Linear Algebra

Principal Component Analysis (PCA)

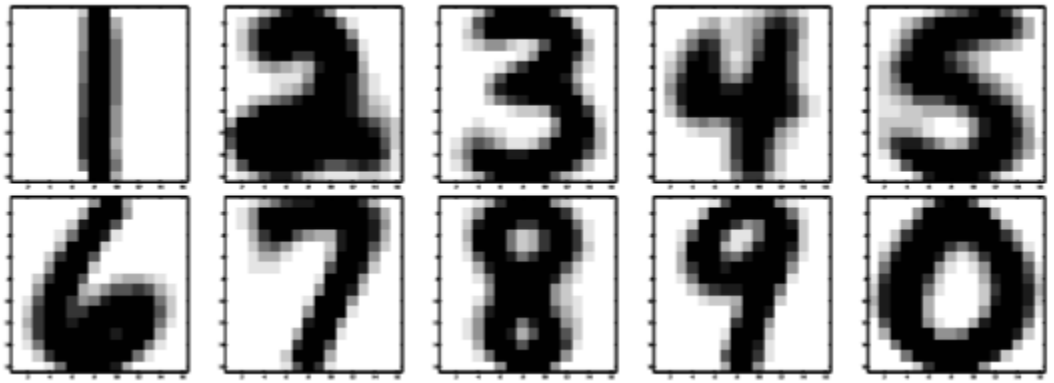
Dr. Anahita Zarei

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

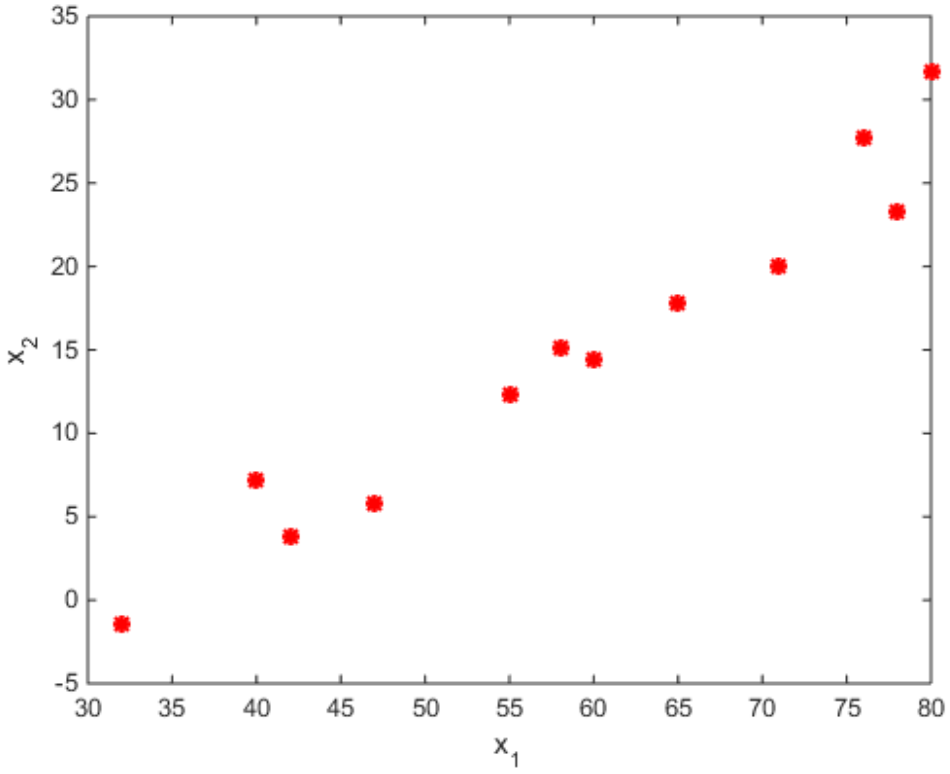
Why PCA?

- 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 **90** 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?

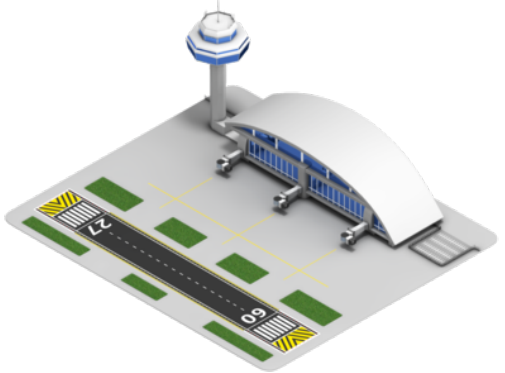
- **The true dimensionality of data is often lower than the observed dimensionality**.
- Example: You get a data set from a weather center that's collected over a period of 12 months at a particular region.
  - Data has the following format : (x_1, x_2)
  - What's the dimension of the data?
  - It turns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

