
Dimensionality Reduction

Mean, Variance and Standard Deviation

- Consider one feature/attribute x .
- Suppose that we have n examples of patterns that all belong to the same class.
- Let the different values for the feature x be $x^{(1)}, x^{(2)}, \dots, x^{(n)}$
- There are two important statistics that we can use to characterize this collection of examples- the mean \bar{X} (or, μ) and the variance σ^2 .
- The mean is the arithmetic average or the center of mass:

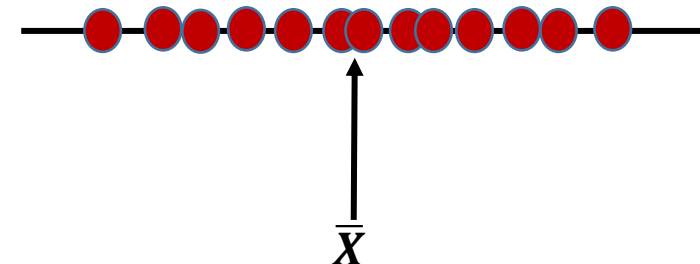
$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

$X =$

Attribute
4
6
2

$\bar{X} =$

4



Mean, Variance and Standard Deviation

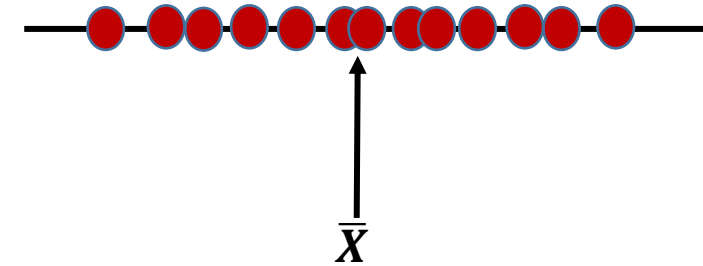
- The mean is the arithmetic average or the center of mass:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

- In general, if the data fall in one cluster/group, we expect the mean to be more or less in the center of that cluster. That is, the mean represents a typical value.
- The variance is a measure of the size of the cluster -- how much departure there is from the typical value.

- It is defined as the arithmetic average of the square of the deviations from the mean.

$$\sigma^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$



X =

Attribute 1
4
6
2

$\sigma^2 =$

$$\frac{(4 - 4)^2 + (6 - 4)^2 + (2 - 4)^2}{3}$$

Mean, Variance and Standard Deviation

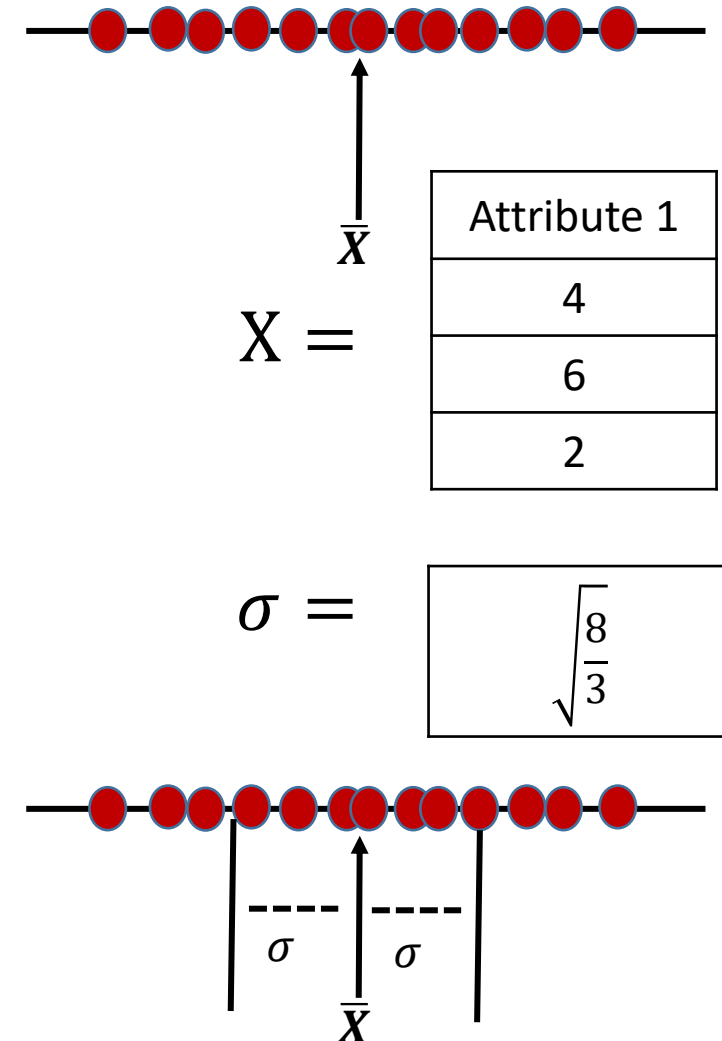
- The variance is a measure of the size of the cluster -- how much departure there is from the typical value.

- It is defined as the arithmetic average of the square of the deviations from the mean.

$$\sigma^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

- The square root of the *variance* is the *standard deviation* σ

- Where the mean measures the location of the center of the cluster, the standard deviation measures its "radius".



Covariance

- Variance – measure of the deviation from the mean for points in one dimension, e.g., heights
- Covariance – a measure of how much each of the dimensions varies from the mean with respect to each other
- Covariance is measured between 2 dimensions to see if there is a relationship between the 2 dimensions, e.g., number of hours studied and grade obtained.
- The covariance between one dimension and itself is the variance

X	Y
10	3
5	2
3	1

Mean	6	2
------	---	---

$$Var(X) = \frac{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}{n}$$

$$cov(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n}$$

$$cov(X, Y) = \frac{(10 - 6)(3 - 2) + (5 - 6)(2 - 2) + (3 - 6)(1 - 2)}{3}$$
$$cov(X, Y) = \frac{4 + 0 + 3}{3} = \frac{7}{3}$$

Covariance

- What is the interpretation of covariance calculations?
- Say you have a 2-dimensional data set
 - **X: number of hours studied for a subject**
 - **Y: marks obtained in that subject**
- And assume the covariance value (between X and Y) is:
104.53
- What does this value mean?

Covariance

- Exact value is not as important as its sign.
- A positive value of covariance indicates that **both dimensions increase or decrease together**, e.g., as the number of hours studied increases, the grades in that subject also increase.
- A negative value indicates while **one increases the other decreases**, or vice-versa, e.g., active social life vs. performance in CS Dept.
- If covariance is zero: the two dimensions are **independent of each other**, e.g., heights of students vs. grades obtained in a subject.
- Covariance calculations are used to find relationships between dimensions in high dimensional data sets (usually greater than 3) where visualization is difficult.

Covariance Matrix

- Representing covariance among dimensions as a matrix, e.g., for 3 dimensions:

$$C = \begin{bmatrix} \text{cov}(X, X) & \text{cov}(X, Y) & \text{cov}(X, Z) \\ \text{cov}(Y, X) & \text{cov}(Y, Y) & \text{cov}(Y, Z) \\ \text{cov}(Z, X) & \text{cov}(Z, Y) & \text{cov}(Z, Z) \end{bmatrix}$$

- **Properties:**
 - **Diagonal: variances of the variables**
 - $\text{cov}(X, Y) = \text{cov}(Y, X)$ hence matrix is symmetrical about the diagonal
 - **m-dimensional data will result in $m \times m$ covariance matrix**

Data Dimensionality

- From a theoretical point of view, increasing the number of features should lead to better performance.
- In practice, the inclusion of more features leads to worse performance (i.e., **curse of dimensionality**).
- The number of training examples required increases **exponentially** with dimensionality.

The curse of dimensionality

- Real data usually have **thousands, or millions** of dimensions
 - E.g., web documents, where the dimensionality is the vocabulary of words
 - Facebook graph, where the dimensionality is the number of users
- Huge number of dimensions causes problems
 - Data becomes very **sparse**, some algorithms become meaningless
 - The **complexity** of several algorithms depends on the dimensionality and they become infeasible.

Dimensionality Reduction

- Usually the data can be described with fewer dimensions, without losing much of the meaning of the data.
 - The data reside in a space of lower dimensionality
- Essentially, we assume that some of the data is noise, and we can approximate the useful part with a lower dimensionality space.
 - Dimensionality reduction does not just reduce the amount of data, it often brings out the useful part of the data

Dimensionality Reduction

- Significant improvements can be achieved by first mapping the data into a lower-dimensional space.

$$x = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \dots \\ a_p \end{bmatrix} \rightarrow \text{reduce dimensionality} \rightarrow y = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \dots \\ b_k \end{bmatrix}$$

$(k \ll p)$

- Dimensionality can be reduced by
 - Combining features using a linear or non-linear transformations.
 - Selecting a subset of features (i.e., feature selection).

