# Review

Ngoc Hoang Luong

University of Information Technology (UIT), VNU-HCM

March 27, 2023

# References

The contents of the slides are from: Gaston Sanchez and Ethan Marzban: *All Models Are Wrong: Concepts of Statistical Learning* - `https://allmodelsarewrong.github.io/duality.html`

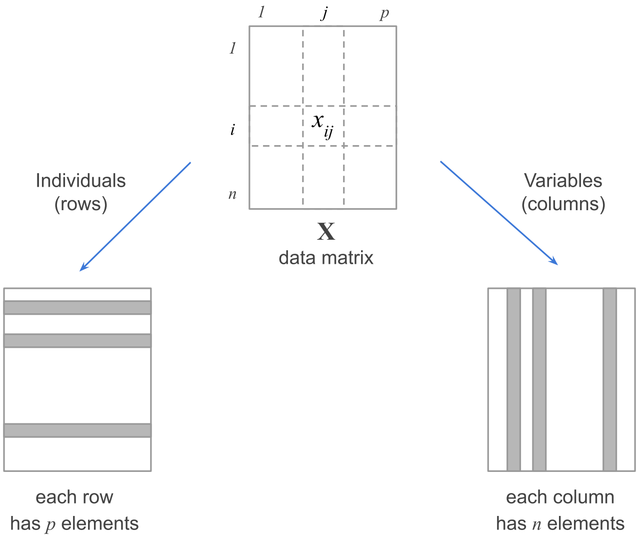
# Basic Notations

- Assume our data to be in a tabular format, which can be represented as a mathematical **matrix** object.
- An example of a data matrix $\boldsymbol{X}$ of size $n \times p$:

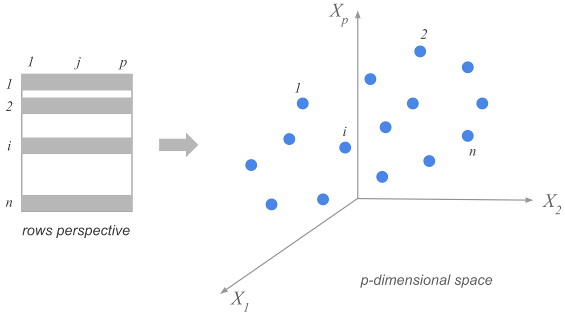
$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix}$$

- We assume the rows of a data matrix correspond to the data items/individuals/objects.
- We assume the columns of a data matrix correspond to the variables/features observed on the individuals.
- $x_{ij}$ represents the value observed for the $j$-th variable on the $i$-th individual.

# Duality of a Data Matrix

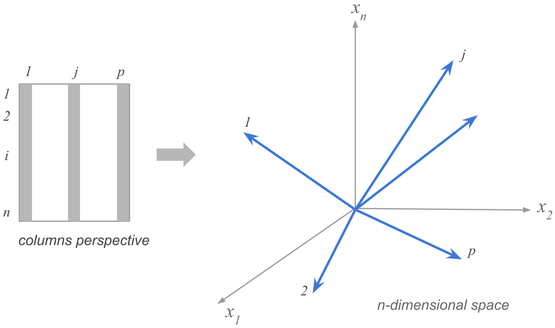


# Duality of a Data Matrix - Row Space



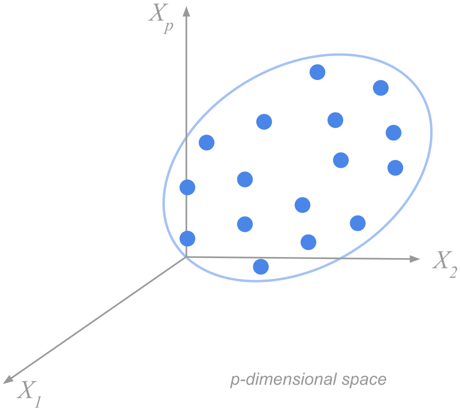

- Each row of the data matrix has $p$ elements. We can regard each row as a single point in a $p$-dimensional space (with $p$ axes).
- All together they form a cloud of points.

# Duality of a Data Matrix - Column Space




- Because each variable has $n$ elements, we can regard the set of $p$ variables as objects in an $n$-dimensional space (with $n$ axes).
- Each variable is represented as an arrow (or a vector). In data pre-processing, we apply transformations on variables, that can change their scales (shrinking or stretching) without modifying their directions.