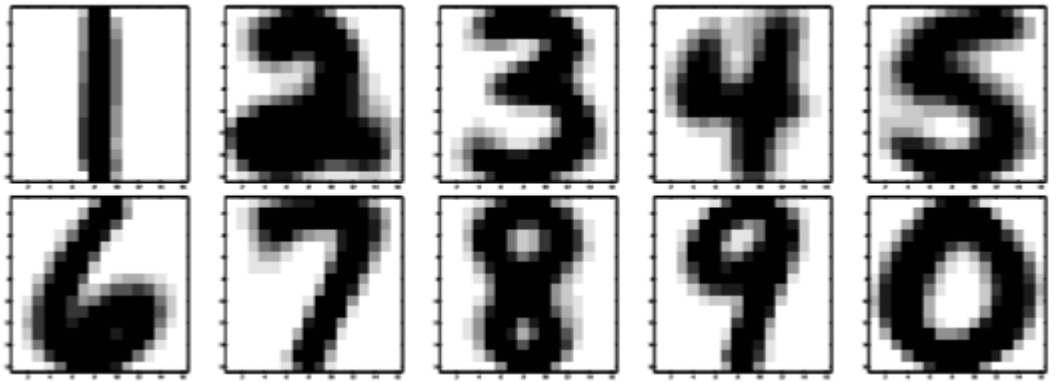
- Although, at the surface these all seem like different attributes, there's a single factor that can explain lots of these observations: temperature!
- **A machine learning algorithm should look for the single variable that counts for the others rather than looking at every individual one.**



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.**

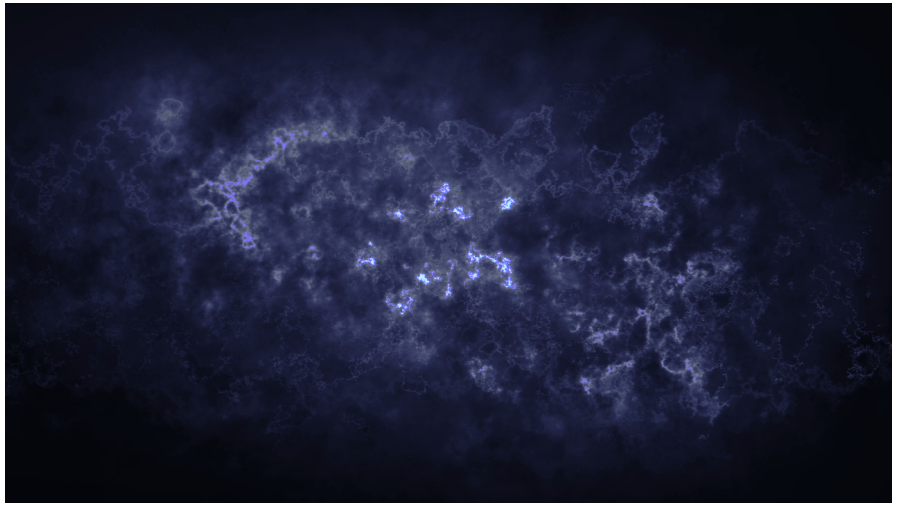


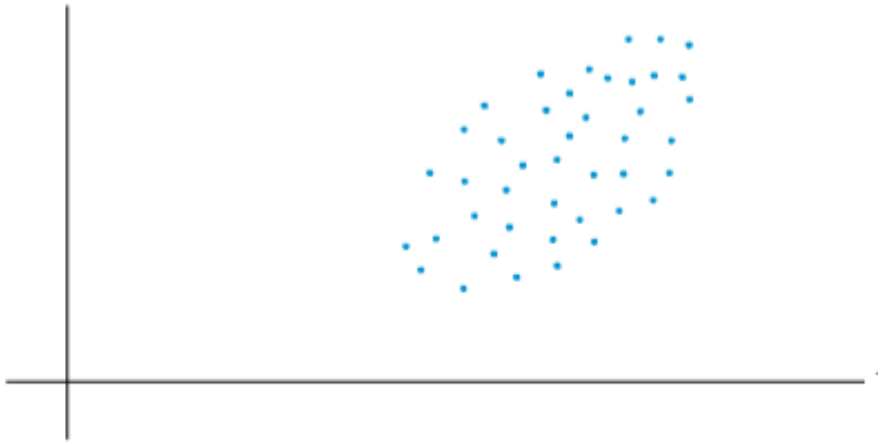
How to Reduce Dimensionality?

- 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

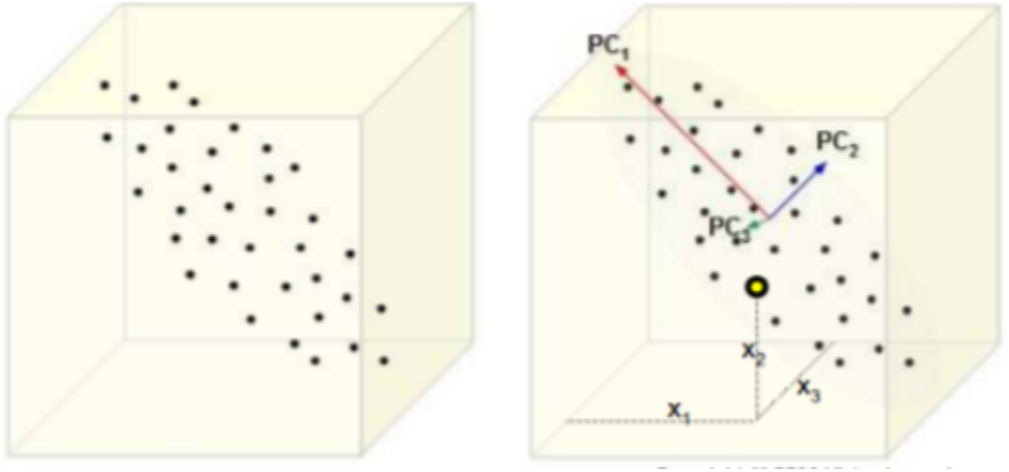
- In a dataset with k numeric attributes, you can visualize the data as a cloud of points in k-dimensional space:
  - the stars in the galaxy,
  - a swarm of flies frozen in time,
  - a two-dimensional scatter plot on paper.
- The attributes represent the coordinates of the space.
- But the axes you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