Dimensionality Reduction

- **Linear** combinations are particularly attractive because they are simple to compute and analytically tractable.
- Given $x \in \mathbb{R}^p$, the goal is to find an $p \times k$ matrix **U** such that:

$$y = U^T x \in \mathbb{R}^k \text{ where } k \ll p$$

$$x = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \dots \\ a_p \end{bmatrix} \rightarrow \text{reduce dimensionality} \rightarrow y = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \dots \\ b_k \end{bmatrix} \quad (k \ll p)$$

Example of a problem

- We collected p parameters about 100 students:
 - Height
 - Weight
 - Hair color
 - Average grade
 - ...
- We want to find the most important parameters that best describe a student.

Example of a problem

- Each student is described using a vector length p :
 - (180, 70, 'purple', 84,...)
- We have $n = 100$ such vectors. Let's put them in one matrix, where each column is one student vector.
- So we have a $p \times n$ matrix. This will be the input of our problem.

Which parameters can we ignore?

- Constant parameter (number of heads)
 - 1,1,...,1.
- Constant parameter with some noise - (thickness of hair)
 - 0.003, 0.005, 0.002, ..., 0.0008 → low variance
- Parameter that is linearly dependent on other parameters (head size and height)
 - $Z = aX + bY$

Which parameters do we want to keep?

- Parameter that doesn't depend on others (e.g. eye color), i.e. uncorrelated.
- Parameter that changes a lot (grades)
 - The opposite of noise
 - High variance
- Questions:
 - How we describe 'most important' features using math?
 - Variance
 - How do we represent our data so that the most important features can be extracted easily?
 - Change of basis

Change of Basis !!!

- Let X and Y are matrices related by a linear transformation P .
- X is the original recorded data set and Y is a re-representation of that data set.

$$P^T X = Y$$

- Let's define;
 - p_i are the row of P .
 - x_i are the columns of X .
 - y_i are the columns of Y .

Change of Basis !!!

- Let X and Y are matrices related by a linear transformation P .
- X is the original recorded data set and Y is a re-representation of that data set.

$$P^T X = Y$$

- What does this mean?
 - P is a matrix that transforms X into Y .
 - Geometrically, P is a rotation and a stretch (scaling) which transforms X into Y .
 - The rows of P , $\{p_1, p_2, \dots, p_m\}$ are a set of new basis vectors for expressing the columns of X .

Change of Basis !!!

- Changing the basis doesn't change the data – only its representation.
- Changing the basis is actually projecting the data vectors on the basis vectors.
- Geometrically, P is a rotation and a stretch of X .
 - If P basis is orthonormal (length = 1) then the transformation P is only a rotation.

What does “best express” the data mean ?!!!

- As we've said before, we want to filter out noise and extract the relevant information from the given data set.
- Hence, the representation we are looking for will decrease both noise and redundancy in the data set at hand.

Principal Component Analysis

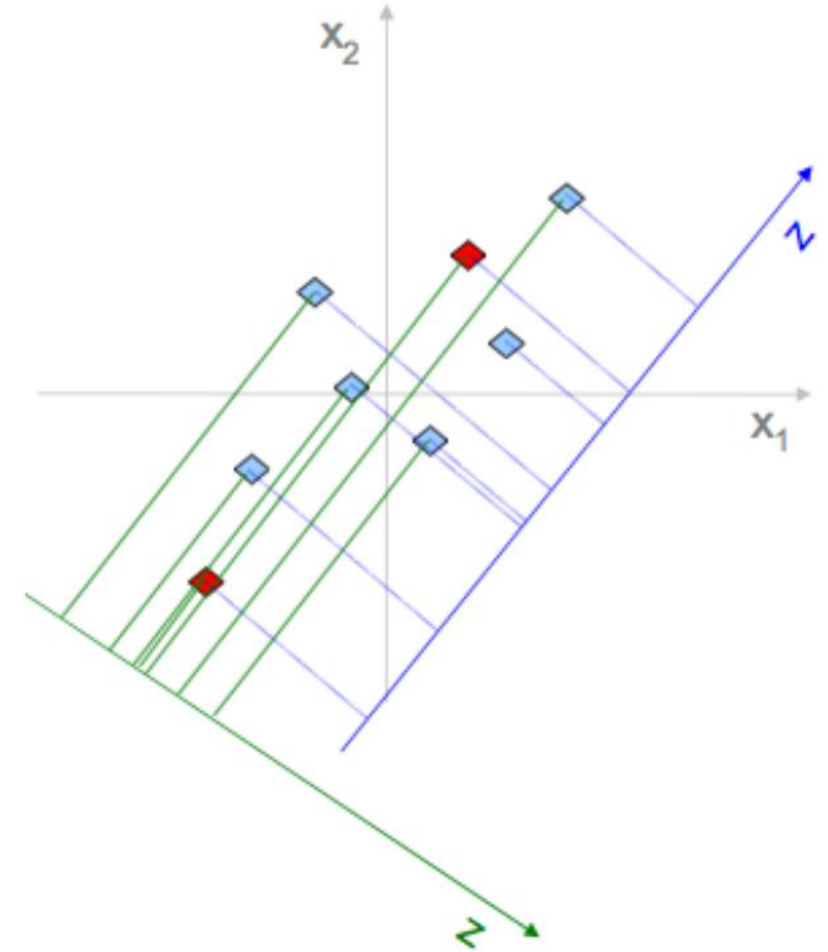
Maximum Variance Projection Method

Principal component analysis

- **Principal component analysis (PCA) is a way to reduce data dimensionality by finding a new set of variables, smaller than the original set of variables, that nonetheless retains most of the sample's information.**
- **The new variables, called principal components (PCs), are uncorrelated, and are ordered by the fraction of the total information each retains.**
- **PCA projects the data in the least square sense— it captures big (principal) variability in the data and ignores small variability.**

Why Variance

- Example: reduce 2-dimensional data to 1d
 - $\{x_1, x_2\} \rightarrow e$ (along new axis)
- Pick E to maximize variability
- Reduces cases when two points are close in e -space but very far in (x, y) -space
- Minimizes distances between original points and their projections



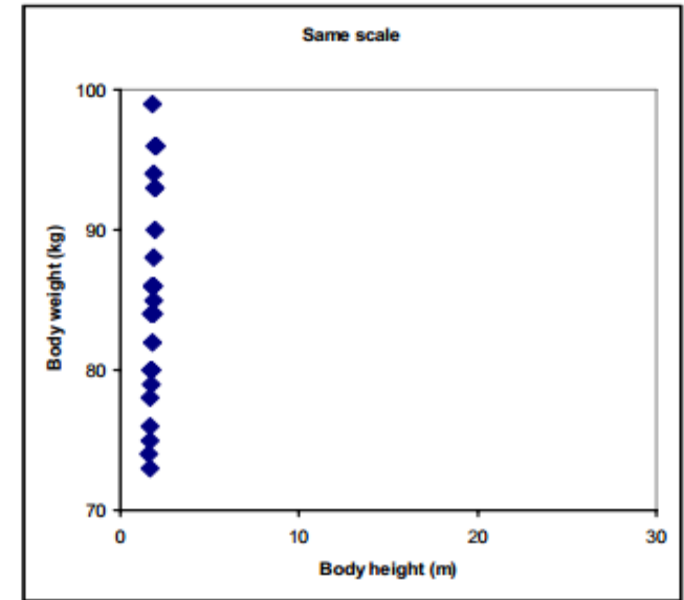
Pre-Processing of Data

- **Scaling**

- Variables often have substantially different numerical range
- A variable with large range has a large variance, whereas a variable with a small range has a small variance
- Since PCA is a maximum variance projection method, it follows that a variable with a large variance is more likely to be expressed in modeling than a low-variance variable

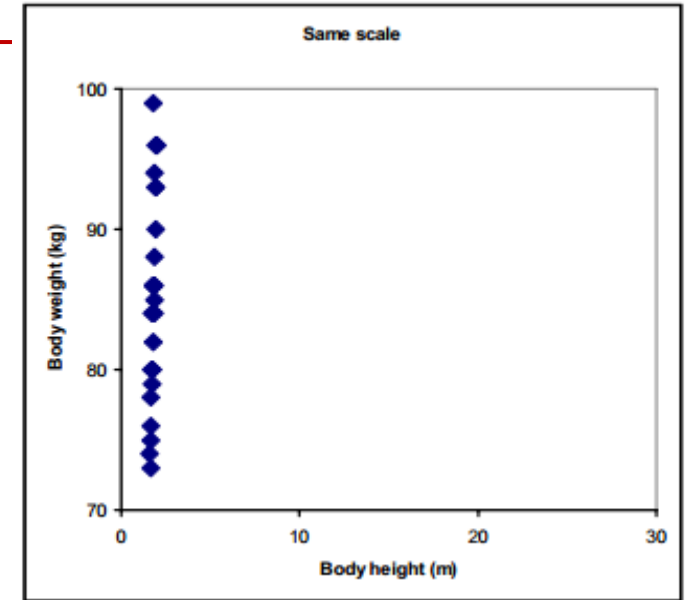
- **Example**

Height (m)	1.8	1.61	1.68	1.75	1.74	1.67	1.72	1.98	1.92	1.7	1.77	1.92
Weight (kg)	86	74	73	84	79	78	80	96	90	80	86	93
Height (m)	1.6	1.85	1.87	1.94	1.89	1.89	1.86	1.78	1.75	1.8	1.68	
Weight (kg)	75	84	85	96	94	86	88	99	80	82	76	



