Linear Algebra

Principal Component Analysis (PCA)

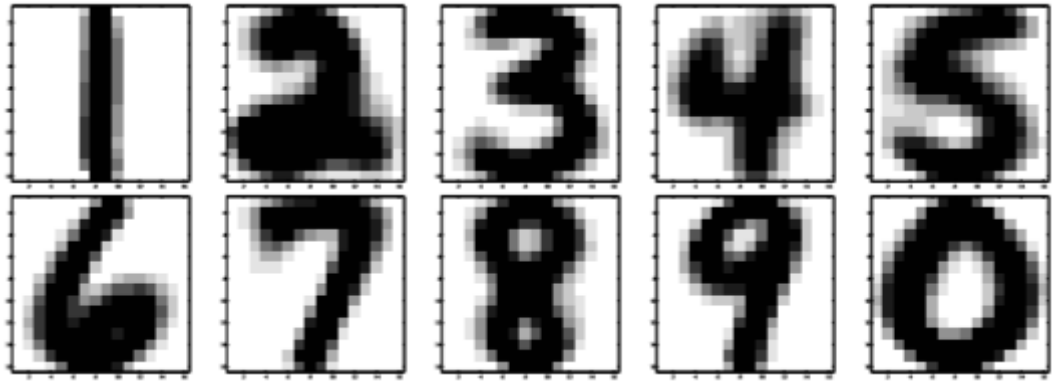
Dr. Anahita Zarei

Overview

- Curse of Dimensionality
- Methods to Reduce Dimensionality
- Principal Component Analysis
  - Mean Deviation Form
  - Covariance Matrix
  - Projection Onto New Dimensions
- Reading: section 7.5 from Lay

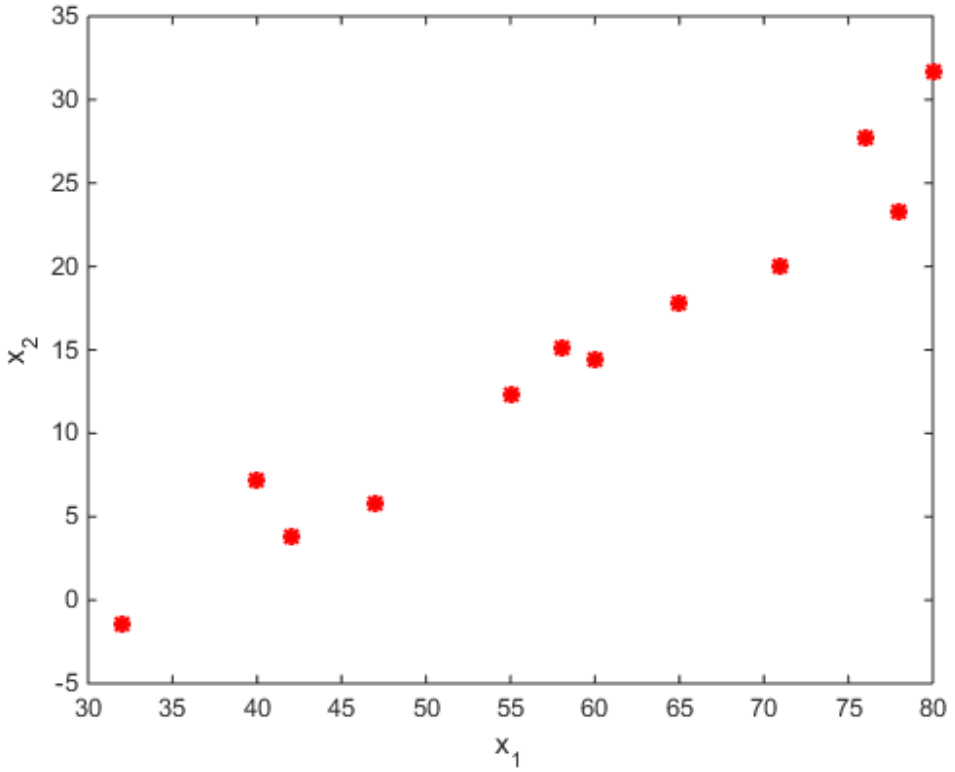
Why PCA?

- PCA has to do with reducing the dimension of multivariate data.
- Datasets are usually high dimensional.
- Homework problems:
  - Heathcare Studies: Predict patients compliance based on **30** attributes.
  - Digit Recognition: Classify the digits into one of 10 classes based on 16x16 (= **256**) pixel-images.
- Any realistic data has a high number of dimension.
  - Any text processing application can potentially deal with **billions** of words.