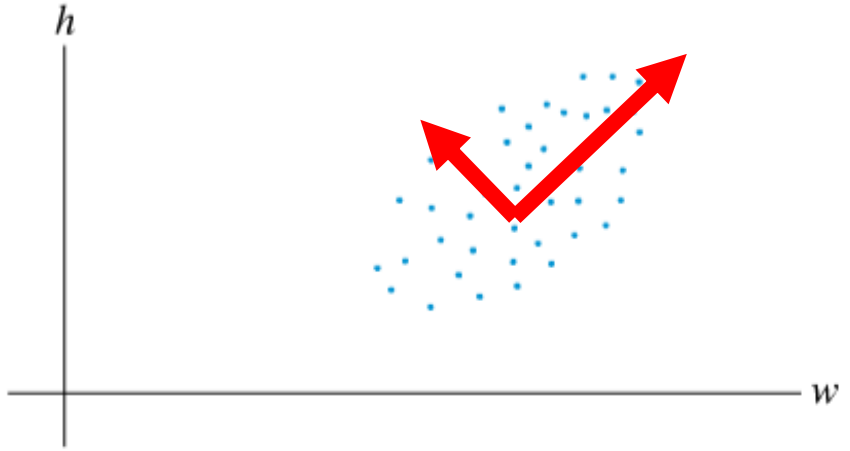


Example

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

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

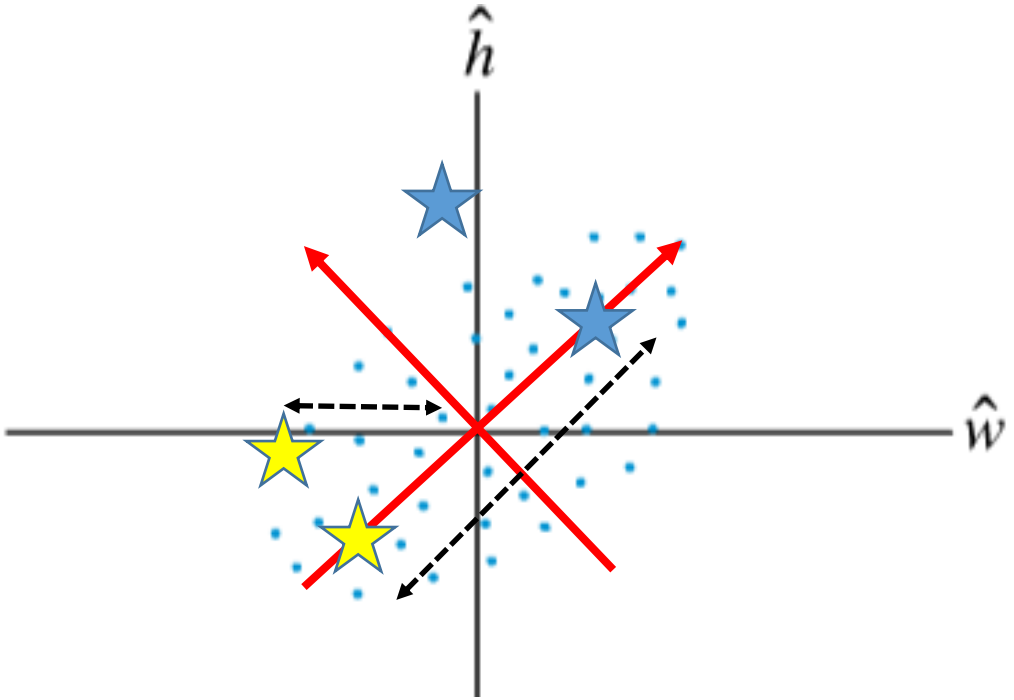
$$\mathbf{X}_1 \quad \mathbf{X}_2 \quad \quad \mathbf{X}_N$$

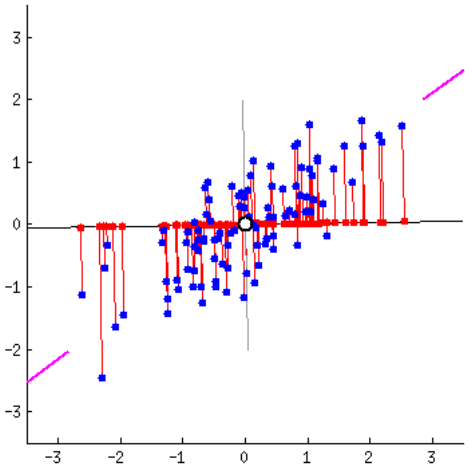


- Note that in two dimensions you only have one choice for the second axis. Its direction is determined by the first axis
- However, in three dimensions it can lie anywhere in the plane perpendicular to the first axis, and in higher dimensions there are even more choices, although it is always constrained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