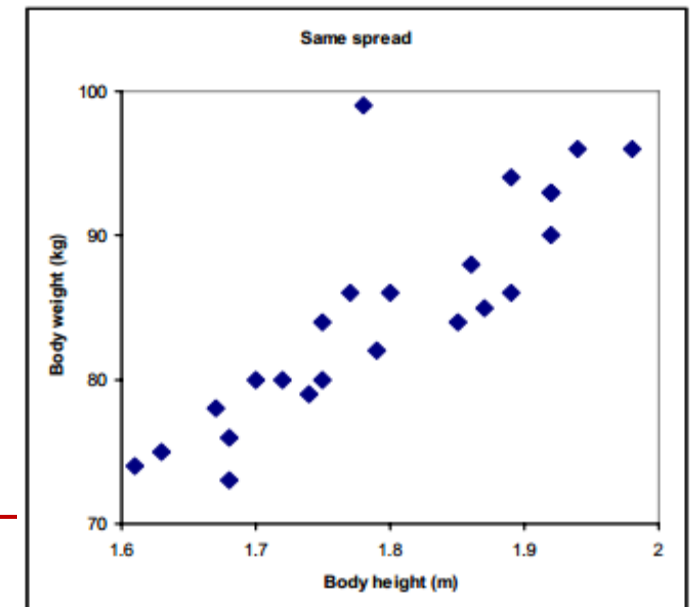
Pre-Processing of Data - Scaling

- We can see that the data points only spread in the vertical direction because body weight has much larger numerical range than body height.



- Lets zoom the Figure

- There is strong correlation between body height and body weight, except for one outlier in the data.



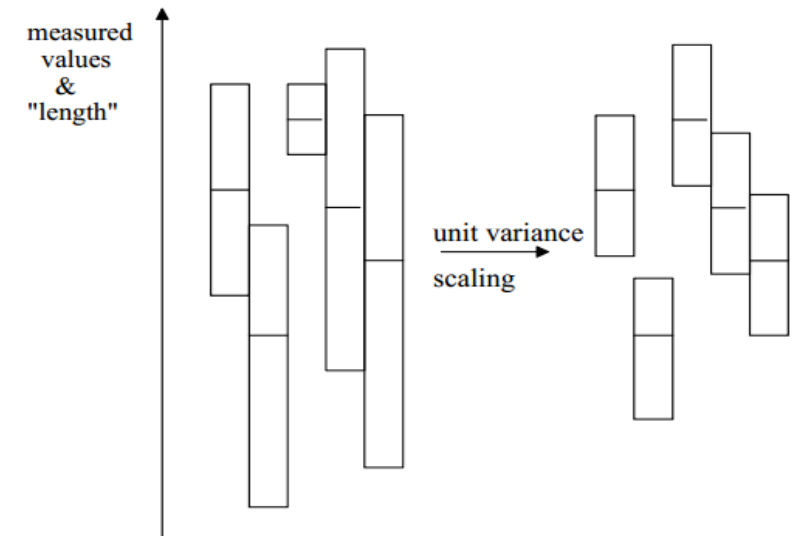
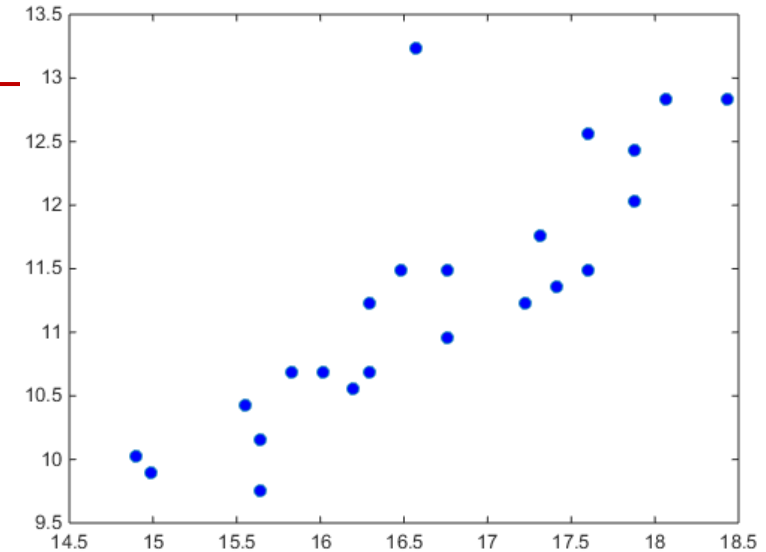
Pre-Processing of Data - Scaling

- **Solution:**

- **Scaling** : In order to give both variable, body weight and height, equal weight in the data, we standardized (scaling or weighting) them.
- There are many ways, but the most common technique is *unit variance*.

- **Unit variance**

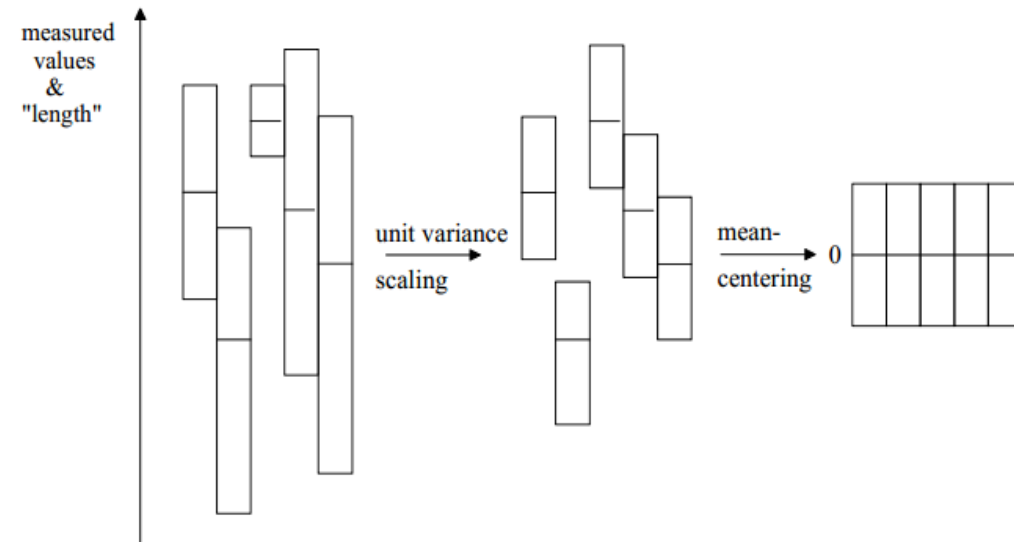
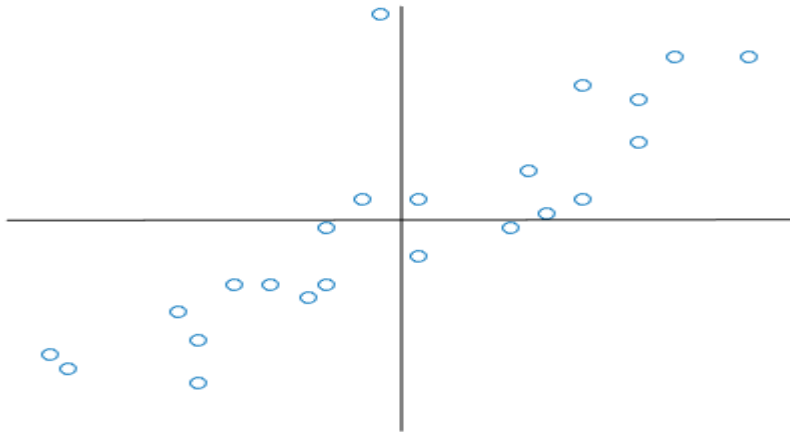
- For each variable, calculate the standard deviation (s_k)
- Scaling weight = inverse of standard deviation $\left(\frac{1}{s_k}\right)$



