§ Mathematical Note I

— Empirical Mean and Covariance —

Let $\vec{X}_1 = \{\vec{x}_i\}_{i=1}^{N}$ be a set of N realizations of the random variable $\vec{X}$, the empirical mean of $\vec{X}$ is:

$$(\mathrm{I}.1) \qquad \overline{\vec{x}} := \frac{1}{N} \sum_{i=1}^{N} \vec{x}_i$$

the arithmetic average of the observations.
The empirical covariance, the measurement of data spread around the empirical mean, is a D×D matrix given by:

$$(\mathrm{I}.2) \quad \Sigma := \frac{1}{N} \sum_{i=1}^{N} (\vec{x}_i - \overline{\vec{x}})(\vec{x}_i - \overline{\vec{x}})^T$$

Now, let $\vec{X}_1 = \{x_i\}_{i=1}^{N}$ and $\vec{Y}_1 = \{y_i\}_{i=1}^{N}$ be the realizations of random variables $\vec{X}$ and $\vec{Y}$, with empirical means:

$$(I.3) \qquad \bar{X} = \frac{1}{N} \sum_{i=1}^{N} \vec{x}_i$$

$$(I.4) \qquad \bar{Y} = \frac{1}{N} \sum_{i=1}^{N} \vec{y}_i$$

and empirical variances:

$$(I.5) \quad \Sigma_x = \frac{1}{N} \sum_{i=1}^{N} (\vec{X}_i - \bar{X})(\vec{X}_i - \bar{X})^T$$

$$(I.6) \quad \Sigma_y = \frac{1}{N} \sum_{i=1}^{N} (\vec{Y}_i - \bar{Y})(\vec{Y}_i - \bar{Y})^T$$

The relationship between these two variables is given by two measures:

+ Covariance

$$Cov(X,Y) = \frac{1}{N^2} \sum_{i=1} (\vec{x}_i - \bar{x})(\vec{Y}_i - \bar{Y})^T$$

+ Correlation

$$Cor(X,Y) \quad Cov(X,Y) \big/ \left((\Sigma_x)^{1/2} \cdot (\Sigma_Y)^{1/2} \right)$$

§ Mathematical Notes II

~ Singular Value Decomposition ~

Singular Value Decomposition (SVD) is the most fundamental theorem of linear algebra, since it can be applied to all matrices.

The SVD of a matrix $\hat{A}$ is a linear map $\Phi : V \to W$, that quantifies a change in the internal geometry of such vector spaces.

<u>SVD THEOREM</u>: Let $\hat{A}$ be a rectangular $m \times n$ matrix of rank $r \in [0, \min(m,n)]$. The SVD is a decomposition of the form:

$$\underset{m \times n}{\boxed{A}} = \underset{m \times m}{\boxed{U}}\;\underset{m \times n}{\boxed{\Sigma}}\;\underset{n \times n}{\boxed{V^T}} \tag{II.1}$$

with an orthogonal matrix $\hat{U} \in \mathbb{R}^{m \times m}$ with column vectors $\vec{u}_i$, $i = 1, \dots, m$, and an orthogonal matrix $\hat{V} \in \mathbb{R}^{n \times n}$ with column vectors $\vec{v}_j$, $j = 1, \dots n$. Moreover $\hat{\Sigma}$ is an $m \times n$ matrix with $\Sigma_{ii} = \sigma_i \geq 0$ and $\Sigma_{ij} = 0$, $i \neq j$.

The diagonal entries σ_i, $i = 1, \dots, M$ are called singular values. The u_i's are the left singular vector and the $\vec{v}_j$ are the right singular vectors.

The singular matrix $\hat{\Sigma}$, has its eigenvalues ordered from greater to smaller $\sigma_1 > \dots > \sigma_n$ and is unique. However, since $\hat{\Sigma}$ is of the same size of $\hat{A}$, it is a rectangular matrix, however it requires zero padding.

The SVD of a matrix is composed by finding two sets of orthonormal basis:
$U=(\vec{u}_1,\dots,\vec{u}_m)$ and $V=(\vec{v}_1,\dots,\vec{v}_n)$ of the codomains $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively. From these ordered basis, we construct the matrix $\hat{U}$ and $\hat{V}$.

We begin with the right-singular vector. Using the spectral theorem, every symmetric matrix has an orthonormal basis of eigenvectors and thus can be diagonalized. Also, we can always construct a symmetric and semipositive matrix from $\hat{A}$ by doing $\hat{A}\hat{A}^T \in \mathbb{R}^{n\times n}$, that can be always diagonalized:

$$(\mathrm{II}.2) \quad \hat{A}^{+}\hat{A} = \hat{P}\hat{D}\hat{P}^T = \hat{P}\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}\hat{P}^T$$

when $\hat{D}$ is an orthogonal matrix which columns are formed by the eigenvectors of $\hat{A}\hat{A}^T$.

