Marcus Rose
Math 123 HW 2
Due 2/2/23

---

1. Let $\Sigma \in \mathbb{R}^{d\times d}$ be symmetric. Let $F : \mathbb{R}^d \to \mathbb{R}$ be the function $F(\mathbf{u}) = \mathbf{u}^T\Sigma\mathbf{u}$, where $\mathbf{u}$ is a column vector. Show

$$\frac{\partial F}{\partial \mathbf{u}} = 2\Sigma\mathbf{u}.$$

Proof First reorganizing such that

$$\frac{\partial F}{\partial \mathbf{u}} = \frac{\partial \mathbf{u}^T\Sigma\mathbf{u}}{\partial \mathbf{u}}$$

Allow $\Sigma\mathbf{u}$ to play the role of row vector. Then, $\mathbf{u}^T\Sigma\mathbf{u}$ is just a linear combination of u_i and $\Sigma_{i,:}u_i$, where $\Sigma_{i,:}$ is the *ith* row of the symmetric matrix. Hence, this is a real-valued function with a gradient that is simply Σu. Therefore,

$$\frac{\partial \mathbf{u}^T\Sigma\mathbf{u}}{\partial \mathbf{u}} = \mathbf{u}^T\Sigma^T + \mathbf{u}^T\Sigma = \mathbf{u}^T(\Sigma + \Sigma^T).$$

Because Σ is symmetric,

$$\mathbf{u}^T(\Sigma + \Sigma^T) = \mathbf{u}^T(\Sigma + \Sigma) = \mathbf{u}^T 2\Sigma \implies \frac{\partial F}{\partial \mathbf{u}} = 2\Sigma\mathbf{u}$$

□

2. The variance for a set of numbers $\{x_i\}_{i=1}^n \in \mathbb{R}$ is $\sigma^2 = \frac{1}{n}\Sigma_{i=1}^n(x_i - \mu)^2$, where μ is the sample mean. For each of the following, prove or give a counter-example:

(a) The variance is *translation invariant*, i.e. the variance of $x_1, ..., x_n$ is the same as the variance of the translated set $x_1 + T, ..., x_n + T$ for any fixed $T \in \mathbb{R}$.

Proof WTS

$$\sigma^2 = \frac{1}{n}\sum_i^n (x_i - \mu)^2 = \sum_i^n ((x_i + T) - \mu_T)^2 = \sigma_T^2,$$

where $\mu_T = \frac{1}{n}\sum_i^n (x_i + T)$.
Note that because $T \in \mathbb{R}$,

$$\mu_T = \frac{1}{n}\sum_i^n x_i + \frac{1}{n}\sum_i^n T = \frac{1}{n}\sum_i^n x_i + \frac{1}{n}Tn = \mu + T.$$

Then,

$$\sigma_T^2 = \frac{1}{n}\sum_i^n ((x_i + T) - (\mu + T))^2 = \frac{1}{n}\sum_i^n (x_i - \mu)^2 = \sigma^2.$$

□

(b) The variance is 0 iff $x_i = C, \forall i = 1, ..., n$ for some constant C (i.e. $\sigma^2 = 0$ iff all data points are equal).

Proof (⇒) If variance is zero, then all $x_i = C$. If all samples in the set are C, then the sample mean μ is necessarily C. I.e.

$$\sigma^2 = 0 = \frac{1}{n}\sum_i^n (x_i - \mu)^2 = \sum_i^n (x_i - \mu) \implies x_i = \mu,$$

as there cannot exist a sum of squares whose value is negative. And if this variance is zero in the first place (i.e. $\mu = x_i$), all $x_i = C = \mu$.
(⇐) If $x_i = C \forall i$, then the variance is 0.
If all data is the same, then $x_i = C = \mu =$. So,

$$\sigma^2 = \frac{1}{n}\sum_i^n (x_i - \mu)^2 = \frac{1}{n}\sum_i^n (C - C)^2 = 0.$$

□

(c) The variance is additive, i.e. $x_1, ..., x_n$ have variance σ_x^2 and $y_1, ..., y_m$ have variance σ_y^2, then the concatenated set has variance $\sigma_x^2 + \sigma_y^2$.

Proof Let X, Y be sampled from the sets in question (simply to write the proof in a more organized fashion, i.e. X, Y are random variables. Then, $\sigma_x^2 + \sigma_y^2 = \mathbb{E}[(X + Y)(\mu_X + \mu_Y)]^2$, which is

$$\mathbb{E}[(X - \mu_X + Y - \mu_Y)^2] = \mathbb{E}[X - \mu_X)^2 + 2(X - \mu_X)(Y - \mu_Y) + (Y - \mu_Y)^2] = \sigma_X^2 + 2\text{Cov}(X, Y) + \sigma_Y^2$$

Hence, variance is not additive.