Pre-Processing of Data

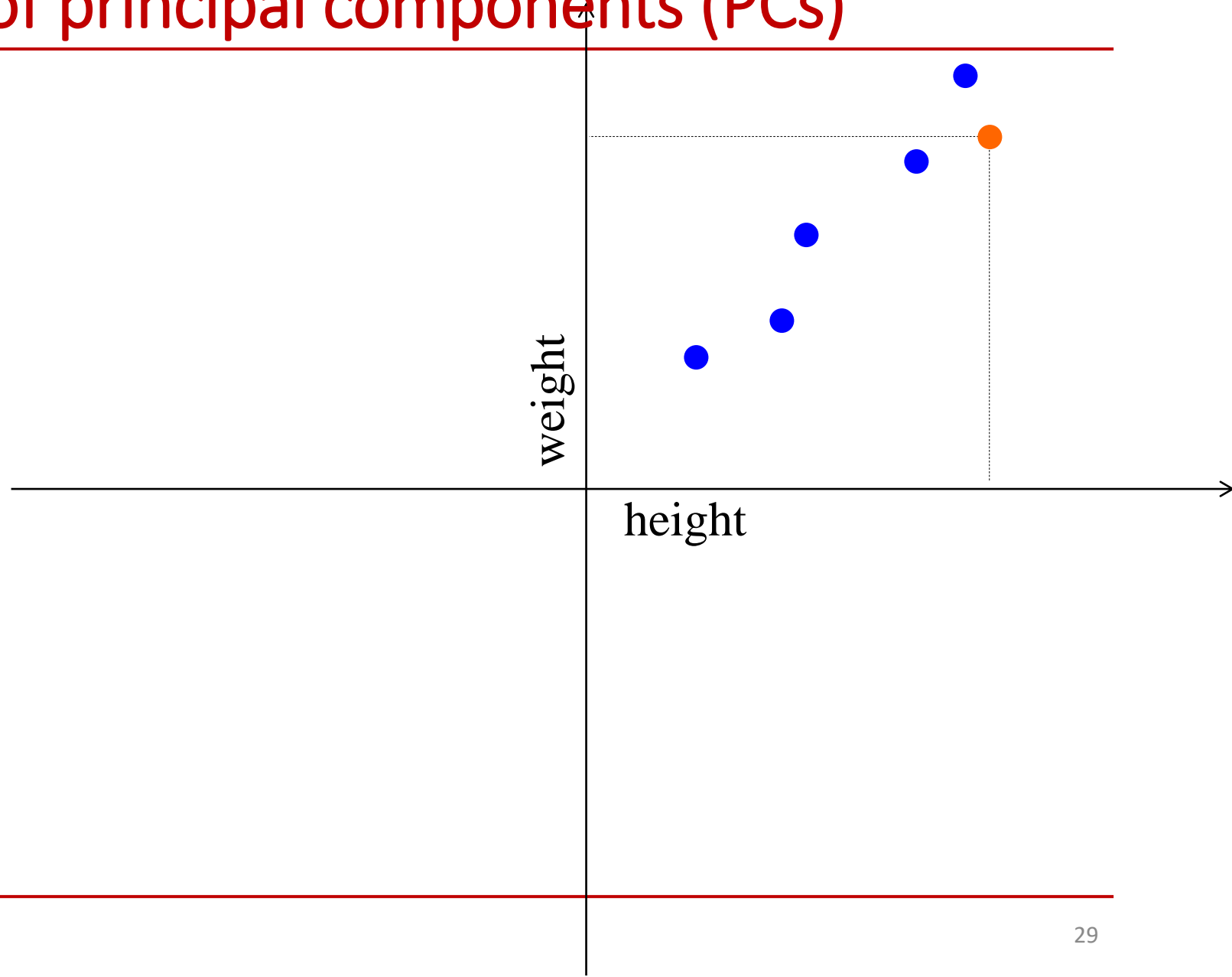
- **Mean-Centering**

- This improves the interpretability of the model.
- Calculate average value of each variable and then subtract from the data.
- Normalization is not needed.
 - However, mean centering is essential for performing Principal Component Analysis, as it gives direction of variability across the mean of the samples by creating the covariance matrix.
 - Without centering, one would be looking at variations about the origin.



Geometric picture of principal components (PCs)

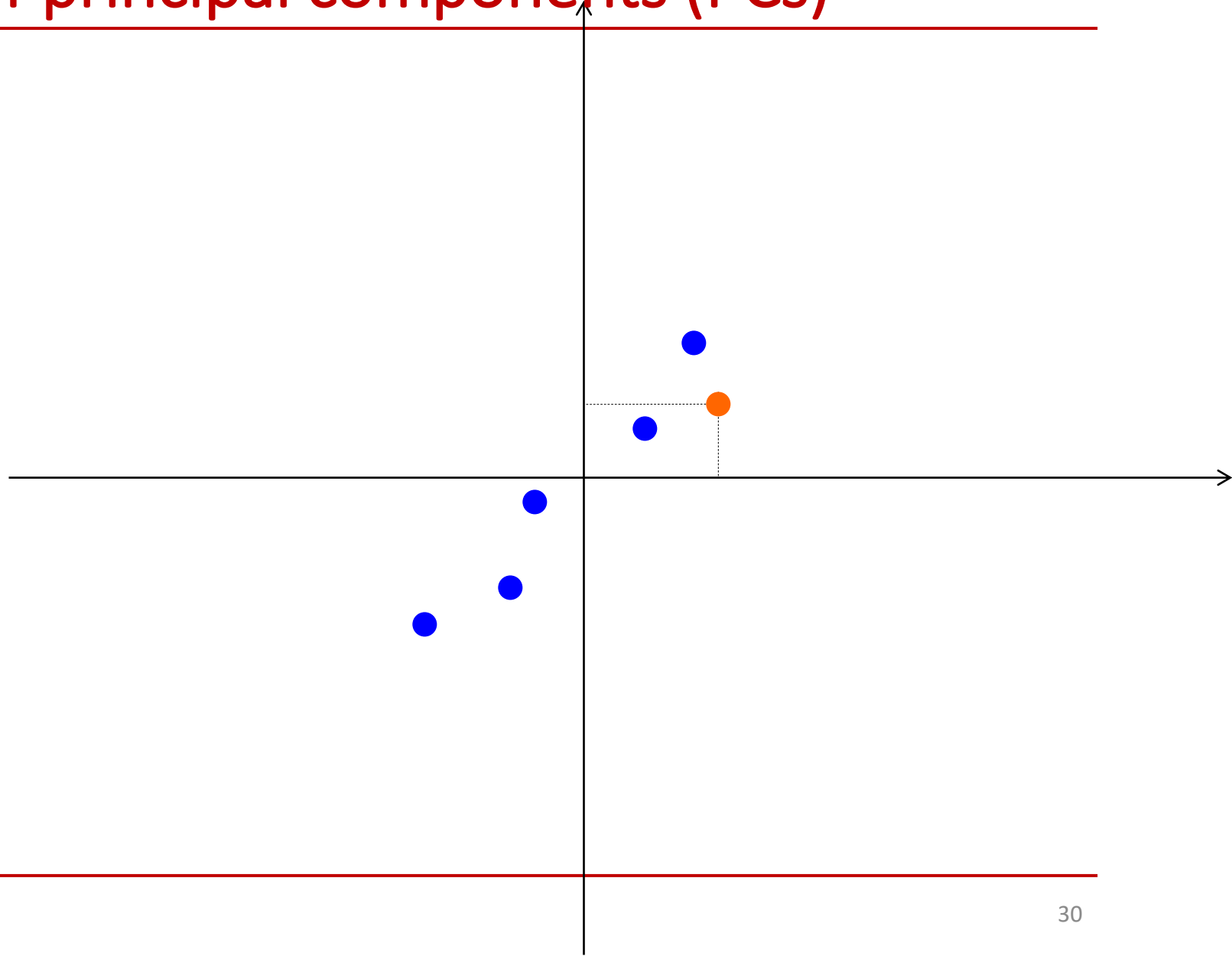
- Here is a small dataset of *opponents* we have to fight.
- Each data object is represented by its X-Y location in 2D space.
- A randomly chosen object is shown in orange.



