

# Lecture 07: Random Matrices I

Nikola Zlatanov

Innopolis University

Advanced Statistics

20-th of March to 27-th of March, 2023

# Motivation

- The vast majority of applications in data science, ML, and computer vision are based on data from random vectors and/or random matrices.
- An image mathematically is represented as a matrix.
- In fact, all images, digitally stored or processed by computers, are simply the storage or processing of matrices.
- Computers do not know that these are images. For them, these are matrices.
- Hence, we now need to develop statistics for vectors and matrices and then see how these statistics are applied in practice.

## Mean of Random Vectors

- Let  $\mathbf{X} = [X_1, X_2, \dots, X_d]^T$ , i.e.,  $\mathbf{X} \in \mathbb{R}^d$ , be a randomly generated  $d$ -dimensional vector according to some distribution  $f_{\mathbf{X}}(\mathbf{x})$ .
- How to compute the mean and variance of random vectors?
- The mean of  $\mathbf{X}$  is given by the following vector

$$E[\mathbf{X}] = \left[ E[X_1], E[X_2], \dots, E[X_d] \right]^T = [\mu_1, \mu_2, \dots, \mu_d]^T \quad (1)$$

where  $\mu_i = E[X_i]$ , for  $i = 1, 2, \dots, d$ .

- How about variance? Is it also going to be a vector?

# Variance of Random Vectors

- If  $d = 1$ , then  $\mathbf{X} = X_1$  and then we know how to compute the variance as

$$\begin{aligned}\text{VAR}[\mathbf{X}] \Big|_{d=1} &= E\left[(X_1 - E[X_1])^2\right] = E\left[(X_1 - E[X_1])(X_1 - E[X_1])\right] \\ &= E\left[X_1^2\right] - E[X_1]^2\end{aligned}\tag{2}$$

- But what do we do if  $d > 1$ ? How can we have a square of a vector to obtain something like  $E\left[(\mathbf{X} - E[\mathbf{X}])^2\right]$

# Covariance Matrix

- Well, for  $d > 1$ , the variance becomes a matrix that is called *the covariance matrix*, defined as

$$\begin{aligned}\text{COV}(\mathbf{X}) &= E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T] \\ &\stackrel{(a)}{=} E[\mathbf{X}\mathbf{X}^T] - E[\mathbf{X}]E[\mathbf{X}^T] \\ &= E[\mathbf{X}\mathbf{X}^T] - \boldsymbol{\mu}\boldsymbol{\mu}^T,\end{aligned}\tag{3}$$

where (a) you need to prove at home.

Hand-drawn diagram illustrating the dimensions of the covariance matrix. It shows a vertical vector of size  $d \times 1$  multiplied by a horizontal vector of size  $1 \times d$ , resulting in a square matrix of size  $d \times d$ .

# Covariance Matrix

- The  $(i, j)$ -th element of the matrix  $\text{COV}(\mathbf{X})$  is given by

$$\begin{aligned} [\text{COV}(\mathbf{X})]_{ij} &= E[(X_i - E[X_i])(X_j - E[X_j])] \\ &= E[X_i X_j] - E[X_i]E[\mathbf{X}_j], \end{aligned} \quad \text{X} \quad (4)$$

which is the definition of covariance between RVs  $X_i$  and  $X_j$

- The  $(i, i)$ -th element of  $\text{COV}(\mathbf{X})$ , i.e., the main diagonal elements of the matrix  $\text{COV}(\mathbf{X})$ , is given by

$$\begin{aligned} [\text{COV}(\mathbf{X})]_{ii} &= E[(X_i - E[X_i])^2] \\ &= E[X_i^2] - E[X_i]^2 = \text{VAR}(X_i) \end{aligned} \quad (5)$$

- $\text{COV}(\mathbf{X})$  is a symmetric matrix since  $[\text{COV}(\mathbf{X})]_{ij} = [\text{COV}(\mathbf{X})]_{ji}$  holds, which follows from (6) and (7).