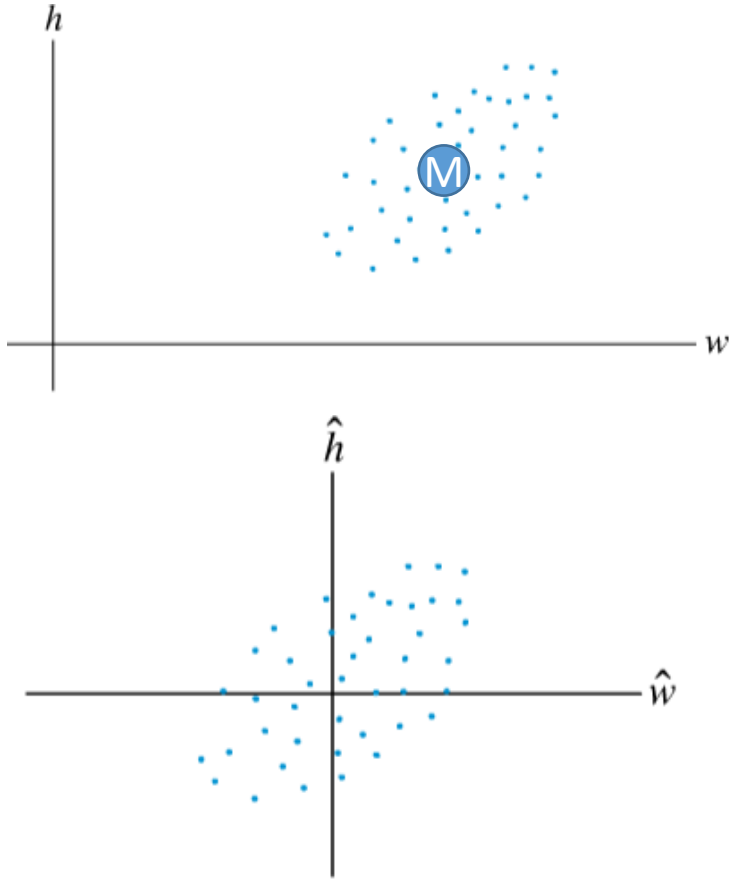
How to Get the Principal Components?

- 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:

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

$$\uparrow \quad \uparrow \qquad \uparrow$$

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



- 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
```

```
> B = sweep(x,1,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 $\begin{bmatrix} var(x_1) & \cdots & cov(x_1, x_d) \\ \vdots & \ddots & \vdots \\ cov(x_d, x_1) & \cdots & var(x_d) \end{bmatrix}$
- The main diagonal of the covariance matrix, contains the variances. e.g. $var(x_1)$ 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

- $sample\ cov(x_1, x_2) = \frac{1}{N-1}\sum_{i=1}^{N}(x_{1i} - m_1)(x_{2i} - m_2)$
- Since we already centered the data, m_1 and m_2 are zero
- $sample\ cov(x_1, x_2) = \frac{1}{N-1}\sum_{i=1}^{N} x_{1i}x_{2i}$
- 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:

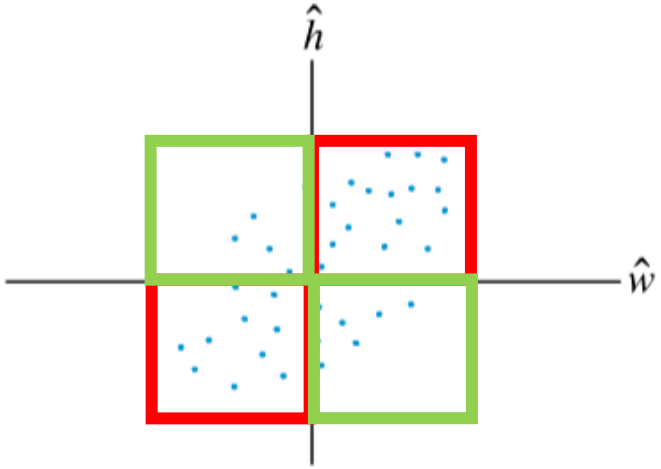
$$s = \frac{1}{N-1}BB^T$$

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 of the points in this diagram contribute positively or negatively to the covariance?
- Points in the first and third quadrants vary together (they're both above or below the mean simultaneously). Points in the second and forth quadrants vary 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:

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

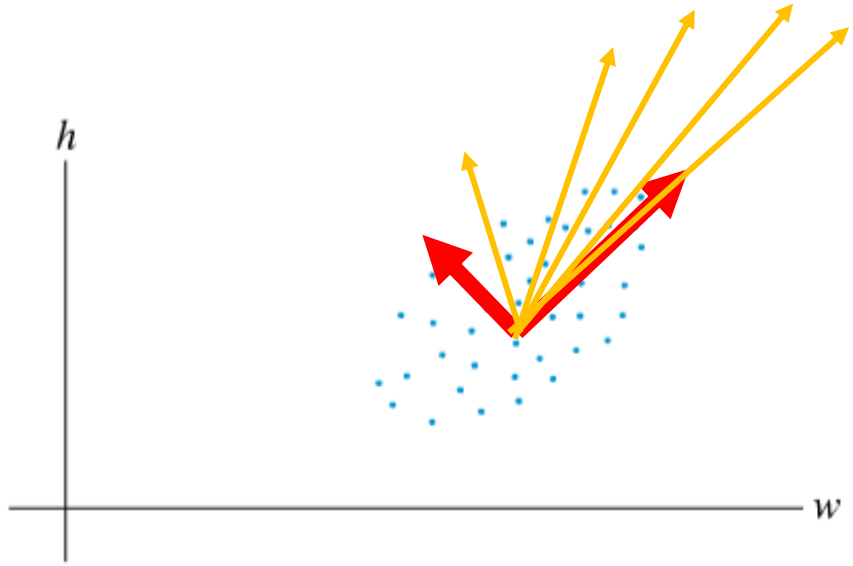
- Determine the covariance matrix.

```
> 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 vector 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.

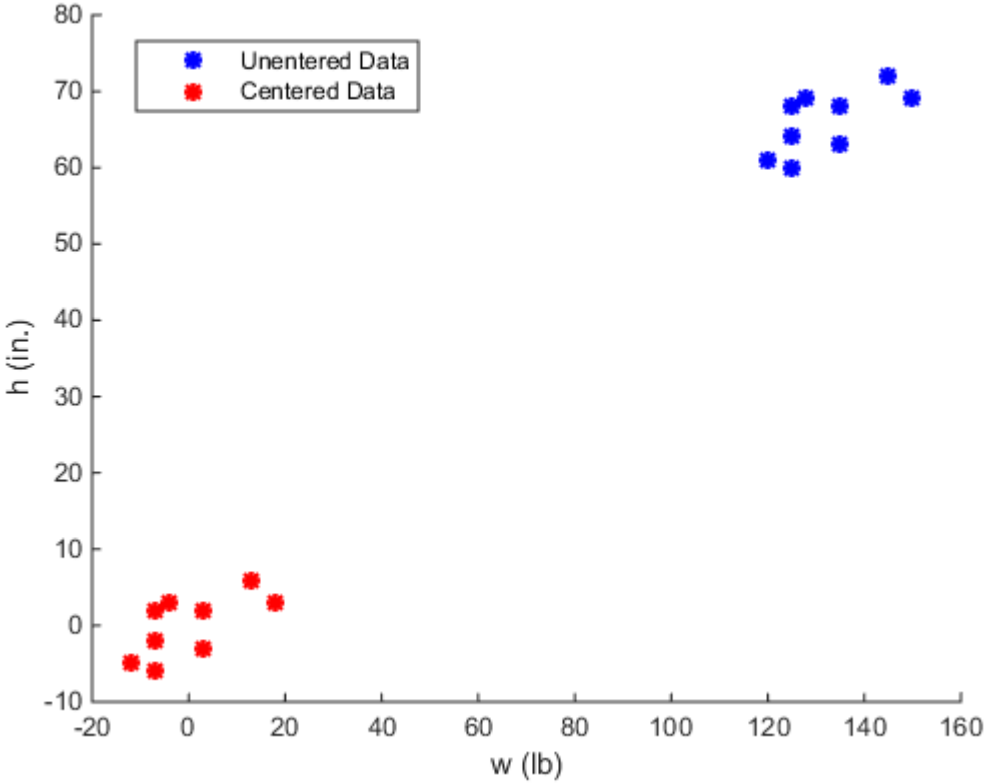
Example

- 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
```

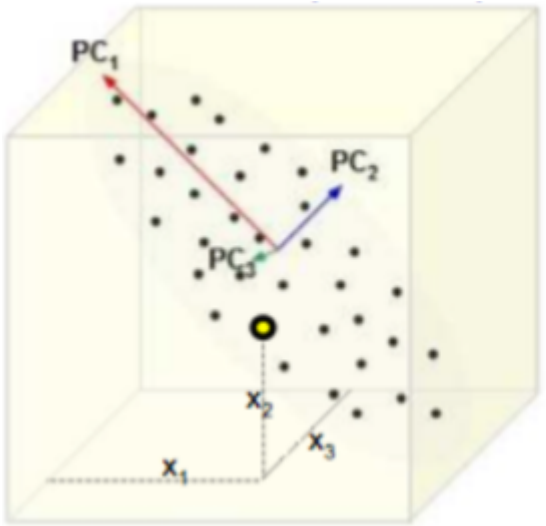
```
> a1 = matrix(1, nrow = 2)
> a1
     [,1]
[1,]    1
[2,]    1
> a2 = S%*%a1
> a2 = a2/norm(a2,'2')
> a2
       [,1]
w 0.9460978
h 0.3238811
> a3 = S%*%a2
> a3 = a3/norm(a3,'2')
> a3
       [,1]
w 0.9585130
h 0.2850487
> a4 = S%*%a3
> a4 = a4/norm(a4,'2')
> a4
       [,1]
w 0.9594534
h 0.2818675
>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

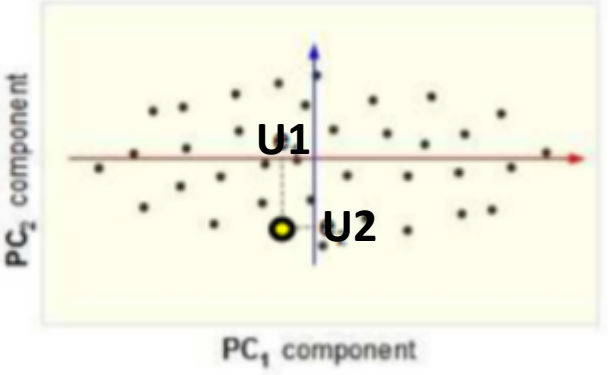
```
> eigen(S)
$values
[1] 110.283477   8.966523