Why Reducing the Dimensionality?

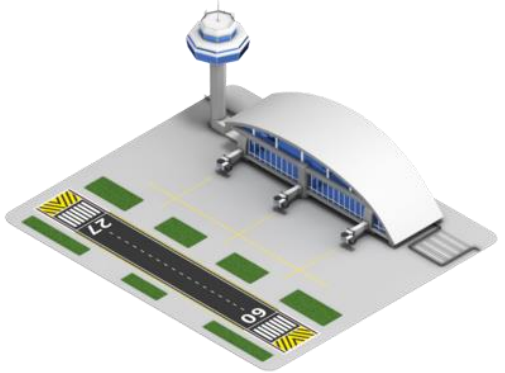
- **The true dimensionality of data is often lower than the observed dimensionality**.
- Example: You get a data set from a weather center that's collected over a period of 12 months at a particular region.
  - Data has the following format : (x_1, x_2)
  - What's the dimension of the data?
  - It turns out that x1 is temperature in Fahrenheit and x2 is temperature in Celsius.
  - So your data has only 1 dimension.



Why Reducing the Dimensionality?

- Example: You get a data set from a monitoring agency with the following attributes:
  - x_1: num. of traffic accidents
  - x_2: num. of school closures
  - x_3: num. of delayed flights
  - x_4: num. of wild fires
  - x_5: num. of patients with heat stroke

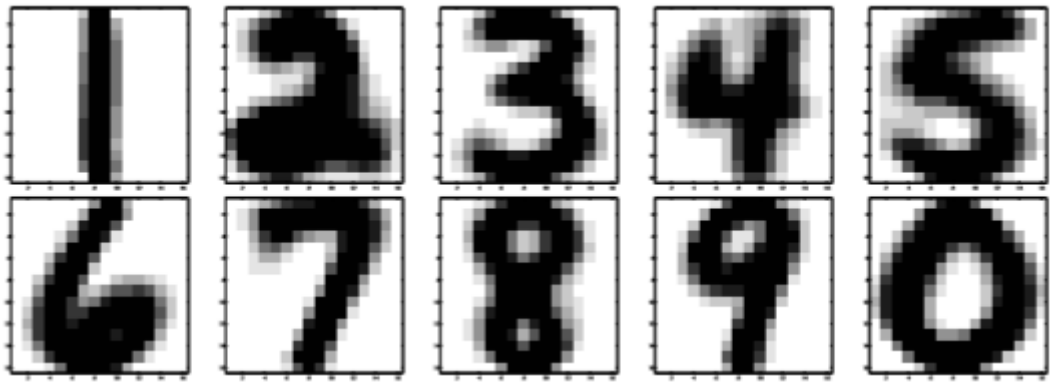
- Although, at the surface these all seem like different attributes, there's a single factor that can explain lots of these observations: temperature!
- **A machine learning algorithm should look for the single variable that counts for the others rather than looking at every individual one.**



Why Reducing the Dimensionality?

- Example: Handwritten digits in MNIST data set contains 16x16 images where each pixel can have a value of 0 or 1.
- This will result in 2^{256} possible events.
- However, many of these results will never happen, and true dimensionality is much smaller.

- **Data sets may have redundant attributes that don't contribute much to learning algorithms.**
- **Using the original representation 'wastes' the machine learning algorithm on the outcomes that will never happen.**

