IE452/IE552: Algebraic and Geometric Methods in Data Analysis

Project Report

Batıhan Akça – 21502824
Emre Can Durmuş – 21601443

Question 1:

For constructing a suitable subspace, the Random Projection is used where:

$$f: R^d \to R^k$$

$$f(v) = (u_{1.}v, u_{2.}v, \dots, u_{k.}v)$$

- k Gaussian Vectors $u_i \sim Normal(\mu = 0, \sigma^2 = 1)$ are generated.
- v is one of the rows/vectors from the dataset $D_{7352\ X\ 561}$.

$$D_{7352\ X\ 561} * U_{561\ X\ k} = R_{7352\ X\ k}$$

$$U_{561\ X\ k}\ contains\ u_i\ random\ vectors\ as\ colums.$$

- For verification of pairs' differences Johnson-Lindenstrauss Theorem is used:
- Where $\varepsilon = 0.1$ for our problem

$$\left[(1-\varepsilon)\sqrt{k}\left|v_i - v_j\right| \leq \left|f(v_i) - f(v_j)\right| \leq (1+\varepsilon)\sqrt{k}\left|v_i - v_j\right|\right]$$

$$with\ probability\ at\ least\ 1 - 3/2n$$

$$Upper\ bound\ for\ k : \sim O\,\frac{\left(\ln \frac{N}{\sqrt{0.05}}\right)}{0.01}\ is\ given$$

$$and\ also\ by\ Johnson - Lindenstrauss\ Theorem$$

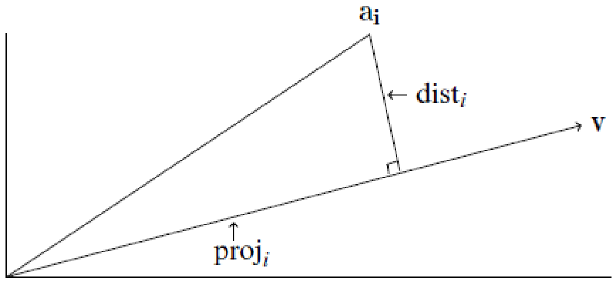
$$Lower\ bound\ for\ k \geq \frac{3(\ln N)}{c\varepsilon^2}\ \ is\ considered, c > 0$$

For k = 150 , 96.45% of the pairwise distance differences have founded between the bounds according to Johnson-Lindenstrauss Theorem.

Question 2:

$$As\ the\ measure\ of\ the\ best\ fits,\ minimizing\ sum\ of\ dist_i^{\ 2}\ \ is\ considered.$$

$$By\ Pythagorean\ Theorem$$



$$dist_i{}^2 = \left\|\overrightarrow{a_i}\right\| - (length\ of\ projection)^2$$

$$For\ k\ = 403\ \ V_k\ fit\ ratio\ to\ full\ SVD\ fit\ ratio\ calculated\ as\ 0.100756$$

$$For\ k\ = 403\ \ V_k\ fit\ \ dist_i{}^2 calculated\ as\ 9.925$$

- Since a random projection is used in the first method its $dist_i{}^2$ fit measure turned out to be higher than K-SVD fit.
- K-SVD fit converged to 0 as it increased to 561 which is the rank of the dataset matrix D.

<u>Question 3:</u>

$$\Sigma\ is\ selected\ with\ all\ vectors\ as\ the\ eigenvalues\ of\ full\ SVD\ of\ D$$

$$\Sigma = \begin{pmatrix} \sigma_1(D) & \cdots & \sigma_1(D) \\ \vdots & \ddots & \vdots \\ \sigma_d(D) & \cdots & \sigma_d(D) \end{pmatrix}$$

$$P\left[\frac{\sigma_{max}(D')}{\sqrt{N}} \geq 1.05\ \sigma_{max}(\sqrt{\Sigma}) + \sqrt{\frac{tr(\Sigma)}{n}}\right] = \ 0$$

$$P\left[\frac{\sigma_{min}(D')}{\sqrt{N}} \geq 0.95\ \sigma_{min}(\sqrt{\Sigma}) - \sqrt{\frac{tr(\Sigma)}{n}}\right] = \ 1$$

$$Where\ N = 7352,$$

$$\sigma_{max}(\sqrt{\Sigma}) = \ 1882.8573, \qquad \sigma_{min}(\sqrt{\Sigma}) = 2.8626 * 10^{-16},$$

$$\sqrt{tr(\Sigma)} = 79.4942\ and\ n = 561$$