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

Coordinates in the New System

- Assume that **after centering** the data points, the yellow point in the figure has coordinate $X_1 = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ (left figure).
- The new dimension is denoted by $Y_1 = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ (right figure).
- Assume that the first two principal components are vectors U_1 and U_2 (red and blue arrows).
- To find y_1 we need to project vector X_1 onto U_1. (What's the dimension of U_1?)
- $y_1 = P_{U_1}^{X_1} = U_1.X_1 = U_1^T X_1$
- Similarly: $y_2 = P_{U_2}^{X_1} = U_2.X_1 = U_2^T X_1$
- Let $P = \begin{bmatrix} \uparrow & \uparrow \\ U_1 & U_2 \\ \downarrow & \downarrow \end{bmatrix}$ where P is a 3×2 matrix.
- $\boldsymbol{Y_1} = \begin{bmatrix} U_1^T X_1 \\ U_2^T X_1 \end{bmatrix} = \begin{bmatrix} \leftarrow & U_1^T & \rightarrow \\ \leftarrow & U_2^T & \rightarrow \end{bmatrix} X_1 = \boldsymbol{P^T X_1}$





Projection of All Data Points in Matrix Form

- Assume that all data points are in the **mean deviation** form:
- $X = [X_1 \quad \dots \quad X_N]$ where X is a $d \times N$ matrix. (d is the dimension of the original data).
- Assume that P denotes the first m principal components:

$$P = [U_1 \quad \dots \quad U_m]$$

- The new set of coordinates can be found from:

$$\boldsymbol{Y = P^T X}$$

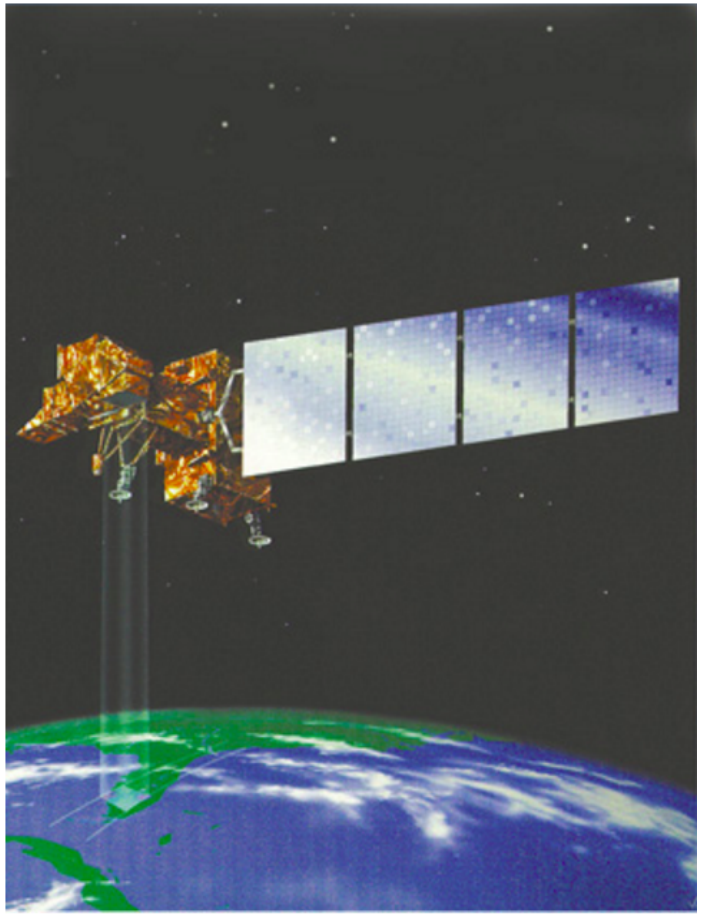
- What's the dimension of each element in the above expression?
  - X is $\mathrm{d} \times N$
  - Y is $m \times N$
  - P is $\mathrm{d} \times m$; P^{T} is $\mathrm{m} \times \mathrm{d}$

PCA: A Linear Combination

- Revisiting the expression for new coordinates, let $U_1 = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$ and $U_2 = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$
- Recall that $y_1 = P_{U_1}^{X_1} = U_1.X_1 = U_1^T X_1$ and $y_2 = P_{U_2}^{X_1} = U_2.X_1 = U_2^T X_1$
- We can write $y_1 = a_1x_1 + a_2x_2 + a_3x_3$ and $y_2 = c_1x_1 + c_2x_2 + c_3x_3$.
- Therefore, we can view the new dimension y as a linear combination of original coordinates where the weights are given by elements in the eigenvectors U_1 and U_2.