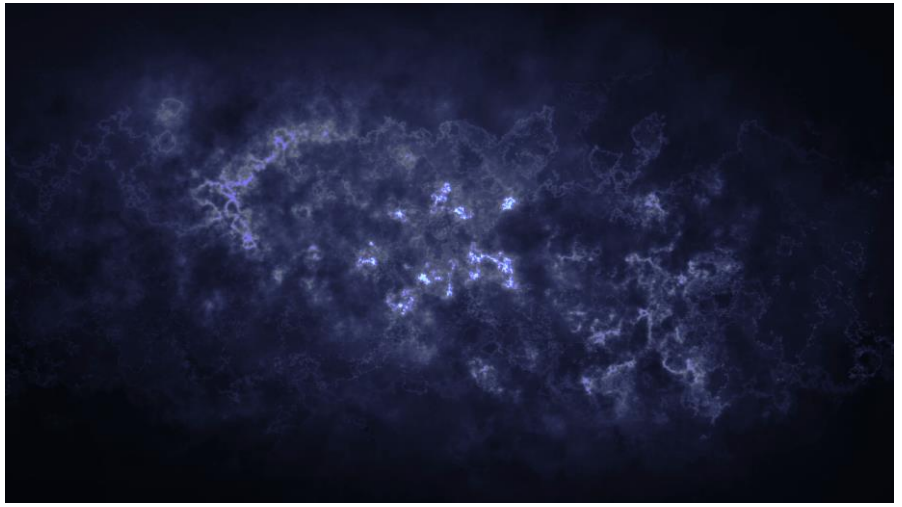


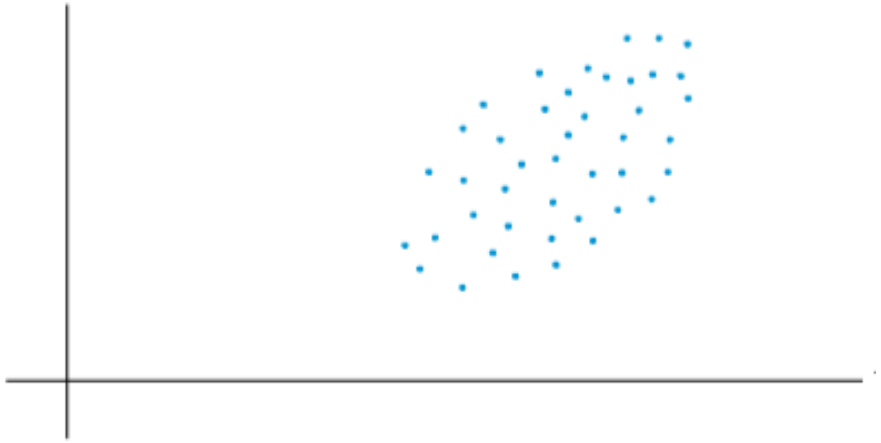
How to Reduce Dimensionality?

- Goal:
  - Try to preserve as much structure in the data as possible
  - Try to select/generate features that are discriminative
- Methods:
  - Use expert knowledge
    - E.g. An expert tells you that one of the variable can count for some other attributes
  - Feature selection
    - Simplest reduction method.
    - Select a subset of d available attributes that contribute the most in information gain.
    - Throw away rest of the attributes.
  - Feature extraction
    - Use all of the original data and combine them in some way to form a smaller set of new attributes.
    - The new attributes don't maintain the same physical meaning as in the original data set.

Attributes and Coordinates

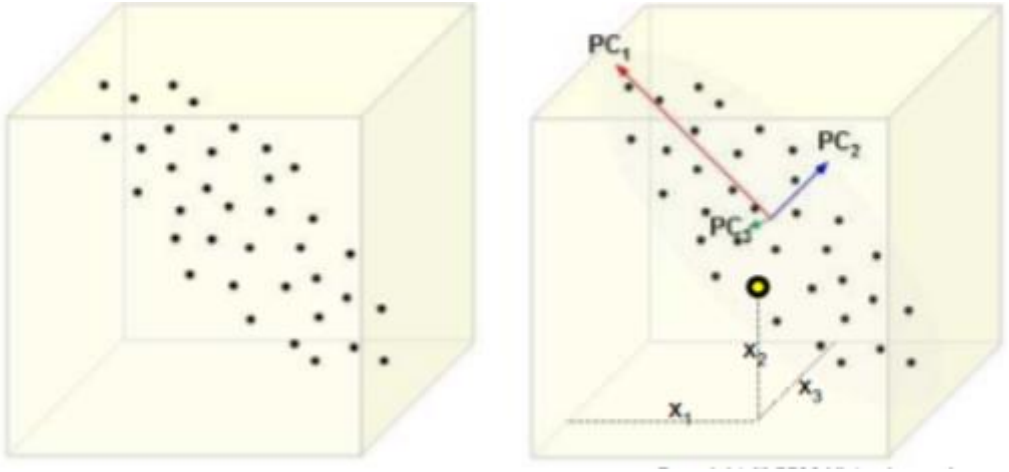
- In a dataset with k numeric attributes, you can visualize the data as a cloud of points in k-dimensional space:
  - the stars in the galaxy,
  - a swarm of flies frozen in time,
  - a two-dimensional scatter plot on paper.
- The attributes represent the coordinates of the space.
- But the axes you use, the coordinate system itself, is arbitrary.





