Nico Espinosa Dice
Math189R Spring 2020
Homework 5
March 2, 2020

1 (Murphy 12.5 - Deriving the Residual Error for PCA) It may be helpful to reference section 12.2.2 of Murphy.

(a) Prove that

$$\left\| \mathbf{x}_i - \sum_{j=1}^{k} z_{ij} \mathbf{v}_j \right\|^2 = \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v}_j.$$

Hint: first consider the case when $k = 2$. Use the fact that $\mathbf{v}_i^\top \mathbf{v}_j$ is 1 if $i = j$ and 0 otherwise. Recall that $z_{ij} = \mathbf{x}_i^\top \mathbf{v}_j$.

(b) Now show that

$$J_k = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v}_j \right) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \lambda_j.$$

Hint: recall that $\mathbf{v}_j^\top \mathbf{\Sigma} \mathbf{v}_j = \lambda_j \mathbf{v}_j^\top \mathbf{v}_j = \lambda_j$.

(c) If $k = d$ there is no truncation, so $J_d = 0$. Use this to show that the error from only using $k < d$ terms is given by

$$J_k = \sum_{j=k+1}^{d} \lambda_j.$$

Hint: partition the sum $\sum_{j=1}^{d} \lambda_j$ into $\sum_{j=1}^{k} \lambda_j$ and $\sum_{j=k+1}^{d} \lambda_j$.

(a)

We have:

$$\left\|\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right\|^2 = (\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j)^T(\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j)$$

(property of Euclidean distance)

$$= (\mathbf{x}_i^T - (\sum_{j=1}^{k} z_{ij}\mathbf{v}_j)^T)(\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j)$$

(property of matrix transpose)

$$= \mathbf{x}_i^T\mathbf{x}_i - \mathbf{x}_i^T\sum_{j=1}^{k} z_{ij}\mathbf{v}_j - (\sum_{j=1}^{k} z_{ij}\mathbf{v}_j^T)\mathbf{x}_i + (\sum_{j=1}^{k} z_{ij}\mathbf{v}_j)^T(\sum_{j=1}^{k} z_{ij}\mathbf{v}_j)$$

(distributive property)

$$= \mathbf{x}_i^T\mathbf{x}_i - \mathbf{x}_i^T\sum_{j=1}^{k} z_{ij}\mathbf{v}_j - (\sum_{j=1}^{k} z_{ij}\mathbf{v}_j^T)\mathbf{x}_i + (\sum_{j=1}^{k} \mathbf{v}_j^T z_{ij}^T)(\sum_{j=1}^{k} z_{ij}\mathbf{v}_j)$$

(property of transpose of product)

$$= \mathbf{x}_i^T\mathbf{x}_i - \mathbf{x}_i^T\sum_{j=1}^{k} z_{ij}\mathbf{v}_j - (\sum_{j=1}^{k} z_{ij}\mathbf{v}_j^T)\mathbf{x}_i + \sum_{j=1}^{k} \mathbf{v}_j^T z_{ij}^T z_{ij}\mathbf{v}_j$$

$$= \mathbf{x}_i^T\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}z_{ij} - \sum_{j=1}^{k} z_{ij}z_{ij}^T + \sum_{j=1}^{k} \mathbf{v}_j^T\mathbf{v}_j z_{ij}^T z_{ij}$$

(since $z_{ij} \in \mathbb{Z}$ and multiplication is commutative in the real numbers)

$$= \mathbf{x}_i^T\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}z_{ij} - \sum_{j=1}^{k} z_{ij}z_{ij}^T + \sum_{j=1}^{k} \mathbf{v}_j^T\mathbf{v}_j(\mathbf{v}_j^T\mathbf{x}_i)(\mathbf{x}_i^T\mathbf{v}_j)$$

(since $z_{ij} = \mathbf{x}_i^T\mathbf{v}_j$)

$$= \mathbf{x}_i^T\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}z_{ij} - \sum_{j=1}^{k} z_{ij}z_{ij}^T + \sum_{j=1}^{k} \mathbf{v}_j^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_j$$