# Covariance Matrix

- Hence, the entire matrix  $\text{COV}(\mathbf{X})$  looks like this:

$$\text{COV}(\mathbf{X}) = \begin{bmatrix} \text{VAR}[x_1] & \dots & \text{COV}(x_1, x_2) & \text{COV}(x_1, x_3) & \dots & \text{COV}(x_1, x_k) \\ \text{COV}(x_2, x_1) & \text{VAR}[x_2] & \dots & \dots & \dots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \mu_{ji} & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \mu_{ij} & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{COV}(x_k, x_1) & \dots & \dots & \dots & \dots & \text{VAR}[x_k] \end{bmatrix}$$

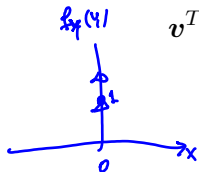
Diagram illustrating the structure of the Covariance Matrix  $\text{COV}(\mathbf{X})$ . The matrix is symmetric, with diagonal elements representing variances ( $\text{VAR}[x_i]$ ) and off-diagonal elements representing covariances ( $\text{COV}(x_i, x_j)$ ). The matrix is partitioned into blocks, with the top-left block labeled  $\text{COV}(\mathbf{X})$  and the bottom-right block labeled  $\text{VAR}(\mathbf{X})$ . The matrix is symmetric, with diagonal elements representing variances ( $\text{VAR}[x_i]$ ) and off-diagonal elements representing covariances ( $\text{COV}(x_i, x_j)$ ). The matrix is partitioned into blocks, with the top-left block labeled  $\text{COV}(\mathbf{X})$  and the bottom-right block labeled  $\text{VAR}(\mathbf{X})$ . The matrix is symmetric, with diagonal elements representing variances ( $\text{VAR}[x_i]$ ) and off-diagonal elements representing covariances ( $\text{COV}(x_i, x_j)$ ). The matrix is partitioned into blocks, with the top-left block labeled  $\text{COV}(\mathbf{X})$  and the bottom-right block labeled  $\text{VAR}(\mathbf{X})$ .

# Covariance Matrix

- A positive semi-definite matrix  $\mathbf{A}$  satisfies

$$\mathbf{v}^T \mathbf{A} \mathbf{v} \geq 0, \quad \forall \mathbf{v} \neq \mathbf{0} \quad (6)$$

- Fact: The matrix  $\text{COV}(\mathbf{X})$  is a positive semi-definite matrix.  
Proof:



$$\begin{aligned} \mathbf{v}^T \text{COV}(\mathbf{X}) \mathbf{v} &= \mathbf{v}^T E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T] \mathbf{v} \\ &= E[\mathbf{v}^T (\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T \mathbf{v}] \\ &\stackrel{(a)}{=} E[Y Y] = E[Y^2] \stackrel{(b)}{\geq} 0 \end{aligned} \quad (7)$$

where (a) comes from  $Y = \mathbf{v}^T (\mathbf{X} - E[\mathbf{X}]) = (\mathbf{X} - E[\mathbf{X}])^T \mathbf{v}$  and (b) holds since expectation of a squared RV is always non-negative



## Normal Random Vector

- Let  $\mathbf{Z} = [Z_1, Z_2, \dots, Z_d]$ , where  $Z_i \sim N(0, 1)$  are i.i.d.
- Then,  $E[\mathbf{Z}] = \mathbf{0} = [0, 0, \dots, 0]$  and  $\text{COV}(\mathbf{Z}) = \mathbf{I}$ , where  $\mathbf{I}$  is the identity matrix.
- In such case, we use the following notation  $\mathbf{Z} \sim N(\mathbf{0}, \mathbf{I})$ .
- The PDF of  $\mathbf{Z}$  is given by

$$\begin{aligned}
 f_{\mathbf{Z}}(\mathbf{z}) &= \prod_{i=1}^d f_{Z_i}(z_i) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right) \\
 &= \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\sum_{i=1}^d z_i^2}{2}\right) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{\|\mathbf{z}\|_2^2}{2}\right) \quad (8)
 \end{aligned}$$