Geometric picture of principal components (PCs)

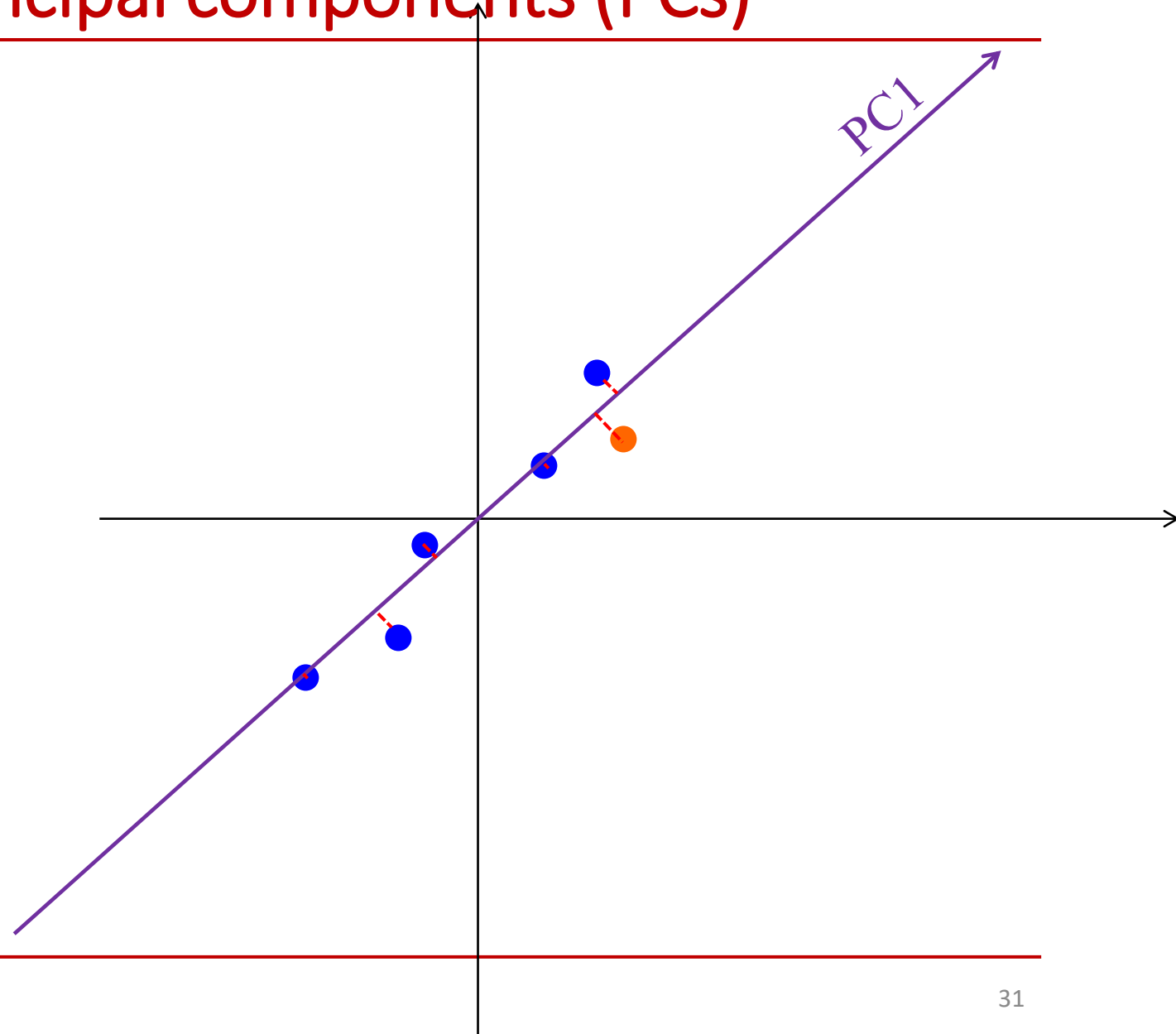
Let us z-normalize the data...

Each data object is still represented by its X-Y location in 2D space



Geometric picture of principal components (PCs)

- Let us rotate the axis to find the highest variance

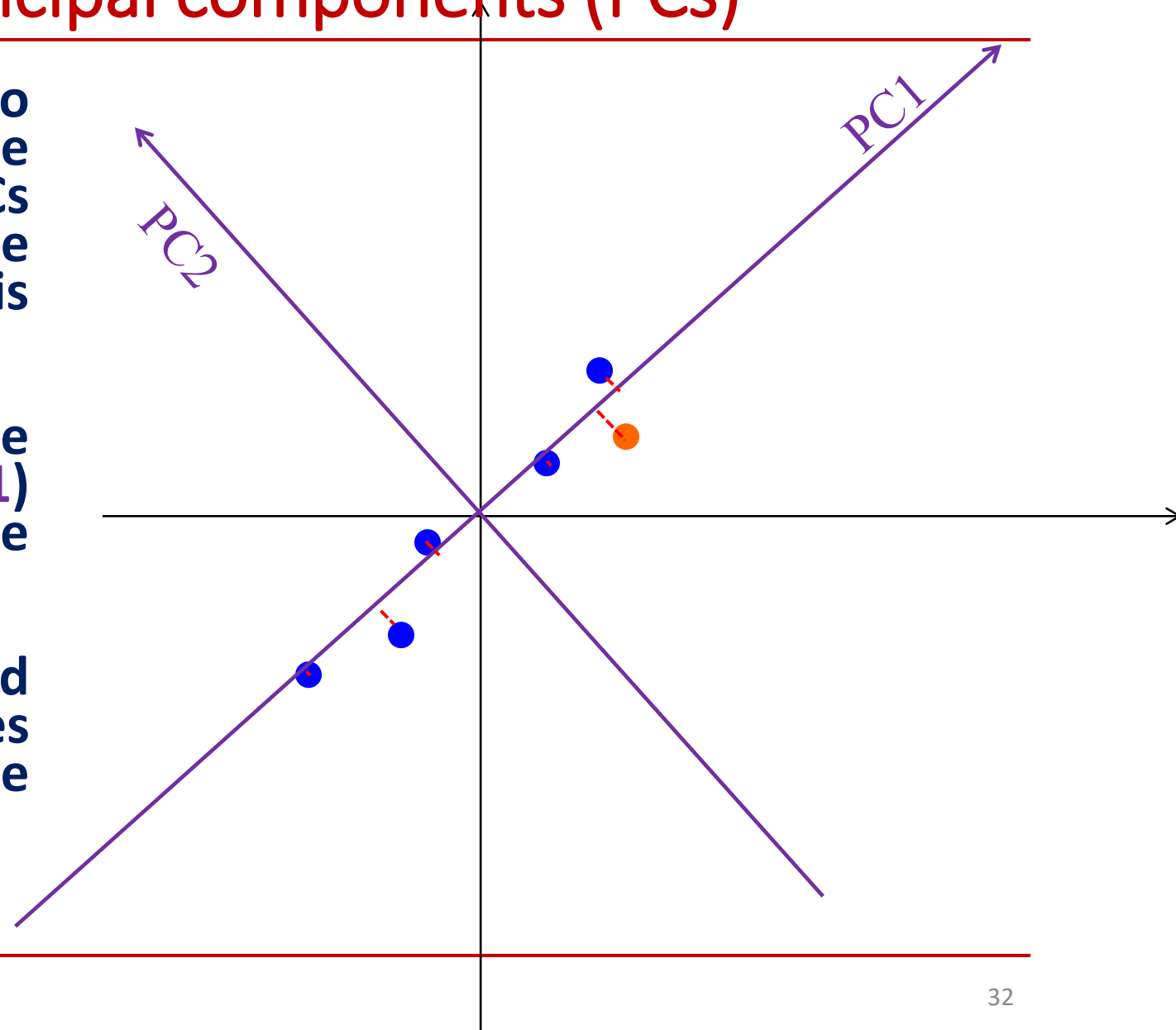


Geometric picture of principal components (PCs)

The idea is to rotate the axes so that the new axes (also called the principal components, i.e., PCs for short) are such that the variance of the data on each axis goes down from axis to axis.

The first new axis is called the first principal component (**PC1**) and it is in the direction of the greatest variance in the data.

Each new axis is constructed orthogonal to the previous ones and along the direction with the largest remaining variance.