(since $\mathbf{v}_j^T\mathbf{v}_j = 1$)

$$= \mathbf{x}_i^T\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}z_{ij} - \sum_{j=1}^{k} z_{ij}z_{ij}^T + \sum_{j=1}^{k} \mathbf{v}_j^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_j$$

$$= \mathbf{x}_i^T\mathbf{x}_i - 2\sum_{j=1}^{k} z_{ij}^T z_{ij} + \sum_{j=1}^{k} z_{ij}^T z_{ij}$$

$$= \mathbf{x}_i^T\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}^T z_{ij}$$

$$= \mathbf{x}_i^T\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_j.$$

Thus, we see that $\left\|\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right\|^2 = \mathbf{x}_i^T\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_j$. ∎

(b)

We have:

$$J_k = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i^\top\mathbf{x}_i - \sum_{j=1}^{k}\mathbf{v}_j^\top\mathbf{x}_i\mathbf{x}_i^\top\mathbf{v}_j\right)$$

(by definition)

$$= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top\mathbf{x}_i) - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\mathbf{v}_j^\top\mathbf{x}_i\mathbf{x}_i^\top\mathbf{v}_j$$

(distributive property)

$$= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top\mathbf{x}_i) - \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n}\mathbf{v}_j^\top\mathbf{x}_i\mathbf{x}_i^\top\mathbf{v}_j$$

$$= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top\mathbf{x}_i) - \sum_{j=1}^{k}\mathbf{v}_j^\top\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i\mathbf{x}_i^\top)\mathbf{v}_j$$

(since $\mathbf{v}_j^\top$ does not depend on i)

$$= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top\mathbf{x}_i) - \sum_{j=1}^{k}\mathbf{v}_j^\top\mathbf{\Sigma}\mathbf{v}_j$$

(since $\mathbf{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top\mathbf{x}_i)$ by Murphy, Section 12.2.1 under Equation 12.28)

$$= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top\mathbf{x}_i) - \sum_{j=1}^{k}\lambda_j$$

(given: $\mathbf{v}_j^\top\mathbf{\Sigma}\mathbf{v}_j = \lambda_j$).

Thus, we see that $J_k = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top\mathbf{x}_i) - \sum_{j=1}^{k}\lambda_j$. ∎

(c)

We know that $J_d = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top\mathbf{x}_i) - \sum_{j=1}^{d}\lambda_j$ by the equation above, and we are also given

that $J_d = 0$. Thus, we know that:

$$0 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{x}_i) - \sum_{j=1}^{d}\lambda_j, \text{ so}$$

$$\sum_{j=1}^{d}\lambda_j = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{x}_i).$$

Next, let us consider:

$$\begin{aligned}
J_k &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{x}_i) - \sum_{j=1}^{k}\lambda_j \\
&\text{(proven above)} \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{x}_i) - (\sum_{j=1}^{d}\lambda_j - \sum_{j=k+1}^{d}\lambda_j) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{x}_i) - (\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{x}_i) - \sum_{j=k+1}^{d}\lambda_j) \\
&\text{(since } \sum_{j=1}^{d}\lambda_j = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i^\top \mathbf{x}_i)) \\
&= \sum_{j=k+1}^{d}\lambda_j.
\end{aligned}$$

Thus, we see that the error from only using $k < d$ terms is given by $J_k = \sum_{j=k+1}^{d}\lambda_j$.

■

2 (ℓ_1-Regularization) Consider the ℓ_1 norm of a vector $\mathbf{x} \in \mathbb{R}^n$:

$$\|\mathbf{x}\|_1 = \sum_i |\mathbf{x}_i|.$$

Draw the norm-ball $B_k = \{\mathbf{x} : \|\mathbf{x}\|_1 \leq k\}$ for $k = 1$. On the same graph, draw the Euclidean norm-ball $A_k = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq k\}$ for $k = 1$ behind the first plot. (Do not need to write any code, draw the graph by hand).