- Note that  $f_{\mathbf{Z}}(\mathbf{z})$  depends only on the norm of  $\mathbf{z}$ . As a result, if I rotate  $\mathbf{z}$  by an orthogonal matrix  $\mathbf{U}$ , as  $\hat{\mathbf{z}} = \mathbf{U}\mathbf{z}$ , then  $f_{\hat{\mathbf{Z}}}(\hat{\mathbf{z}}) = f_{\mathbf{U}\mathbf{Z}}(\mathbf{U}\mathbf{z}) = f_{\mathbf{Z}}(\mathbf{z})$  since  $\|\mathbf{U}\mathbf{z}\|_2^2 = \|\mathbf{z}\|_2^2$

# Normal Random Vector

- Let  $\mathbf{X} = [X_1, X_2, \dots, X_d]$  be defined as

$$\mathbf{X} = \mathbf{A}\mathbf{Z} + \boldsymbol{\mu}, \quad (9)$$

where  $\mathbf{A}$  is a fixed matrix,  $\boldsymbol{\mu}$  is a fixed vector, and  $\mathbf{Z} \sim N(0, \mathbf{I})$ .  
Then,  $\mathbf{X}$  is general Gaussian random vector.

- The mean and covariance of  $\mathbf{X}$  are given by

$$E[\mathbf{X}] = \boldsymbol{\mu} \quad (10)$$

$$\begin{aligned} \text{COV}[\mathbf{X}] &= E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{X} - E[\mathbf{X}])^T] = E[\mathbf{A}\mathbf{Z}(\mathbf{A}\mathbf{Z})^T] \\ &= E[\mathbf{A}\mathbf{Z}\mathbf{Z}^T\mathbf{A}^T] = \mathbf{A}E[\mathbf{Z}\mathbf{Z}^T]\mathbf{A}^T = \mathbf{A}\mathbf{I}\mathbf{A}^T \\ &= \mathbf{A}\mathbf{A}^T \end{aligned} \quad (11)$$

# Normal Random Vector

- Let us denote  $\text{COV}[\mathbf{X}]$  as

$$\Sigma = \text{COV}[\mathbf{X}] = \mathbf{A}\mathbf{A}^T \quad (12)$$

Then, the PDF of  $\mathbf{X}$ , also known as the general multivariate Gaussian distribution is given by

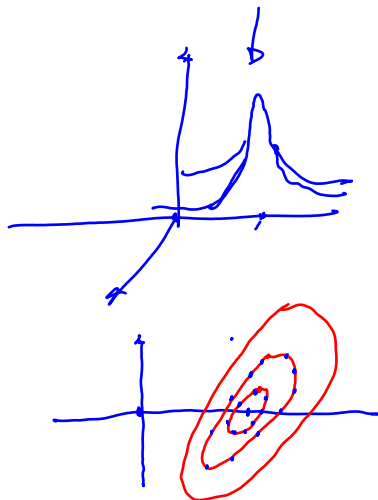
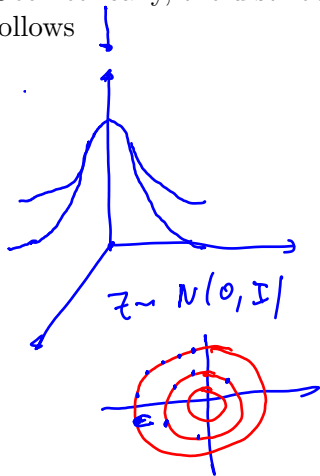
$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)} \exp \left( -\frac{(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}{2} \right), \quad (13)$$

where  $\Sigma^{-1}$  is the inverse of  $\Sigma$ .