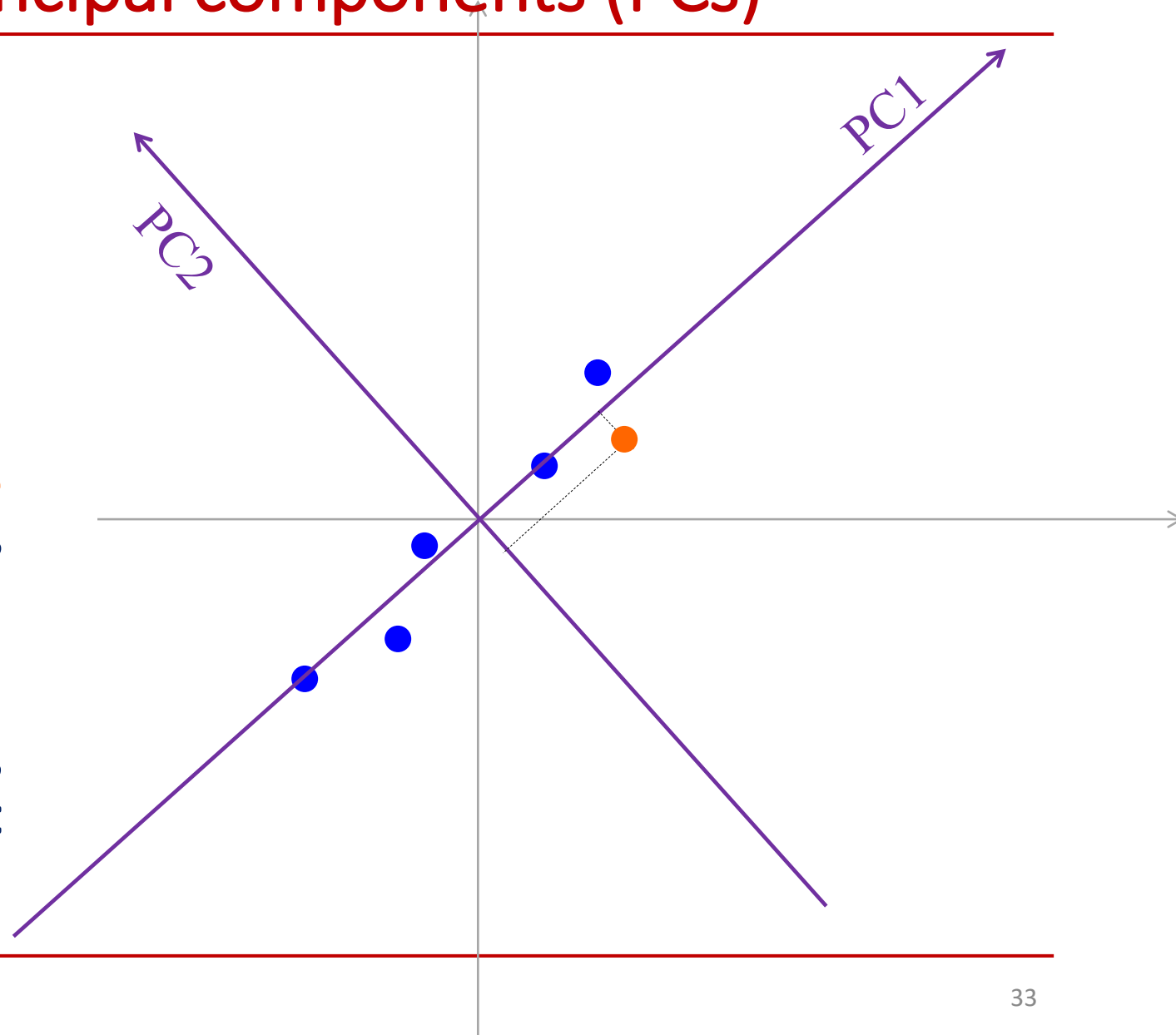
Geometric picture of principal components (PCs)

Each data object is still represented by its location in 2D space.

However, instead of X-Y space, we are now in PC1-PC2 space.

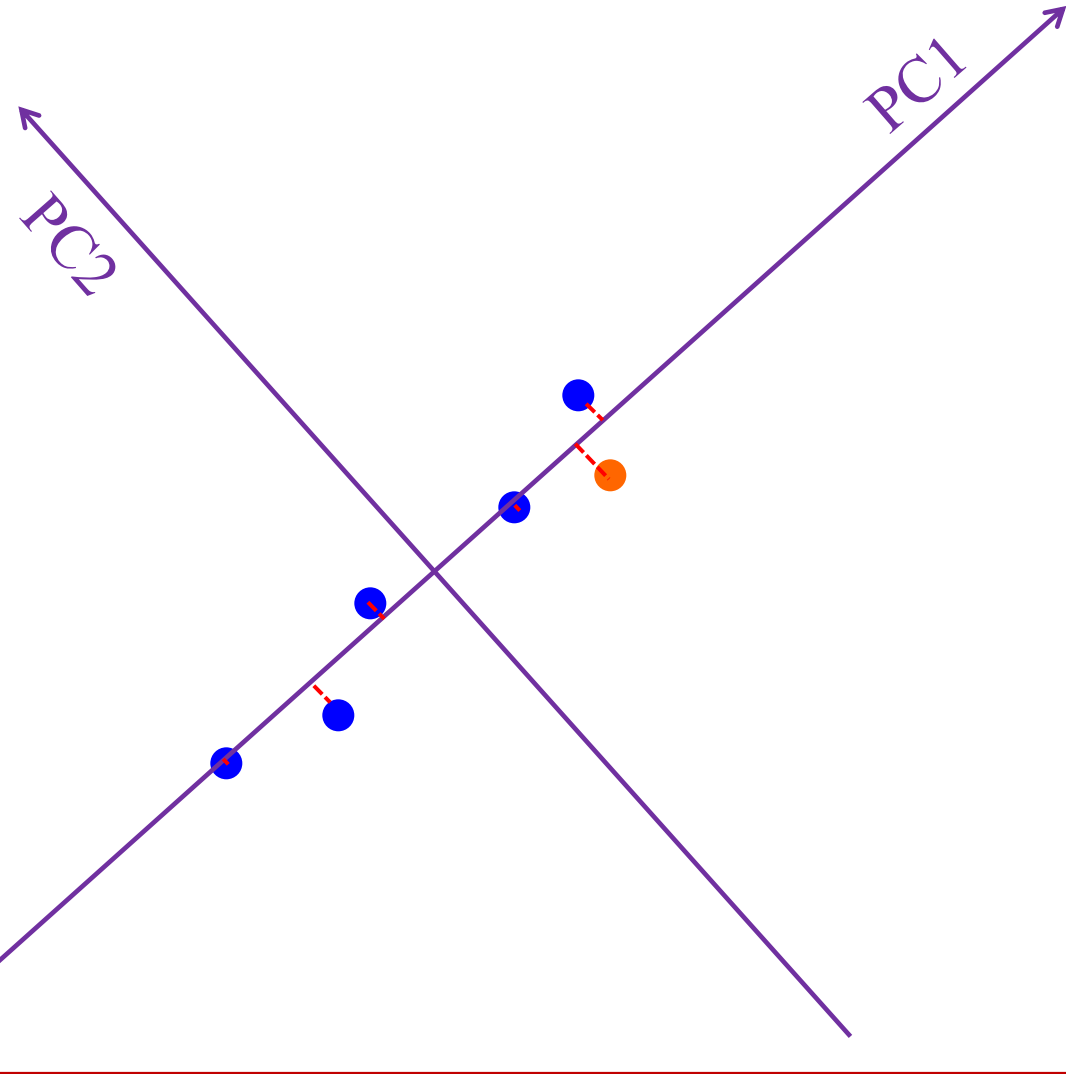
Note that for our **orange** example, the value in PC1 is large, and in PC2 is small.

This is true on average for all data points. Moreover, it is true by definition, this is what PCA does!



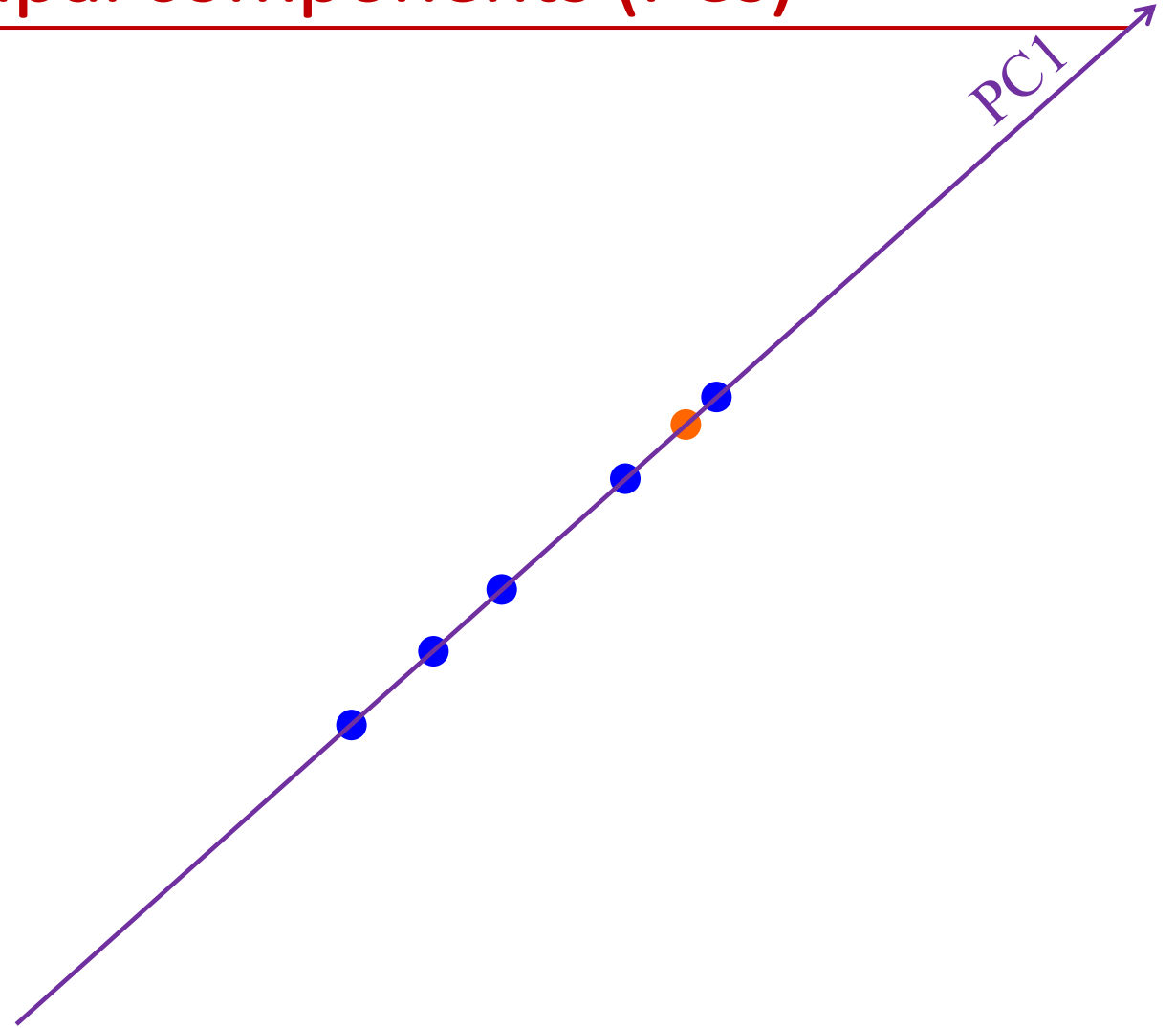
Geometric picture of principal components (PCs)