Idea of PCA

- In machine learning there often is a preferred coordinate system, defined by the very data itself.
- The idea of principal components analysis is to use a special coordinate system that depends on the cloud of points as follows:
  - Place the first axis in the direction of greatest variance of the points to **maximize the variance** along that axis.
  - Choose the second axis in **perpendicular** to the first one, the way that maximizes the variance along it.
  - Continue, choosing each axis to maximize its share of the remaining variance.
  - Find the new coordinate by projecting the data points onto new set of axes.

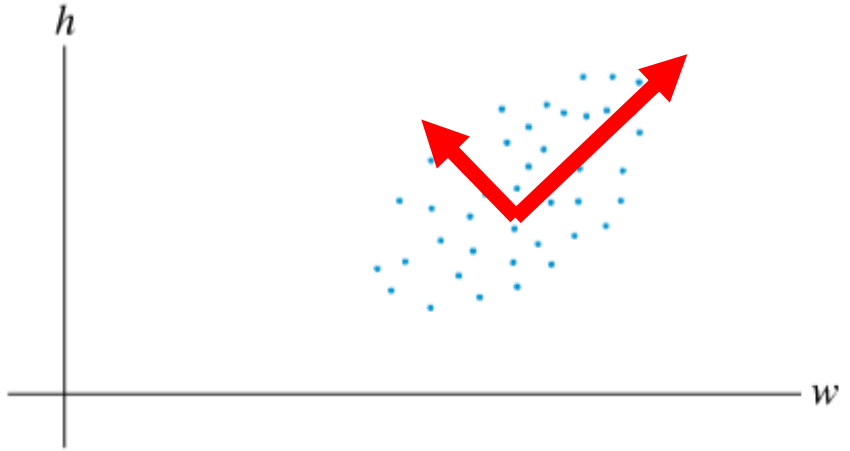


Example

An example of two-dimensional data is given by a set of weights and heights of N high school students. Let X_j denote the observation vector in R^2 that lists the weight and height of the j^{th} student. If w denotes weight and h height, then the matrix of observations has the form

$$\begin{bmatrix} w_1 & w_2 & \cdots & w_N \\ h_1 & h_2 & \cdots & h_N \end{bmatrix}$$

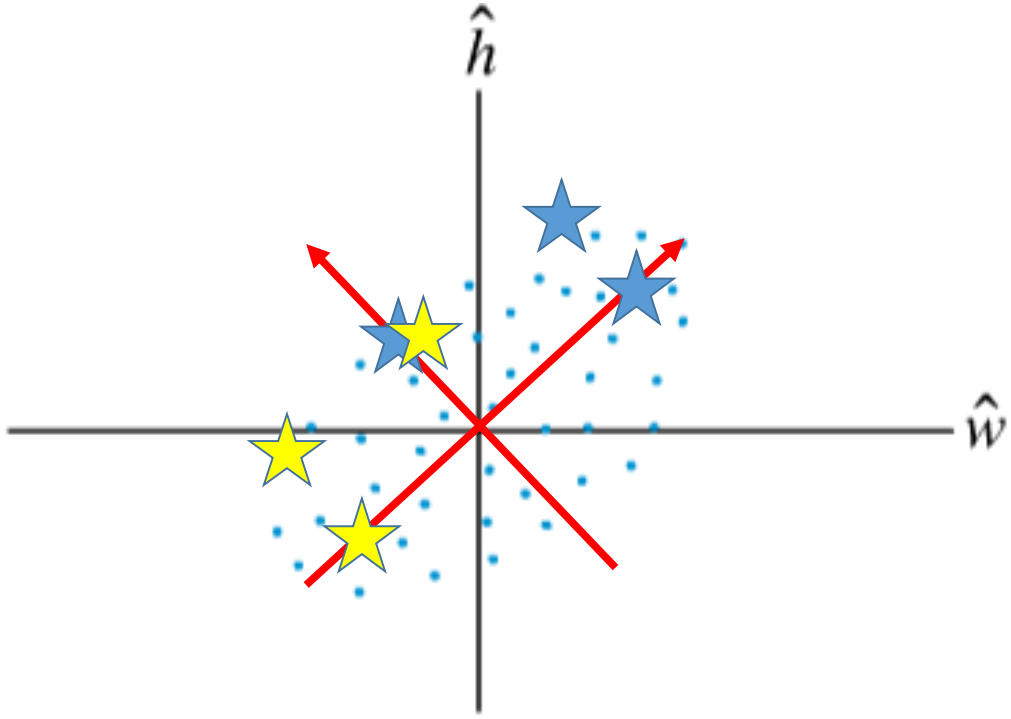
$$\mathbf{X}_1 \quad \mathbf{X}_2 \quad \quad \mathbf{X}_N$$



- Note that in two dimensions you only have one choice for the second axis. Its direction is determined by the first axis
- However, in three dimensions it can lie anywhere in the plane perpendicular to the first axis, and in higher dimensions there are even more choices, although it is always constrained to be perpendicular to the first axis

Why Greatest Variability?