- Note that the general multivariate Gaussian distribution of  $\mathbf{X}$ ,  $f_{\mathbf{X}}(\mathbf{x})$ , depends on two parameters the mean  $\boldsymbol{\mu}$  vector and the covariance matrix  $\Sigma$ .
- We denote the general Gaussian vector as  $\mathbf{X} \sim N(\boldsymbol{\mu}, \Sigma)$ .

# Geometrical Interpretation

Geometrically, the distributions of  $\mathbf{Z}$  and  $\mathbf{X}$  can be drawn as follows as follows

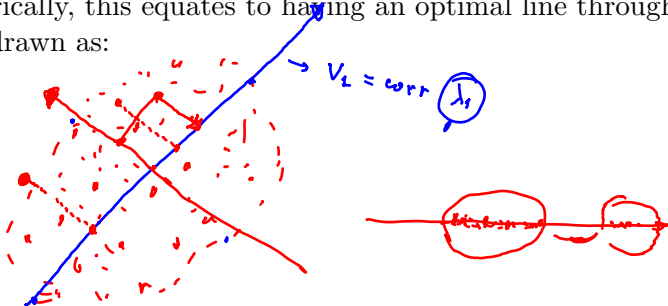


# Principal Component Analysis (PCA)

- Let  $\mathbf{X} \in \mathbb{R}^d$ , where  $d$  is very large, has a covariance matrix  $\Sigma$  and  $\mu = \mathbf{0}$ .
- We would like to reduce the dimension of  $\mathbf{X}$ ,  $d$ , such that we have very little loss of information.
- Ideally, we would like a mapping  $F : \mathbf{X} \rightarrow \mathbf{Y}$ , where  $\mathbf{Y} \in \mathbb{R}^k$  and  $k \ll d$ .
- Note that we did that in the previous lecture, using the Johnson–Lindenstrauss Lemma.
- Note that the Johnson–Lindenstrauss algorithm is blind (does not care) about the statistics of the data.
- Maybe, if we know the statistics of the data, say the covariance matrix,  $\Sigma$ , we will do a more accurate reduction of dimensions.
- Maybe, if we can reduce the dimension  $d$  to  $1 \leq k \leq 3$ , then we can even visualise the data. Note that we humans cannot visualise above 3 dimensions. This is where PCA is very useful, i.e., as a tool to visualise data.

# PCA

- Let us start with the extreme case: Let us reduce the dimension  $d$  to one.  $d \rightarrow k=1$
- Hence, we need a one dimensional RV  $Y$  that best “explains” the  $d$ -dimensional random vector  $X$ .
- Geometrically, this equates to having an optimal line through data points, drawn as:



# Principal Component Analysis (PCA)

- Mathematical, we need to find a unit vector  $\mathbf{v}$  which maximizes the variance of the projection of data points on  $\mathbf{v}$
- Mathematically, we need to maximize

$$\text{VAR}[\langle \mathbf{X}, \mathbf{v} \rangle],$$

where

$$\langle \mathbf{a}, \mathbf{b} \rangle = \sum_{i=1}^d a_i b_i$$

$\mathbf{a}^T, \mathbf{b}$

denotes the inner product of vectors  $\mathbf{a}$  and  $\mathbf{b}$ .

- On the other hand,

$$\begin{aligned} \text{VAR}[\langle \mathbf{X}, \mathbf{v} \rangle] &= E[\langle \mathbf{X}, \mathbf{v} \rangle^2] - \langle E[\mathbf{X}^T], E[\mathbf{v}] \rangle^2 = E[\langle \mathbf{X}, \mathbf{v} \rangle^2] \\ &= E[\mathbf{v}^T \mathbf{X} (\mathbf{v}^T \mathbf{X})^T] = E[\mathbf{v}^T \mathbf{X} \mathbf{X}^T \mathbf{v}] = \mathbf{v}^T E[\mathbf{X} \mathbf{X}^T] \mathbf{v} \\ &= \mathbf{v}^T \Sigma \mathbf{v} \end{aligned} \tag{14}$$