Since $\hat{A}$ posses a SVD decomposition:

(II.3) $\hat{A} = \hat{U}\hat{\Sigma}\hat{V}^T$

One can write:

(II.4) $$\hat{A}^T\hat{A} = (\hat{U}\hat{\Sigma}\hat{V}^T)^T(\hat{U}\hat{\Sigma}\hat{V}^T)$$

$$= (\hat{V}\hat{\Sigma}^T\hat{U}^T)(\hat{U}\hat{\Sigma}\hat{V}^T)$$

$$= \hat{V}\hat{\Sigma}^T\underbrace{\hat{U}^T\hat{U}}_{=\mathbb{1}}\hat{\Sigma}\hat{V}^T$$

$$= \hat{V}\hat{\Sigma}^T\Sigma\hat{V}^T$$

Since $\Sigma = \mathrm{diag}(\sigma_1 \dots \sigma_N)$

$$= \hat{V}\begin{bmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_N^2 \end{bmatrix}\hat{V}^T$$

$$= \hat{P}\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{bmatrix}\hat{P}^T$$

this implies:

$$\hat{P} = \hat{V}$$

$$\lambda_i = \sigma_i^2$$

To obtain the left-singular vectors, we follow a similar procedure with $\hat{A}^T\hat{A}$.

Now, we need to connect the two steps. Using that the image of $\vec{v}_i$ under $\hat{A}$ has to be orthonormal to:

$$(\mathrm{II}.4)\quad \hat{A}\vec{v}_i = \hat{A}_{m\times n}\, \bar{v}_{i\,1\times n} = [\quad]_{m\times 1}$$

if $m>n$, then the image is a basis of $\mathbb{R}^m$, so:

$$(\mathrm{II}.5):\ \underbrace{\vec{u}_i}_{\text{Basis of } \mathbb{R}^m \text{ under } \hat{A}^T\hat{A}} = \frac{\hat{A}\vec{v}_i}{\|\hat{A}\vec{v}_i\|} = \frac{1}{\sqrt{\lambda_i}}\hat{A}\vec{v}_i = \frac{1}{\sigma_i}\hat{A}\vec{v}_i$$

resulting in the singular value equation.

$$(\mathrm{II}.6)\quad \hat{A}\vec{v}_i = \sigma_i\vec{u}_i \quad , i=1,\dots,n$$

in matrix form:

$$(\mathrm{II}.7)\quad \hat{A}\hat{V} = \hat{U}\hat{\Sigma}$$

$$\hat{A}\underbrace{\hat{V}\hat{V}^T}_{=\mathbb{1}} = \hat{U}\hat{\Sigma}\hat{V}^T$$

$$\hat{A} = \hat{U}\hat{\Sigma}\hat{V}^T \quad \blacksquare$$

§ Mathematical Notes III

~ Matrix Approximation ~

The SVD decomposition allow us to represent a matrix $\hat{A}$ as a sum of low-ranking matrices formed by the left and right singular vectors.

A rank-1 approximation is given by:

$$(\text{III}.1) \quad \hat{A}_i^{(1)} := \vec{u}_i \vec{v}_i^{\,T}$$

thus, the full rank $\hat{A}$ matrix is constructed using $\hat{A}_i^{(1)}$ as building blocks:

$$(\text{III}.2) \quad \hat{A} = \sum_{i=1}^{M} \sigma_i \vec{u}_i \vec{v}_i^{\,T} = \sum_{i=1}^{M} \sigma_i \hat{A}_i^{(1)}$$

This implies that we can approximate $\hat{A}$ truncating in a rank-$k < M$:

$$(\text{III}.3) \quad \hat{A}^{(k)} = \sum_{i=1}^{k} \sigma_i \hat{A}_i^{(1)}$$

In order to measure the error in this truncation we introduce the so called matrix-norm:

*Definition: (Spectral Norm of a Matrix):

For $\vec{x} \in \mathbb{R}^n / \{\vec{0}\}$, the spectral norm of a matrix is defined as:

$$(\text{III}.4) \qquad \|A\|_2 := \max_{\vec{x}} \frac{\|\hat{A}\vec{x}\|_2}{\|\vec{x}\|_2}$$

The spectral norm is a measure of how long any vector $\vec{x}$ can at most become when multiplied by $\hat{A}$.

Definition: The spectral norm of a matrix is it's longest singular value σ_1 /

With these concepts we introduce the second most important result of this lecture, the Eckart-Young theorem

Theorem (Eckart-Young Theorem):

Consider a matrix $\hat{A} \in \mathbb{R}^{m \times n}$ of rank r and let $\hat{B} \in \mathbb{R}^{m \times n}$ be a matrix of rank k. For any $k \leq r$ with $\hat{A}^{(k)} = \sum_{i=1}^{k} \sigma_i \vec{u}_i \vec{v}_i^T$, it holds that:

$$\hat{A}^{(k)} = \underset{r(B^{(k)})}{\operatorname{argmin}} \|\hat{A} - \hat{B}\|_2$$

$$\|\hat{A} - \hat{A}^{(k)}\|_2 = \sigma_{k+1}$$

The Eckart-Young Theorem states strictly how much error we introduce by approximating A by a low-rank approximation of $\hat{A}$. The SVD projection is a full rank approximation of the $\hat{A}$ matrix, and the PCA stands for a low-rank perturbative expansion of $\hat{A}$.