- The dimensions with the greatest variability preserve the distances.
- Distance between data points is a manifestation of the data structure. Why?
- Because we assume nearby things are similar. Similarity is very important for learning algorithms.
- Therefore, the high variance dimensions preserve the structure.
- Note that while the relative distance between some points changes, the line with the largest variability preserve the most distances as accurately as possible, overall.

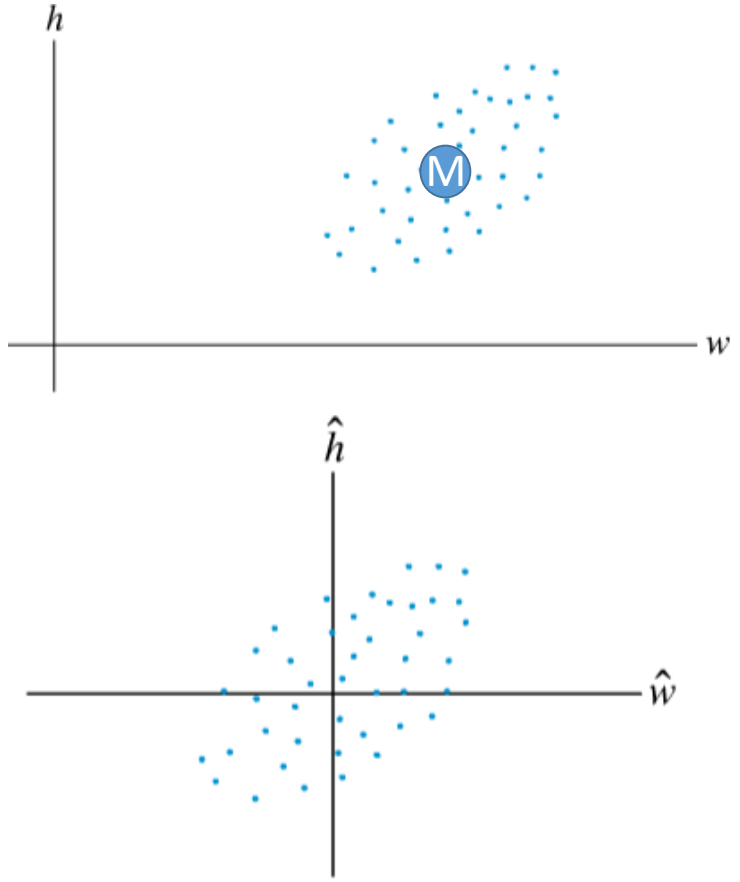


How to Get the Principal Components?

- The first step is to **center** the data points.
- i.e. subtract the mean of each attribute from the corresponding coordinate.
- Example: Consider the matrix of observations:

$$\begin{bmatrix} w_1 & w_2 & \cdots & w_N \\ h_1 & h_2 & \cdots & h_N \end{bmatrix}$$

$$\mathbf{X}_1 \quad \mathbf{X}_2 \quad \quad \mathbf{X}_N$$



- The sample mean, M is given by

$$\mathbf{M} = \frac{1}{N}(\mathbf{X}_1 + \cdots + \mathbf{X}_N)$$

- The sample mean is the point in the center.
- For k = 1…N, let $\hat{\mathbf{X}}_k = \mathbf{X}_k - \mathbf{M}$
- The columns of B have a zero sample mean. $B = [\hat{\mathbf{X}}_1 \quad \hat{\mathbf{X}}_2 \quad \cdots \quad \hat{\mathbf{X}}_N]$
- B is said to be in **mean-deviation** form.