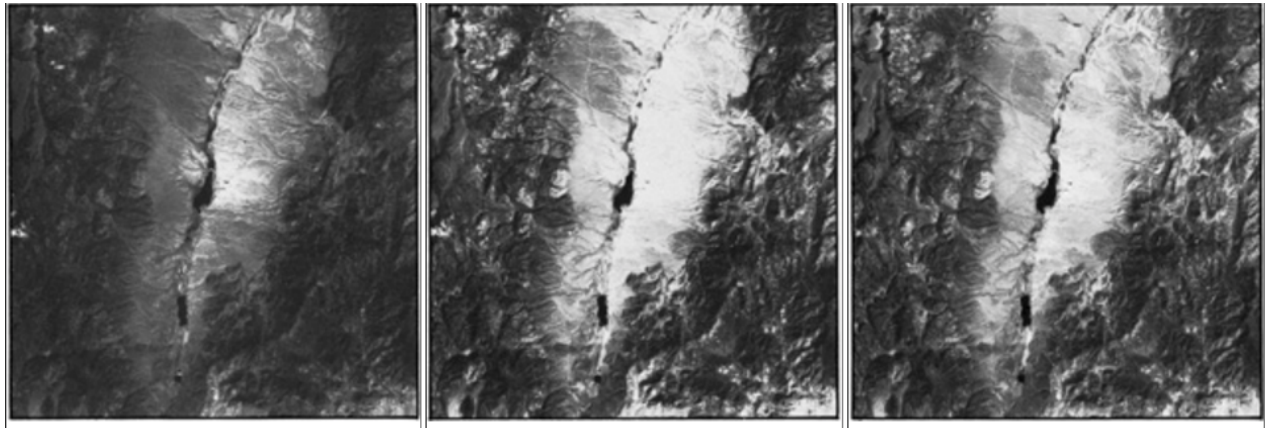
Example: Multichannel Image Processing

- The two Landsat satellites streak across the sky in near polar orbits, recording images of terrain and coastline.
- Every 16 days, each satellite passes over almost every square kilometer of the earth‘s surface, so any location can be monitored every 8 days.
- The Landsat images are useful for many purposes:
  - Developers and urban planners use them to study the rate and direction of urban growth.
  - Governments can detect and assess damage from natural disasters, such as forest fires, lava flows, floods, and hurricanes.
  - Environmental agencies can identify pollution from smokestacks and measure water temperatures in lakes and rivers near power plants.



Example: Multichannel Image Processing

- Sensors aboard the satellite acquire seven simultaneous images of any region on earth to be studied.
- The sensors record energy from separate wavelength bands
  - three in the visible light spectrum
  - four in infrared and thermal bands.
- Each of the seven images is one channel of a multichannel or multispectral image.
- These photos are examples of multichannel images taken over Railroad Valley, Nevada.



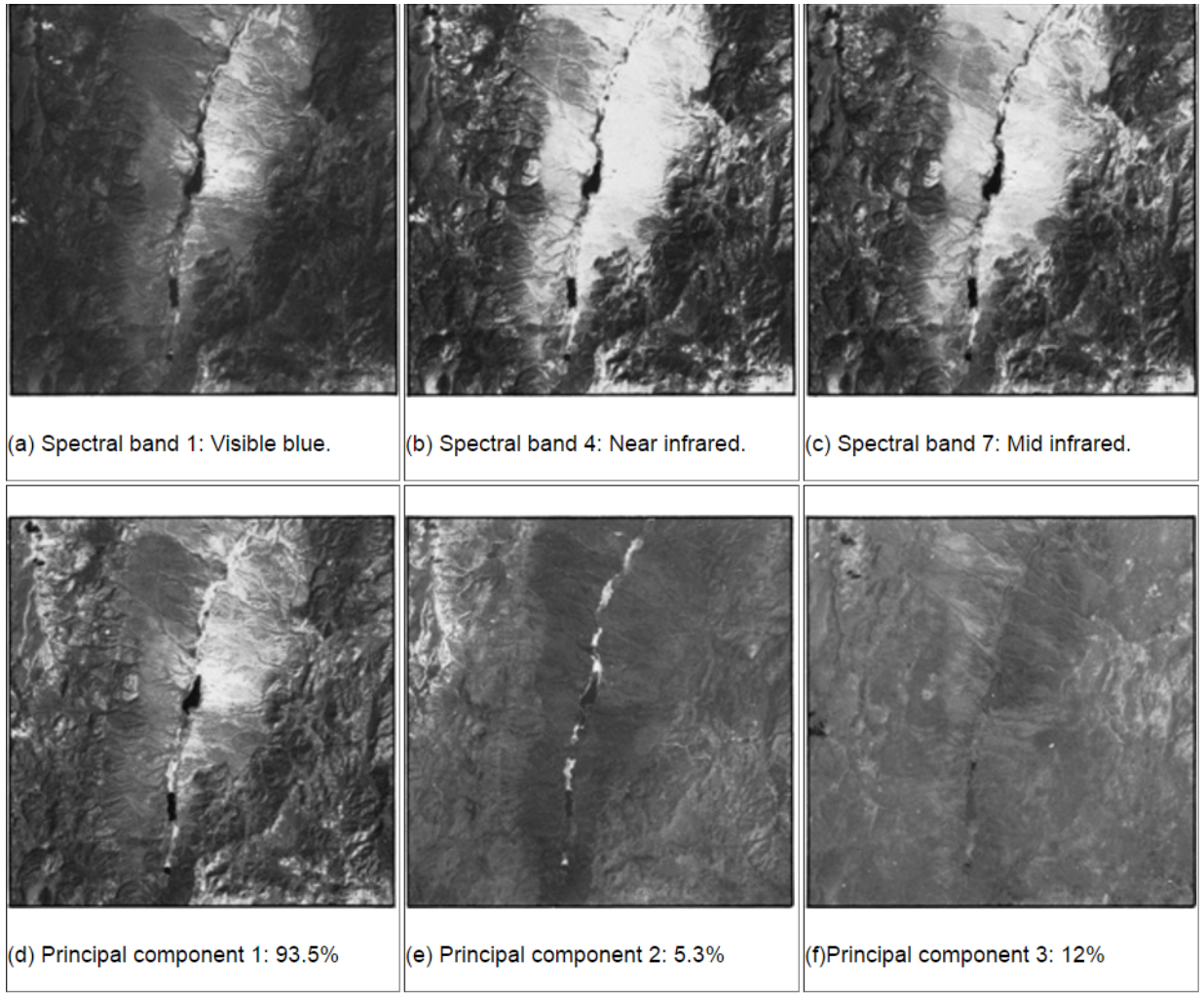
(a) Spectral band 1: Visible blue. (b) Spectral band 4: Near infrared. (c) Spectral band 7: Mid infrared.