□

3. A matrix $A \in \mathbb{R}^d$ is said to be positive semi-definite if $\mathbf{y}^T A\mathbf{y} \geq 0$ for all $\mathbf{y} \in \mathbb{R}^{d\times 1}$. The matrix A is said to be positive definite if it is positive semi-definite and $\mathbf{y}^T A\mathbf{y} = 0$ iff $\mathbf{y} = \mathbf{0}$.

(a) Let $\{x_i\}_{i=1}^n \in \mathbb{R}$ be data. Let Σ be the covariance matrix. Prove Σ is positive semi-definite.

Proof The covariance matrix of $\mathbf{x}$ is given by

$$\text{Cov}\frac{1}{n-1}\mathbb{E}[x_i - \mu]^2 = \frac{1}{n-1}\sum_{i=1}^n [x_i - \mu][x_i - \mu]^T.$$

We are told that $\mathbf{y}$ is a column vector of dimension d. So,

$$\mathbf{y}^T \cdot \text{Cov} \cdot \mathbf{y} = \frac{1}{n-1}\sum_{i=1}^n \mathbf{y}^T[x_i - \mu][x_i - \mu]^T\mathbf{y}.$$

Rewriting the above as an outer product,

$$\mathbf{y} \cdot [x_i - \mu][x_i - \mu] \cdot \mathbf{y} = \frac{1}{n-1}\sum_{i=1}^n \mathbf{y}[x_i - \mu][x_i - \mu] \cdot y = \frac{1}{n-1}\sum_{i=1}^n [\mathbf{y} \cdot (x_i - \mu)]^2.$$

Note (and using similar logic for the proof in 2b), the sum of values squared cannot be less than zero, i.e.

$$\frac{1}{n-1}\sum_{i=1}^{n} \text{Cov}(X)^2 \geq 0.$$

□

(b) Is Σ necessarily positive-definite?

Proof No. If you have an orthogonal y_i dotted with the x_i will be zero, without this being positive semi-definite. I.e. if there exists $\{x_i\}_{i=1}^{n}$ that are all orthogonal to y^T, then $x_i \cdot y^T$ is 0 but the vector was not zero.

□

4. (a) Intuitively (without proof), d^* will be small when there is a disproportionate level of eigenvectors located in a lower dimension relative to the more equitable distribution of eigenvectors that span the full space. That is, the data could be extremely correlated and lie on vectors that span the same dimensional space. This is opposed to eigenvectors that are more equitably distributed across D.

   (b) $d*$ would be large if the data was just noise, i.e. there is little to no correlation and few principal components capture a good amount of the data. d^* might need to be increased to capture more data.

   Similarly, if the structure of the data is flat or equally spread across dimensions, a dimension-reducing regime could fail to reduce the dimension, as dimension-reducing entails cutting off data with associated eigenvalue that would have roughly equal weight with the rest of the data.

   (c) d^* will be exactly $0.95 * D$ when the first $0.95 * D$ principal components of the data capture *exactly* 0.95 of the total data variance (or the remaining principal components capture less than 0.05 of the said variance).

5. The goal is to find the principal components of the given data, and determine if there is dimension-reduction involved.

   The code for producing the principal components is below. The following is just from the lecture code, and the only change is the type of data used.

```

clear all;

load(‘SalinasA.mat’);

% Display data
for k=1:3
    close all;
    imagesc(salinasA(:,:,k));
    title([‘Band Number ’,num2str(k)]);
    pause(1);
end

% Step 1: Center the data
Mu=mean(X);
X=X-Mu;

```

```
% Step 2: Build covariance matrix
Covar=1/size(X,1)*(X'*X);

% Step 3: Use MATLAB built-in PCA

[PC,EigVals]=eig(Covar);
EigVals=diag(EigVals);
[EigVals,IdxSorted]=sort(EigVals,'descend');
PC=PC(IdxSorted,:);

% Plot the first 10 eigenvalues
figure;
plot(EigVals(1:10));

%%

% How many dimensions do we need to preserve 95% of the variance?

TotalVariance=sum(EigVals);

for k=1:length(EigVals)
    TruncatedVariance(k)=sum(EigVals(1:k));
    ProportionVariance(k)=TruncatedVariance(k)/TotalVariance;
end

figure;
plot(ProportionVariance);
title('Proportion of Variance as a Function of Number of PC');

% Find the smallest k such that ProportionVariance(k)>.95
D_reduced=find(ProportionVariance>.95,1,'first');
display(D_reduced)
```

Listing 1: Code for Circle

Hence, just 2 dimensions are needed to preserve 95 percent of the variance of the data. The first three and last three principal components are: $\{0.6658, 0.0476, 0.0044\}, 1\times10^{-7}\{0.2602, 0.2483, 0.1942\}$.

This makes sense because the principal components are much larger with the first eigenvalues. This is due to the fact that 95of the variance is captured in such a small dimension, so naturally the eigenvalues for the later principal components would be very small.