Figure 1: Set of four colinear points A, B, C and D

Homework 4

posted April 9, 2020
due April 26, 2020

1. Four colinear points A, B, C, D are located as above and the distances between them are measured approximatively yielding the following results:

$$AD = 89, AC = 67, BD = 53, AB = 35, \text{ and } CD = 20.$$

We need to determine the length of the segments $r_1 = AB$, $r_2 = BC$, and $r_3 = CD$.

The results are inconsistent because if we use the last three equations

$$\begin{aligned} r_1 + r_2 + r_3 &= 89 \\ r_1 + r_2 &= 67 \\ r_2 + r_3 &= 53 \\ r_1 = 35 \\ r_3 = 20 \end{aligned}$$

we have $r_1 = 35$, $r_2 = 33$ and $r_3 = 20$. However, the first two equations yield $x_1 + x_2 + x_3 - 89 = -1$ and $x_1 + x_2 - 67 = 1$.

Write the above system in matrix form $A\mathbf{r} = \mathbf{b}$, where $\mathbf{r} = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}$, $A \in \mathbb{R}^{5\times 3}$ and $\mathbf{b} \in \mathbb{R}^5$ and determine $\mathbf{r}$ such that $\| A\mathbf{r} - \mathbf{b} \|$ is minimal.

Let $B \in \mathbb{R}^{m\times n}$ be a matrix that contains input data of m experiments involving n variables. Note that the matrix B can be written as a set of m rows $B = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_m \end{pmatrix}$, where $\mathbf{u}_i \in \mathbb{R}^n$ contains the input values of the variables for the i^{th} experiment. Also, B can be written as $B = (\mathbf{b}^1 \cdots \mathbf{b}^n)$, where each column $\mathbf{b}^j$ contains the values of the variable x_j in each of the m experiments.

The *average* of B is the vector $\tilde{\mathbf{u}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{u}_i$, that is, the averge of the rows of the matrix.

The matrix B is *centered* if $\tilde{\mathbf{u}} = (0, 0, \ldots, 0)$.

2. Prove that the matrix $H_m = I_m - \frac{1}{m}\mathbf{1}_m\mathbf{1}'_m \in \mathbb{R}^{m\times m}$ (known as the centering matrix) is symmetric and idempotent, that is, $H'_n = H_n$ and $H_nH_n = H_n$.

3. Prove that the matrix $\hat{B} = H_mB$ is centered.

4. Let $B \in \mathbb{R}^{m\times n}$ and $\mathbf{y} \in \mathbb{R}^m$ the data used in linear regression. Suppose that B is centered and define the matrix $\hat{B} = \begin{pmatrix} B \\ \sqrt{\lambda} I_n \end{pmatrix} \in \mathbb{R}^{(m+n)\times n}$ and $\hat{\mathbf{y}} = \begin{pmatrix} \mathbf{y} \\ \mathbf{0}_n \end{pmatrix} \in \mathbb{R}^{m+n}$.

   Prove that the ordinary regression applied to this data amounts to ridge regression applied to B and $\mathbf{y}$.

5. Study the GLmnet Vignette (a description of the `glmnet` R package) which is posted on the web site. Install this package and also, the package `ggplot2`. The dataset `diamonds` is a part of `ggplot2`. This data gives the price of a diamond as a function of the carat weight, cut, color, etc.

   Apply multiple linear regression to this dataset using the `glmnet` package and study the effect of at least three model formulas that express the price of a diamond on other variables.

   You need to install and upload both `glmnet` and `ggplot2`.