signal) to the plot, using the R function `points`.
- (g) Add the fitted least square regression curve in Part (b) to this plot, using the R function `points`.
- (h) Repeat the above steps for $p = 10$, $p = 50$ and $p = 100$. Show that the mean squared error of the OLS method increases as the dimension p gets larger, by comparing the resulting plots for the increasing values of p . This supports the theory that $\|\hat{\beta} - \beta\| \approx p\sigma^2$, as shown in the lecture notes.
- (i) What happens if we use $p = 101$ or larger such as $p = 200$?