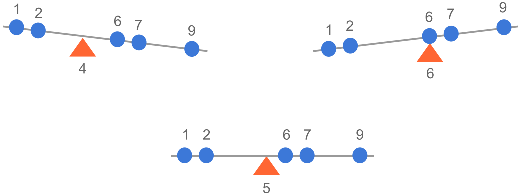
# Duality of a Data Matrix - Cloud of Individuals




- The rows of the data matrix correspond to $n$ individuals/points in a $p$-dimensional space.
- We consider common operations we can apply on the individuals.

# Cloud of Individuals - Average Individual

- If we have only one variable, then all individual points lie in a one-dimensional space, which is a **line**.
- The **average individual** can be defined as the arithmetic average of the values, corresponding to the balancing point.

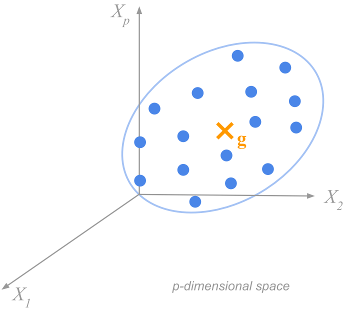



- The average of individuals $x_1, x_2, \ldots, x_n$ is:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\mathbf{x}^\top \mathbf{1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{1} = (1, 1, \ldots, 1)$.

# Cloud of Individuals - Average Individual

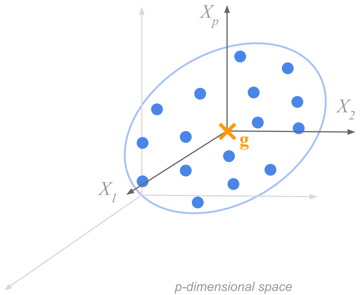



- If we have multiple variables (multivariate), the average individual is the point $\boldsymbol{g}$ with co-ordinates as the averages of all the variables:

$$\boldsymbol{g} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_p)$$

  where $\bar{x}_j$ is the average of the $j$-th variable.
- $\boldsymbol{g}$ is also called the **centroid**, barycenter, or center of gravity of the cloud of points.

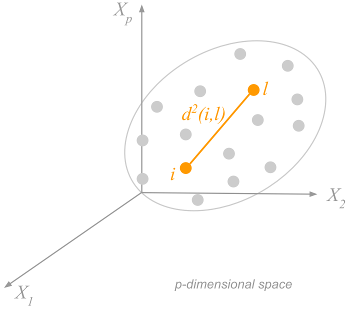
# Cloud of Individuals - Centered Data




- It is convenient to make the centroid of a data set become the origin of the cloud of points.
- Geometrically, this transformation is a shift of the axes in the $p$-dimensional space.
- Algebraically, this transformation is expressing the value of each variable in terms of the deviations from their averages (means).

$$\mathbf{x}_1 - \mathbf{g}, \mathbf{x}_2 - \mathbf{g}, \ldots, \mathbf{x}_n - \mathbf{g}$$

# Cloud of Individuals - Distance between Individuals




- The most common type of distance is the (squared) Euclidean distance.
- With $p$ variables, the squared distance between the $i$-th individual and the $l$-th individual is:

$$\begin{aligned} d^2(i,l) &= (x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \ldots + (x_{ip} - x_{lp})^2 \\ &= (\boldsymbol{x}_i - \boldsymbol{x}_l)^\top (\boldsymbol{x}_i - \boldsymbol{x}_l) \end{aligned}$$

# Cloud of Individuals - Overall Dispersion

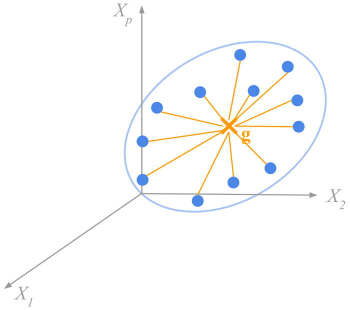
- The centroid is a measure of center of individuals. The dispersion is a measure of spread/scatter among individuals.
- If we have three individuals, we can compute all pairwise distances and sum them up:

$$
\begin{aligned}
d^2(a,a) + d^2(b,b) + d^2(c,c) +& \\
d^2(a,b) + d^2(b,a) +& \\
d^2(a,c) + d^2(c,a) +& \\
d^2(b,c) + d^2(c,b)&
\end{aligned}
$$

- The **overall dispersion** of $n$ individuals is:

$$\sum_{i=1}^{n}\sum_{l=1}^{n} d^2(i,l)$$

## Cloud of Individuals - Inertia




- The **inertia** can be computed by averaging the squared distances between all individuals and the centroid:

$$\frac{1}{n}\sum_{i=1}^{n} d^2(i,g) = \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{g})^\top (\boldsymbol{x}_i - \boldsymbol{g})$$