Example

- Three measurements are made on each of four individuals in a random sample from a population.
- The observation vectors are:

$$\mathbf{X}_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \quad \mathbf{X}_2 = \begin{bmatrix} 4 \\ 2 \\ 13 \end{bmatrix}, \quad \mathbf{X}_3 = \begin{bmatrix} 7 \\ 8 \\ 1 \end{bmatrix}, \quad \mathbf{X}_4 = \begin{bmatrix} 8 \\ 4 \\ 5 \end{bmatrix}$$

- Determine the coordinate of the centered data.

```
> X = matrix(c(1,2,1,4,2,13,7,8,1,8,4,5), nrow = 3)
> X
     [,1] [,2] [,3] [,4]
[1,]    1    4    7    8
[2,]    2    2    8    4
[3,]    1   13    1    5
```

```
> m=rowMeans(X)
> m
[1] 5 4 5
```

```
> M = matrix(rep(m,4), nrow = 3)
> M
     [,1] [,2] [,3] [,4]
[1,]    5    5    5    5
[2,]    4    4    4    4
[3,]    5    5    5    5
```

```
> B = X - M
> B
     [,1] [,2] [,3] [,4]
[1,]   -4   -1    2    3
[2,]   -2   -2    4    0
[3,]   -4    8   -4    0
```

How to Get the Principal Components? – cont.

- The second step is to find the **covariance matrix** for the *d* features.
- A $d \times d$ covariance matrix will look like $\begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{1d} \\ \vdots & \ddots & \vdots \\ \sigma_{d1} & \cdots & \sigma_d^2 \end{bmatrix}$
- The main diagonal of the covariance matrix, contains the variances. e.g. σ_1^2 denotes how spread out the data are along 1st dimension.
- The off diagonal elements indicates if features change together (i.e. if x_1 increases x_2 increases) or in opposite direction (i.e. if x_1 increases x_2 decreases)
- The covariance matrix is symmetric.

Example

- The sample covariance matrix of a data set is as follows: $\begin{bmatrix} 10 & 6 & 0 \\ 6 & 8 & -8 \\ 0 & -8 & 32 \end{bmatrix}$
- Interpret the numbers.
  1. Since the covariance matrix is 3 by 3, the data has a dimension of 3.
  2. The entries in the third dimension has the widest spread of values compare to the first and second dimensions.
  3. The first and second dimensions are positively correlated.
  4. The second and third dimensions are negatively correlated.
  5. The first and third dimensions are **uncorrelated.**
- Analysis of the multivariate data is greatly simplified when most or all of the variables are uncorrelated, that is, when the **covariance matrix of the data is diagonal** or nearly diagonal.

Calculating the Sample Covariance Matrix

- $cov(x_1, x_2) = \frac{1}{N-1}\sum_{i=1}^{N}(x_{1i} - m_1)(x_{2i} - m_2)$
- Since we already centered the data, m_1 and m_2 are zero
- $cov(x_1, x_2) = \frac{1}{N-1}\sum_{i=1}^{N} x_{1i}x_{2i}$
- What's the covariance expression in matrix form?
- Let B be the centered data points: $B = \begin{bmatrix} x_{11} - m_1 & \cdots & x_{1N} - m_1 \\ \vdots & \ddots & \vdots \\ x_{d1} - m_d & \cdots & x_{dN} - m_d \end{bmatrix}$
- Then $BB^T = \begin{bmatrix} x_{11} - m_1 & \cdots & x_{1N} - m_1 \\ \vdots & \ddots & \vdots \\ x_{d1} - m_d & \cdots & x_{dN} - m_d \end{bmatrix} \begin{bmatrix} x_{11} - m_1 & \cdots & x_{d1} - m_d \\ \vdots & \ddots & \vdots \\ x_{1N} - m_1 & \cdots & x_{dN} - m_d \end{bmatrix}$
- Therefore sample covariance will be:

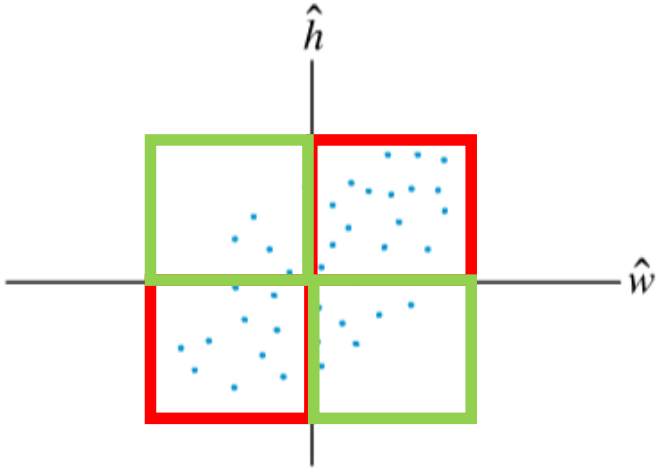
$$s = \frac{1}{N-1}BB^T$$

Sample Covariance Matrix

- Consider the expression

$cov(x_1, x_2) = \frac{1}{N-1}\sum_{i=1}^{N} x_{1i}x_{2i}.$

- Which one the points in this diagram contribute positively or negatively to the covariance?
- Points in the first and third quadrants very together (they're both above or below the mean simultaneously). Points in the second and forth quadrants very in opposite directions.
- Therefore, points in the red squares contribute positively $(\sum_{i=1}^{N} x_{1i}x_{2i} > 0)$and points in green squares contribute negatively $(\sum_{i=1}^{N} x_{1i}x_{2i} < 0)$ to the covariance expression.