# Principal Component Analysis (PCA)

- Hence,

$$\max_{\mathbf{v}} \text{VAR}[\langle \mathbf{X}, \mathbf{v} \rangle] = \max_{\mathbf{v}} \mathbf{v}^T \mathbf{\Sigma} \mathbf{v} = \lambda_1 \quad (15)$$

where  $\lambda_1$  is the largest eigenvalue and the optimal  $\mathbf{v}$ , denoted by  $\mathbf{v}_1$  that achieves the maximization is the eigenvector corresponding to  $\lambda_1$ .



# Principal Component Analysis (PCA)

- Next, let us reduce the dimension  $d$  to two.
- Hence, we need a two dimensional vector  $\mathbf{Y} = [Y_1, Y_2]$  that best “explains” the  $d$ -dimensional random vector  $\mathbf{X}$ .
- Having maximized the variance in one direction, let’s find another direction, orthogonal to the first one, which maximizes the variance in that direction.
- This again equates to line through data points, drawn as:

# Principal Component Analysis (PCA)

- Mathematically, we need

$$\max_{\mathbf{v}, \mathbf{v} \perp \mathbf{v}_1} \text{VAR}[\langle \mathbf{X}, \mathbf{v} \rangle] = \max_{\mathbf{v}, \mathbf{v} \perp \mathbf{v}_1} \mathbf{v}^T \mathbf{\Sigma} \mathbf{v} = \lambda_2 \quad (16)$$

where  $\lambda_1$  is the second largest eigenvalue and the optimal  $\mathbf{v}$ , denoted by  $\mathbf{v}_2$  that achieves the maximization is the eigenvector corresponding to  $\lambda_2$ .

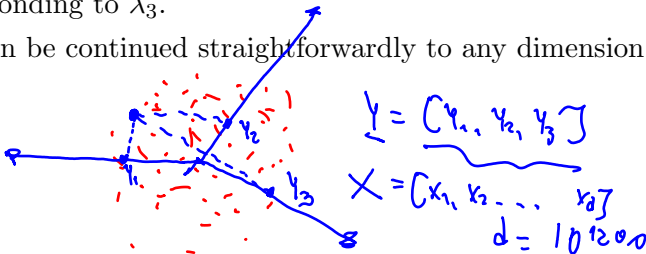
# Principal Component Analysis (PCA)

- If we continue to reducing  $d$  to three dimensions, we have mathematically

$$\max_{\mathbf{v}, \mathbf{v} \perp \mathbf{v}_1, \mathbf{v} \perp \mathbf{v}_2} \text{VAR}[\langle \mathbf{X}, \mathbf{v} \rangle] = \max_{\mathbf{v}, \mathbf{v} \perp \mathbf{v}_1, \mathbf{v} \perp \mathbf{v}_2} \mathbf{v}^T \mathbf{\Sigma} \mathbf{v} = \lambda_3 \quad (17)$$

where  $\lambda_3$  is the third largest eigenvalue and the optimal  $\mathbf{v}$ , denoted by  $\mathbf{v}_3$  that achieves the maximization is the eigenvector corresponding to  $\lambda_3$ .