- We can project the data onto just the **PC1** axis



Geometric picture of principal components (PCs)

- We can project the data onto just the PC1 axis. This means that PC2 no longer exist
- This is a general trick.
- Starting with any N dimensions, we can do PCA, and keep just n dimensions, $n \leq N$, as use the n dimensions for clustering, classifying, indexing, plotting etc.



Algebraic definition of PCs

- Given a sample of n observations on a vector of p variables

$$X = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}^p$$

- Define the first principal component of the sample by the linear transformation

$$Z_1 = a_1^T x_j = \sum_{i=1}^p a_{i1} x_{ij}, j = 1, 2, \dots, n$$

where the vector

$$a_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{p1} \end{bmatrix}, x_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{pj} \end{bmatrix}$$

is chosen such that $\text{var}[Z_1]$ is maximum and $a_1^T a_1 = 1$, where $Z_1 = \{z_1^1, z_2^1, \dots, z_j^1, \dots, z_n^1\}$

Algebraic definition of PCs

- Likewise, define the k^{th} PC of the sample by the linear transformation

$$Z_k = a_k^T x_j = \sum_{i=1}^p a_{ik} x_{ij}, j = 1, 2, \dots, n,$$
$$\mathbf{Z}_k = \{z_1^k, z_2^k, \dots, z_j^k, \dots, z_n^k\}$$

where the vector

$$a_k = \begin{bmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{pk} \end{bmatrix}, x_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{pj} \end{bmatrix}$$

is chosen such that $\text{var}[Z_k]$ is maximum

Subjected to:

$$\text{Cov}(Z_l, Z_k) = 0, \text{ for } k > l \geq 1$$
$$\text{and } a_k^T a_k = 1$$

Algebraic derivation of coefficient vectors a_1

- To find a_1 first note that

$$\text{var}(z_1) = \frac{1}{n} \sum_{i=1}^n (a_1^T x_i - a_1^T \bar{x})^2$$

Algebraic derivation of coefficient vectors a_1

- To find a_1 first note that

$$\begin{aligned} \text{var}(z_1) &= \frac{1}{n} \sum_{i=1}^n (a_1^T x_i - a_1^T \bar{x})^2 \\ &= \frac{1}{n} \sum_{i=1}^n a_1^T (x_i - \bar{x})(x_i - \bar{x})^T a_1 = a_1^T S a_1 \end{aligned}$$

where,

$$S = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T$$

is the covariance matrix. We assume the data is centered,
hence

$$\bar{x} = 0$$

Algebraic derivation of coefficient vectors a_1

$$S = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T, \text{ where } x_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{pi} \end{bmatrix}$$

$$= \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} x_{1i}x_{1i} & x_{1i}x_{2i} & \cdots & x_{1i}x_{pi} \\ x_{2i}x_{1i} & x_{2i}x_{2i} & \cdots & x_{2i}x_{pi} \\ \vdots & \vdots & \ddots & \vdots \\ x_{pi}x_{1i} & x_{pi}x_{2i} & \cdots & x_{pi}x_{pi} \end{bmatrix}, \quad x_i x_i^T = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{pi} \end{bmatrix} [x_{1i} \quad x_{2i} \quad \cdots \quad x_{pi}]$$

$$= \frac{1}{n} \left(\begin{bmatrix} x_{11}x_{11} & x_{11}x_{21} & \cdots & x_{11}x_{p1} \\ x_{21}x_{11} & x_{21}x_{21} & \cdots & x_{21}x_{p1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p1}x_{11} & x_{p1}x_{21} & \cdots & x_{p1}x_{p1} \end{bmatrix} + \begin{bmatrix} x_{12}x_{12} & x_{12}x_{22} & \cdots & x_{12}x_{p2} \\ x_{22}x_{12} & x_{22}x_{22} & \cdots & x_{22}x_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p2}x_{12} & x_{p2}x_{22} & \cdots & x_{p2}x_{p2} \end{bmatrix} + \cdots + \begin{bmatrix} x_{1n}x_{1n} & x_{1n}x_{2n} & \cdots & x_{1n}x_{pn} \\ x_{2n}x_{1n} & x_{2n}x_{2n} & \cdots & x_{2n}x_{pn} \\ \vdots & \vdots & \ddots & \vdots \\ x_{pn}x_{1n} & x_{pn}x_{2n} & \cdots & x_{pn}x_{pn} \end{bmatrix} \right)$$

Algebraic derivation of coefficient vectors a_1

$$= \frac{1}{n} \begin{pmatrix} x_{11}x_{11} & x_{11}x_{21} \cdots & x_{11}x_{p1} & x_{12}x_{12} & x_{12}x_{22} \cdots & x_{12}x_{p2} & & x_{1n}x_{1n} & x_{1n}x_{2n} \cdots & x_{1n}x_{pn} \\ x_{21}x_{11} & x_{21}x_{21} \cdots & x_{21}x_{p1} & x_{22}x_{12} & x_{22}x_{22} \cdots & x_{22}x_{p2} & + \cdots + & x_{2n}x_{1n} & x_{2n}x_{2n} \cdots & x_{2n}x_{pn} \\ x_{p1}x_{11} & x_{p1}x_{21} \cdots & x_{p1}x_{p1} & x_{p2}x_{12} & x_{p2}x_{22} \cdots & x_{p2}x_{p2} & & x_{pn}x_{1n} & x_{pn}x_{2n} \cdots & x_{pn}x_{pn} \end{pmatrix}$$

$$S = \begin{pmatrix} Cov(A_1, A_1) & Cov(A_1, A_2) \cdots & Cov(A_1, A_p) \\ Cov(A_2, A_1) & Cov(A_2, A_2) \cdots & Cov(A_2, A_p) \\ Cov(A_p, A_1) & Cov(A_p, A_2) \cdots & Cov(A_p, A_p) \end{pmatrix}, \text{ where } A_k \text{ is the } k\text{th attribute}$$

$$var(z_1) = a_1^T \left(\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})^T \right) a_1 = a_1^T S a_1$$

Algebraic derivation of coefficient vectors a_k

- To find a_1 :

Maximize: $var(z_1)$

Subjected to: $a_1^T a_1 = 1$

- Let λ be a Lagrange multiplier, then

$$L = a_1^T S a_1 - \lambda(a_1^T a_1 - 1)$$

By differentiating with respect to each element of element of a_1 and writing in vectorized notation

$$\frac{\partial L}{\partial a_1} = S a_1 - \lambda a_1$$

At maximization point

$$\frac{\partial L}{\partial a_1} = 0 \Rightarrow S a_1 = \lambda a_1$$

- Therefore, a_1 is an eigenvector of S , corresponding to the largest eigenvalue $\lambda \equiv \lambda_1$

Algebraic derivation of coefficient vectors a_k

- We have maximized

$$\begin{array}{ll} \text{Maximize:} & \text{var}(z_1) = a_1^T S a_1 \\ \text{Subjected to:} & a_1^T a_1 = 1 \end{array}$$

- To find the next coefficient vector a_2

$$\begin{array}{ll} \text{Maximizing:} & \text{var}(z_2) \\ \text{Subjected to:} & \text{cov}(z_1, z_2) = 0 \\ & \text{and } a_2^T a_2 = 1 \end{array}$$