Show that the optimization problem

$$\begin{aligned} &\text{minimize: } f(\mathbf{x}) \\ &\quad \text{subj. to: } \|\mathbf{x}\|_p \leq k \end{aligned}$$

is equivalent to

$$\text{minimize: } f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p$$

(hint: create the Lagrangian). With this knowledge, and the plots given above, argue why using ℓ_1 regularization (adding a $\lambda\|\mathbf{x}\|_1$ term to the objective) will give sparser solutions than using ℓ_2 regularization for suitably large λ.

The graph of the norm-ball – drawn in blue – and the Euclidean norm-ball – drawn in red – is drawn below.

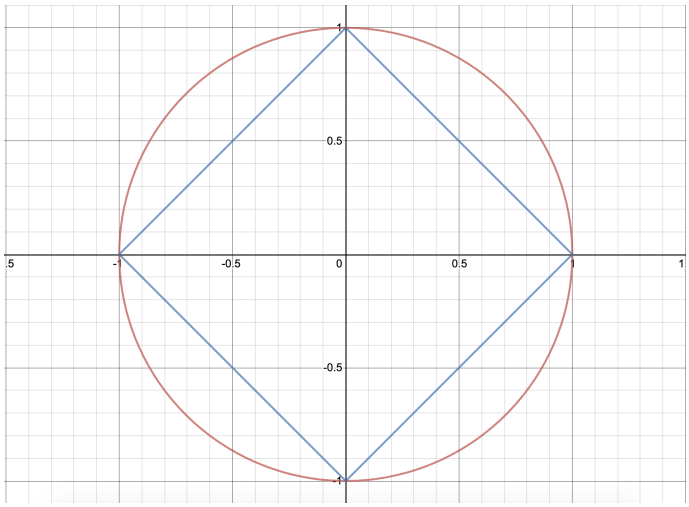

Figure 1: The red line represents the Euclidean norm-ball (ℓ_2 norm). The blue line represents the norm-ball (ℓ_1 norm). I used Desmos Graphing Calculator to plot this graph.

To minimize $f(\mathbf{x})$ subject to $\|\mathbf{x}\|_p \leq k$, we will create a Lagrangian:

$$\begin{aligned} l(\mathbf{x}, \lambda) &= f(\mathbf{x}) - \lambda(\|\mathbf{x}\|_p - k) \\ &\quad \text{(given in Murphy, Section 3.4 under Equation 3.41)} \\ &= f(\mathbf{x}) - \lambda\|\mathbf{x}\|_p + \lambda k. \end{aligned}$$

Since λk is a constant, we know that the λk term will not have an effect on our minimization of the function $f(\mathbf{x})$ subject to the given constraints. Thus, we know that minimizing $l(\mathbf{x}, \lambda) = f(\mathbf{x}) - \lambda(\|\mathbf{x}\|_p - k)$ is equivalent to minimizing $f(\mathbf{x}) - \lambda\|\mathbf{x}\|_p$. ■

We will consider the geometric perspective. As shown in Figure 1, ℓ_2 norm yields a graph that is a circle in two-dimensions, whereas the ℓ_1 norm yields a graph that is a square in two-dimensions, with corners on either the x or y axis. When considering the contours of the constraint surfaces – which are presented in the form of ellipses in Murphy, Chapter 13 – the ellipses will most likely intersect the ℓ_1 square in the corners, since the corners are the points of the square that are farthest from the origin. Since the corners of the square occur on the axes, we know that either the x or y variable is zero at these points. This corresponds to a sparser solution.

Since the ℓ_2 graph is a circle, we know that the constraint surfaces are not more likely to intersect at the points on the circle that are on either the x or y axis. Thus, they are more likely to intersect at points where y and x are both non-zero on the ℓ_2 graph than the ℓ_1 graph. Points where x and y are both non-zero corresponds to a less sparse solution.

Thus, we see that ℓ_1 regularization will give sparser solutions than using ℓ_2 regularization for suitably large λ.