- In uni-dimensional case, $p = 1$, we have: $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

# Cloud of Variables - Mean of a Variable

- The mean (or average) of an $n$-element variable $\bar{x}$ is computed by:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$
$$= \frac{1}{n}x_1 + \frac{1}{n}x_2 + \ldots + \frac{1}{n}x_n$$

- We can generalize the concept of an average as a **weighted aggregation of information**.
- If we denote the weight of the $i$-th individual as $w_i$, then the average is:

$$\bar{x} = w_1x_1 + w_2x_2 + \ldots + w_nx_n$$
$$= \sum_{i=1}^{n} w_ix_i$$
$$= \mathbf{w}^\top \mathbf{x}$$

# Cloud of Variables - Variance of a Variable

- The variance is a measure of spread around the mean. We can take the average of the squared deviations from the mean.

$$Var(X) = \frac{(x_1 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Because the variance has squared unit, we need to take the square root to recover the original unit in which $X$ is expressed. This is the standard deviation.

$$sd(X) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

# Cloud of Variables - Variance of a Variable - Vector Notation

- A variable $X = (x_1, x_2, \ldots, x_n)$ can be denoted as a vector $\mathbf{x}$. The variance of a vector $\mathbf{x}$ can be computed:

$$Var(\mathbf{x}) = \frac{1}{n}(\mathbf{x} - \bar{\mathbf{x}})^\top(\mathbf{x} - \bar{\mathbf{x}})$$

where $\bar{\mathbf{x}}$ is an $n$-element vector of mean values $\bar{x}$.

- If $\mathbf{x}$ is already mean-centered, then

$$Var(\mathbf{x}) = \frac{1}{n}\mathbf{x}^\top\mathbf{x} = \frac{1}{n}\|\mathbf{x}\|^2$$

# Cloud of Variables - Covariance

- The covariance generalizes the concept of variance for two variables.

$$cov(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

where $\bar{x}$ is the mean value of $\mathbf{x}$:

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and $\bar{y}$ is the mean value of $\mathbf{y}$:

$$\bar{y} = \frac{1}{n}(y_1 + y_2 + \ldots + y_n) = \frac{1}{n} \sum_{i=1}^{n} y_i$$

- If the variables are mean-centered, we have

$$cov(\mathbf{x}, \mathbf{y}) = \frac{1}{n}(\mathbf{x}^\top \mathbf{y})$$

# Cloud of Variables - Correlation

- Covariance indicates the direction (positive or negative) of a possible linear relation, but it does not tell how big the relation is.
- Use the standard deviations of the variables to normalize the covariance.
- The coefficient of linear correlation is defined as:

$$cor(\mathbf{x}, \mathbf{y}) = \frac{cov(\mathbf{x}, \mathbf{y})}{\sqrt{var(\mathbf{x})}\sqrt{var(\mathbf{y})}}$$

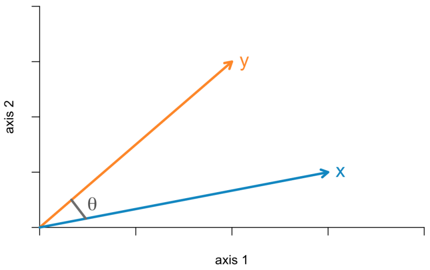
- If the variables are mean-centered, we have:

$$cor(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

- If both $\mathbf{x}$ and $\mathbf{y}$ are standardized, the correlation is:

$$cor(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$$

# Geometry of Correlation



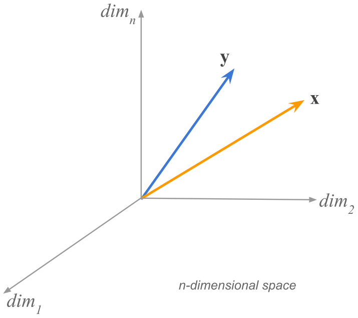

- The inner product of two mean-centered vectors:

$$\mathbf{x}^\top \mathbf{y} = \|\mathbf{x}\| \|\mathbf{y}\| \cos\left(\theta_{\mathbf{x},\mathbf{y}}\right)$$

- The correlation between mean-centered vectors $\mathbf{x}$ and $\mathbf{y}$ is the cosine of the angle between $\mathbf{x}$ and $\mathbf{y}$:

$$\cos\left(\theta_{\mathbf{x},\mathbf{y}}\right) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = cor(\mathbf{x}, \mathbf{y})$$