Example

- Three measurements are made on each of four individuals in a random sample from a population.
- The observation vectors are:

$$\mathbf{X}_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \quad \mathbf{X}_2 = \begin{bmatrix} 4 \\ 2 \\ 13 \end{bmatrix}, \quad \mathbf{X}_3 = \begin{bmatrix} 7 \\ 8 \\ 1 \end{bmatrix}, \quad \mathbf{X}_4 = \begin{bmatrix} 8 \\ 4 \\ 5 \end{bmatrix}$$

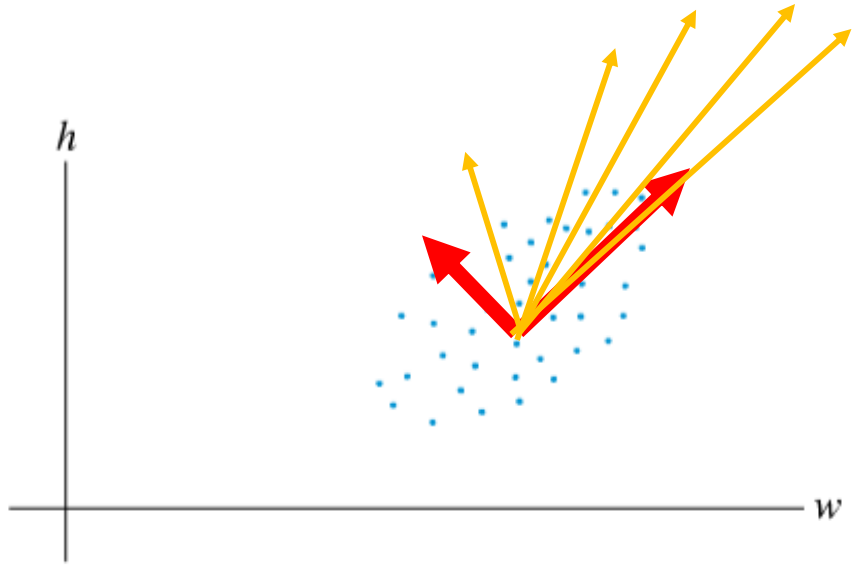
- Determine the covariance matrix.

```
> X = matrix(c(1,2,1,4,2,13,7,8,1,8,4,5), nrow = 3)
> X
     [,1] [,2] [,3] [,4]
[1,]    1    4    7    8
[2,]    2    2    8    4
[3,]    1   13    1    5
> m=rowMeans(X)
> m
[1] 5 4 5
> M = matrix(rep(m,4), nrow = 3)
> M
     [,1] [,2] [,3] [,4]
[1,]    5    5    5    5
[2,]    4    4    4    4
[3,]    5    5    5    5
> B = X - M
> B
     [,1] [,2] [,3] [,4]
[1,]   -4   -1    2    3
[2,]   -2   -2    4    0
[3,]   -4    8   -4    0
```

```
> S=1/(ncol(B)-1)*B%*%t(B)
> S
     [,1] [,2] [,3]
[1,]   10    6    0
[2,]    6    8   -8
[3,]    0   -8   32
```

Covariance Matrix Property

- The covariance matrix has the following property:
- If we take any vector on the plane, multiply it by the covariance matrix, and multiply that outcome by the covariance matrix again and repeat this process, the outcome will eventually point in the direction of the greatest variance.
- In other words, multiplying by the covariance matrix turns any vector toward the dimension of the greatest variance in the data.
- This rotation will stop once the vector comes to the steady state. (i.e. it doesn't spin 360 degrees on the plane).



How to Get the Principal Components? – cont.

- Based on observations from previous slide, we need to look for the vector of convergence to find the first principal component.
- Put differently, we should find the factor that doesn't get turned upon multiplication by covariance matrix.
- The above statement is equivalent to $\boldsymbol{Su_1 = \lambda u_1}$.
- Therefore, final step in finding the principal components is to find the eigenvectors of the covariance matrix.
- A d-dimensional data set has a dxd covariance matrix.
- The eigenvectors with the largest eigenvalues will correspond the dimension of greatest variance.