- First, note that

$$\text{cov}(z_1, z_2) = \text{cov}(a_1^T X, a_2^T X) = \frac{1}{n} a_1^T X (a_2^T X)^T = a_1^T S a_2 = \lambda a_1^T a_2$$

- Then let γ and φ be Lagrange multipliers, and maximize

$$L = a_2^T S a_2 - \gamma(a_2^T a_2 - 1) - \varphi a_2^T a_1$$

Algebraic derivation of coefficient vectors a_k

$$\mathbf{L} = a_2^T S a_2 - \lambda(a_2^T a_2 - 1) - \phi a_2^T a_1$$

$$\frac{\partial L}{\partial a_2} = S a_2 - \gamma a_2 - \phi a_1$$

At maximization point $\frac{\partial L}{\partial a_2} = 0$,

$$\Rightarrow S a_2 - \gamma a_2 - \phi a_1 = 0$$

By multiplying both side with a_1^T

$$\Rightarrow a_1^T S a_2 - \gamma a_1^T a_2 - \phi a_1^T a_1 = 0$$

$$\Rightarrow \lambda a_1^T a_2 - \gamma a_1^T a_2 - \phi a_1^T a_1 = 0$$

$$\Rightarrow \lambda a_1^T a_2 - \gamma a_1^T a_2 - \phi a_1^T a_1 = 0 \Rightarrow \lambda \cdot 0 - \gamma \cdot 0 - \phi a_1^T a_1 = 0 \Rightarrow \phi = 0$$

Algebraic derivation of coefficient vectors a_k

- We find that a_2 is also an eigenvector of S
- whose eigenvalue $\lambda = \lambda_2$ is the second largest.
- In general

$$\text{var}(z_k) = a_k^T S a_k = \lambda_k$$

- The k^{th} largest eigenvalue of S is the variance of the k^{th} PC.
- The k^{th} PC retains the k^{th} greatest fraction of the variation in the sample.

Main steps for computing PCs

- Pre-process the data.
- Form the covariance matrix S .
- Compute its eigenvectors: $\{a_i\}_{i=1}^p$
- Use the first d eigenvectors $\{a_i\}_{i=1}^d$ to form the d PCs.
- The transformation P is given by
$$P \leftarrow [a_1, a_2, \dots, a_d]$$

PCA Assumptions

- PCA assumes that all basis vectors $\{p_1, \dots, p_m\}$ are orthonormal.
- Hence, in the language of linear algebra, PCA assumes P is an orthonormal matrix.
- Secondly, PCA assumes the directions with the largest variances are the most “important” or in other words, most principal.

PCA Example

- **Data**

2.5	0.5	2.2	1.9	3.1	2.3	2	1	1.5	1.1
2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

- **PCA process –STEP 1**

- **Zero Mean data**

0.69	-1.31	0.39	0.09	1.29	0.49	0.19	-0.81	-0.31	-0.71
0.49	-1.21	0.99	0.29	1.09	0.79	-0.31	-0.81	-0.31	-1.01

PCA process –STEP 2

- Calculate the covariance matrix

$$\text{cov} = \begin{pmatrix} 0.616555556 & 0.615444444 \\ 0.615444444 & 0.716555556 \end{pmatrix}$$

- Since the non-diagonal elements in this covariance matrix are positive, we should expect that both the x and y variable increase together

PCA process –STEP 3

- Calculate the eigenvectors and eigenvalues of the covariance matrix

$$\text{eigenvalues} = \begin{matrix} 0.0490833989 \\ 1.2840277100 \end{matrix}$$

$$\text{eigenvectors} = \begin{matrix} -0.735178656 & -0.677873399 \\ 0.677873399 & -0.735178656 \end{matrix}$$

PCA process –STEP 3

- Once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives you the components in order of significance.

PCA process –STEP 4

- **Feature Vector**

Feature Vector = (eig1 eig2 eig3 ... eign)

- **We can either form a feature vector with both of the eigenvectors:**

-.677873399 -.735178656

-.735178656 .677873399

- **Or, we can choose to leave out the smaller, less significant component and only have a single column:**

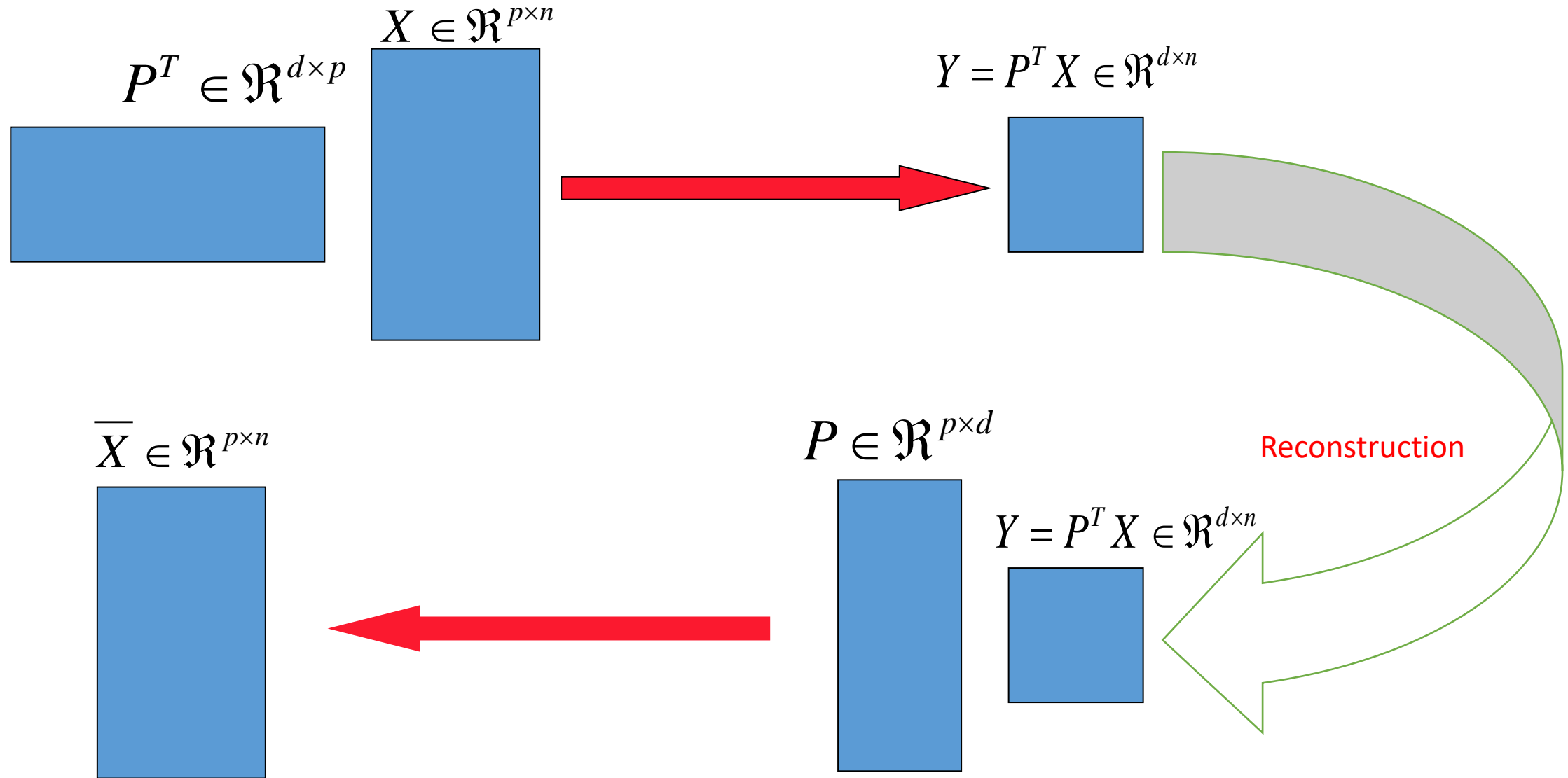
- 0.677873399

- 0.735178656

PCA process –STEP 5

- Reduce dimensionality and form *feature vector* the eigenvector with the *highest* eigenvalue is the *principal component* of the data set.

Reconstruction of Original Data




Optimality property of PCA

Main theoretical result:

- The matrix P consisting of the first d eigenvectors of the covariance matrix S solves the following min problem:

$$\min_{P \in \mathbb{R}^{p \times d}} \|X - P(P^T X)\|_F^2 \text{ subject to } P^T P = I_d$$


$$\|X - \bar{X}\|_F^2$$

reconstruction error

- PCA projection minimizes the reconstruction error among all linear projections of size d .

Reference

- **Feature Extraction, K. Ramachandra Murthy**
- **A Tutorial on Principal Component Analysis, Aly A. Farag Shireen Elhabian University of Louisville, CVIP Lab**