# Cloud of Variables - Orthogonal Projections

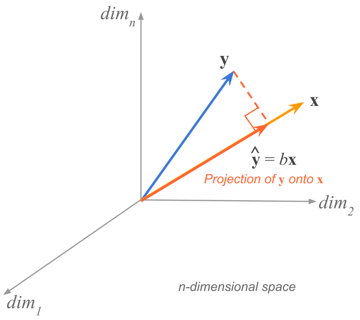



- Consider 2 variables $\mathbf{x}$ and $\mathbf{y}$. Can we approximate $\mathbf{y}$ in terms of $\mathbf{x}$?
- The approximation of $\mathbf{y}$, denoted by $\hat{\mathbf{y}}$, means finding a scalar $b$:

$$\hat{\mathbf{y}} = b\mathbf{x}$$

- To get $\hat{\mathbf{y}}$, we minimize the squared difference between $\mathbf{y}$ and $\hat{\mathbf{y}}$.

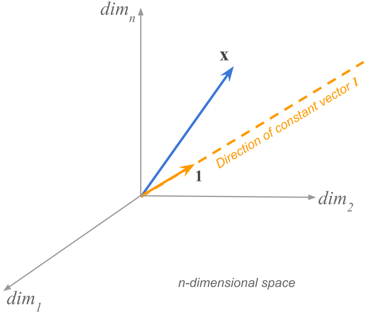
# Cloud of Variables - Orthogonal Projection




- We project $\mathbf{y}$ orthogonally onto $\mathbf{x}$:

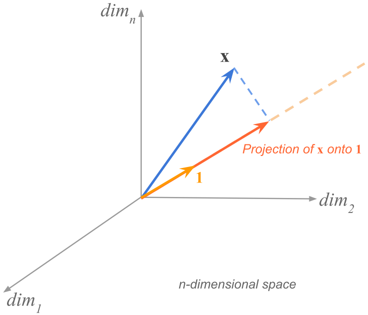
$$\hat{\mathbf{y}} = b\mathbf{x} = \mathbf{x}\left(\frac{\mathbf{y}^\top \mathbf{x}}{\mathbf{x}^\top \mathbf{x}}\right) = \mathbf{x}\left(\frac{\mathbf{y}^\top \mathbf{x}}{\|\mathbf{x}\|^2}\right)$$
$$= \mathbf{x}(\mathbf{x}^\top \mathbf{x})^{-1}\mathbf{x}^\top \mathbf{y}$$

# Cloud of Variables - The Mean as an Orthogonal Projection




- A variable $X = (x_1, x_2, \ldots, x_n)$ can be denoted as a vector $\mathbf{x}$ in an $n$-dimensional space.
- Consider the constant vector $\mathbf{1} = (1, 1, \ldots, 1)$.
- What is the orthogonal projection of $\mathbf{x}$ onto $\mathbf{1}$?

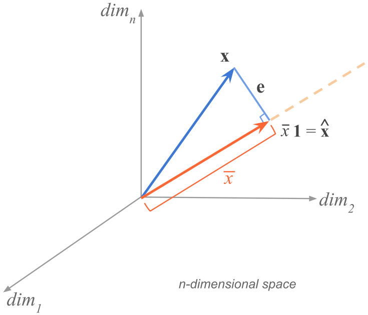
# Cloud of Variables - The Mean as an Orthogonal Projection




- The projection is:

$$
\begin{aligned}
\hat{\mathbf{x}} = b\mathbf{1} = \mathbf{1}\left(\frac{\mathbf{x}^\top\mathbf{1}}{\mathbf{1}^\top\mathbf{1}}\right) &= \mathbf{1}\left(\frac{\mathbf{x}^\top\mathbf{1}}{\|\mathbf{1}\|^2}\right) \\
&= \left(\frac{x_1 \cdot 1 + x_2 \cdot 1 + \ldots + x_n \cdot 1}{1 \cdot 1 + 1 \cdot 1 + \ldots + 1 \cdot 1}\right)\mathbf{1} = \frac{x_1 + x_2 + \ldots + x_n}{n}\mathbf{1} \\
&= \bar{x}\mathbf{1}
\end{aligned}
$$

# Cloud of Variables - The Mean as an Orthogonal Projection




- The mean of the variable $X$, denoted by $\bar{x}$, is the scalar we multiply with $\mathbf{1}$ to obtain the vector projection $\hat{\mathbf{x}}$.