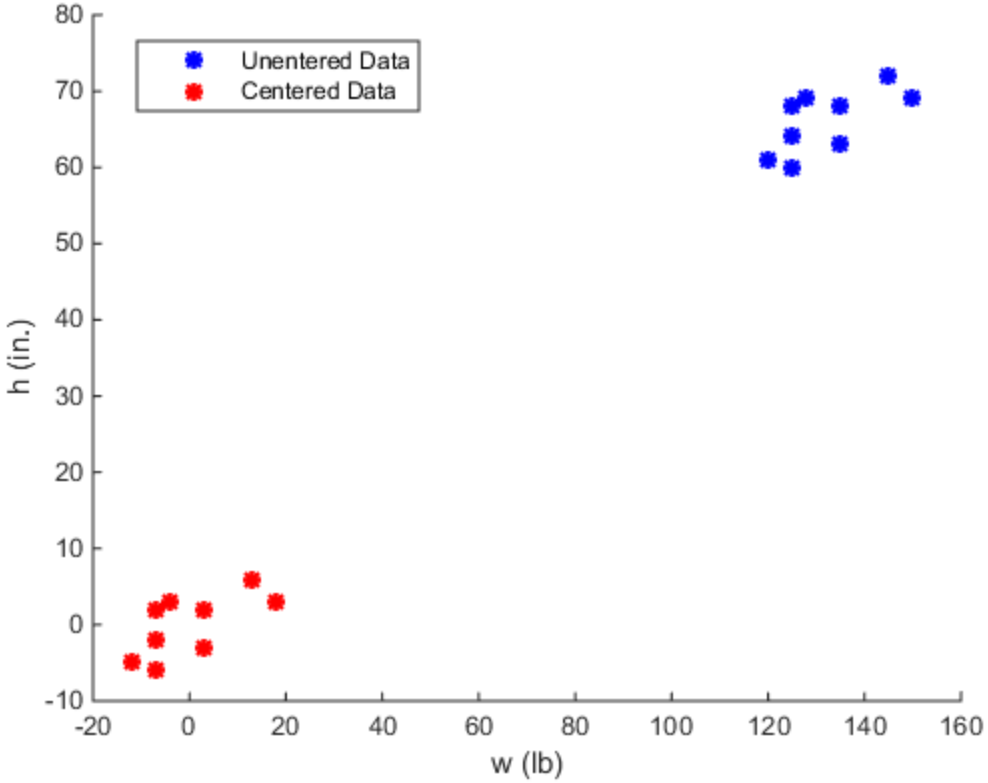
Example

- The following data is a set of weights and heights of 9 high school students.
- w = [120 125 125 135 145 135 128 125 150
- h = [61 60 64 68 72 63 69 68 69];

a) Find the covariance matrix.

b) Choose a random vector on the plane and determine after how many iterations it converges (normalize after each iteration) What's the direction of convergence?

c) Determine the principal components.



Solution

```
> w = c(120,125,125,135,145,135,128,125,150)
> h = c(61,60,64,68,72,63,69,68,69)
> X = rbind(w,h)
> X
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
w  120  125  125  135  145  135  128  125  150
h   61   60   64   68   72   63   69   68   69
> m = rowMeans(X)
> m
  w   h
132  66

> B = X - matrix(rep(m,ncol(X)), nrow = 2)
> B
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]
w  -12   -7   -7    3   13    3   -4   -7   18
h   -5   -6   -2    2    6   -3    3    2    3

> S = (1/(ncol(B)-1))*B%*%t(B)
> S
        w      h
w 102.250 27.375
h  27.375 17.000
```

```
> a1 = matrix(1, nrow = 2)
> a1
     [,1]
[1,]    1
[2,]    1
> a2 = S%*%a1
> a2 = a2/norm(a2,'2')
> a2
        [,1]
w 0.9460978
h 0.3238811
> a3 = S%*%a2
> a3 = a3/norm(a3,'2')
> a3
        [,1]
w 0.9585130
h 0.2850487
> a4 = S%*%a3
> a4 = a4/norm(a4,'2')
> a4
        [,1]
w 0.9594534
h 0.2818675
> a5 = S%*%a4
> a5 = a5/norm(a5,'2')
> a5
        [,1]
w 0.9595293
h 0.2816087
```

```
> eigen(S)
$values
[1] 110.283477   8.966523

$vectors
           [,1]       [,2]
[1,] -0.9595361  0.2815858
[2,] -0.2815858 -0.9595361
```