- This can be continued straightforwardly to any dimension  $k \leq d$



# Principal Component Analysis (PCA)

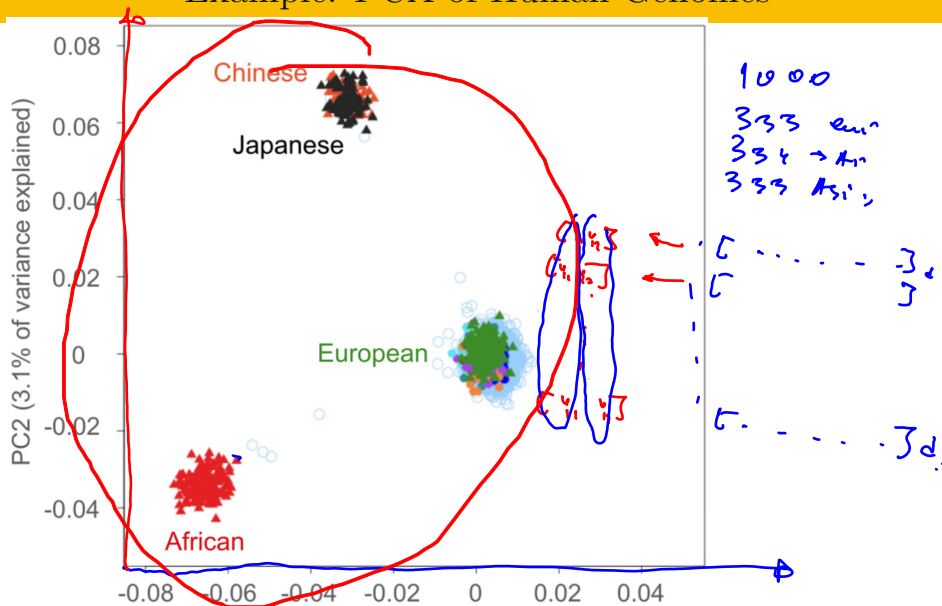
- The General PCA Algorithm: To reduce the dimension of  $\mathbf{X} \in \mathbb{R}^d$  from  $d$  to  $k$ , and thereby obtain  $\mathbf{Y} \in \mathbb{R}^k$ , we project  $\mathbf{X}$  into the subspace defined by the  $k$  eigenvectors of the covariance matrix  $\mathbf{\Sigma}$  that correspond to the  $k$ -th largest eigenvalues of  $\mathbf{\Sigma}$ .
- PCA is usefull since it maximizes the variability of  $\mathbf{X} \in \mathbb{R}^d$  in  $k \leq d$  dimensional space.

# Principal Component Analysis (PCA)

- What PCA tells us about the covariance matrix  $\Sigma$  is that the spectrum of the matrix, i.e., the distribution of its eigenvalues, is a hidden information that resolves the information in  $\mathbf{X}$  into sub-components expressed by the eigenvalues of the covariance matrix  $\Sigma$ .
- One very important fact is that to do PCA, using the shown algorithm, the covariance matrix,  $\Sigma$ , has to be known.

$$\underline{X} = \begin{bmatrix} \lambda_1 \\ \downarrow \\ \underline{v_1} \end{bmatrix}, \begin{bmatrix} \lambda_2 \\ \downarrow \\ \underline{v_2} \end{bmatrix}, \dots, \begin{bmatrix} \lambda_n \\ \downarrow \\ \underline{v_n} \end{bmatrix}$$

# Example: PCA of Human Genomes



# Principal Component Analysis (PCA)

- But, how accurate is the PCA, when the covariance matrix  $\Sigma$  is unknown but needs to be estimated from the data itself?
- How close should the estimated covariance matrix  $\hat{\Sigma}$  and the actual covariance matrix  $\Sigma$  need to be in order for PCA based on  $\hat{\Sigma}$  to be an approximate PCA on  $\Sigma$ ?
- Maybe, we can never have enough samples to estimate  $\Sigma$  accurately enough such that PCA based on  $\hat{\Sigma}$  to be an approximate PCA on  $\Sigma$ ?
- Maybe the genome image we showed does not have any connection to reality due to the low number of samples from which the covariance matrix was estimated?

Handwritten notes:

$$d \rightarrow \mathcal{O}(d) \rightarrow 10^6$$

$$S = \mathcal{O}(10^6)$$

$$S = \rightarrow \mathcal{O}(10^6)$$

PCA