Example: Multichannel Image Processing

- The seven Landsat images of one fixed region typically contain much redundant information, since some features will appear in several images.
- One goal of multichannel image processing is to view the data in a way that extracts information better than studying each image separately.
- PCA is an effective way to **suppress redundant information** and provide in only one or two composite images most of the information from the initial data.
- Roughly speaking, the goal is to find a special linear combination of the images, that is**, a list of weights that at each pixel combine all seven corresponding image values into one new value**.
- The weights are chosen in a way that makes the range of light intensities—the scene **variance**—in the composite image (called the first principal component) **greater** than that in any of the original images.

Example: Multichannel Image Processing

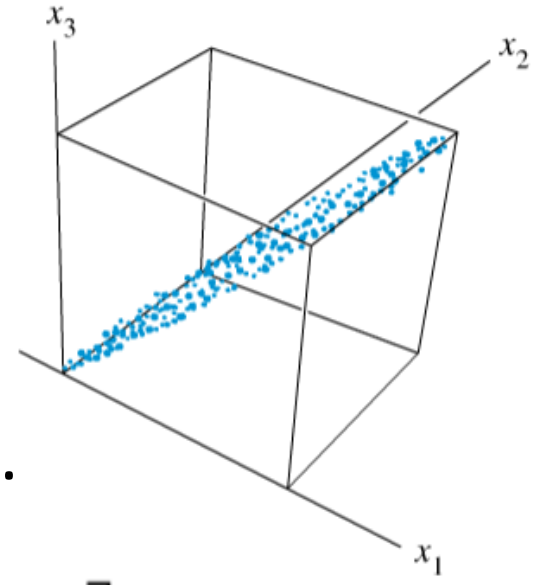
- Images from three Landsat spectral bands are shown in (a) – (c) .
- The total information in the three bands is rearranged in the three principal component images in (d) – (f) .
- The first component (d) explains 93.5% of the scene variance present in the initial data.
- In this way, the three-channel initial data have been reduced to one-channel data, with a loss in some sense of only 6.5% of the scene variance.



(a) Spectral band 1: Visible blue. (b) Spectral band 4: Near infrared. (c) Spectral band 7: Mid infrared.

(d) Principal component 1: 93.5% (e) Principal component 2: 5.3% (f)Principal component 3: 12%

Example: Multichannel Image Processing



- Typically, the image is 2000 × 2000 pixels, so there are 4 million pixels in the image.
- The data for the images in previous slide form a matrix with 3 rows and 4 million columns.
- The data can be visualized as a cluster of 4 million points in R^3.
- The associated covariance matrix is $S = \begin{bmatrix} 2382.78 & 2611.84 & 2136.20 \\ 2611.84 & 3106.47 & 2553.90 \\ 2136.20 & 2553.90 & 2650.71 \end{bmatrix}$
- Use PCA to reduce the 3-channel images to 1-channel image.
- Write the expression for new dimension as a linear combination of the original dimensions.

Solution

```
> s = matrix(c( 2382.78, 2611.84, 2136.20, 2611.84, 3106.47, 2553.90, 2136.20, 2553.90, 2650.7
1), nrow = 3)
> s
         [,1]    [,2]    [,3]
[1,] 2382.78 2611.84 2136.20
[2,] 2611.84 3106.47 2553.90
[3,] 2136.20 2553.90 2650.71
```

```
> eigen(s)
$values
[1] 7614.23008  427.62511   98.10481

$vectors
           [,1]       [,2]       [,3]
[1,] -0.5417295 -0.4893606  0.6834145
[2,] -0.6294758 -0.3026230 -0.7156672
[3,] -0.5570363  0.8178909  0.1441008
```

$$\lambda_1 = 7614.23 \qquad \lambda_2 = 427.63 \qquad \lambda_3 = 98.10$$

$$\mathbf{u}_1 = \begin{bmatrix} .5417 \\ .6295 \\ .5570 \end{bmatrix} \quad \mathbf{u}_2 = \begin{bmatrix} -.4894 \\ -.3026 \\ .8179 \end{bmatrix} \quad \mathbf{u}_3 = \begin{bmatrix} .6834 \\ -.7157 \\ .1441 \end{bmatrix}$$

$$y_1 = .54x_1 + .63x_2 + .56x_3$$

Example

- Given the following data set, find PCA and transform the data points. Find the total variance (sum of variances in each dimension) in the old and new data sets.

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

Solution

```
> d = matrix(c(1,2,1,4,2,13,7,8,1,8,4,5), nrow = 3)
> d
     [,1] [,2] [,3] [,4]
[1,]    1    4    7    8
[2,]    2    2    8    4
[3,]    1   13    1    5
> m = rowMeans(d)
> m
[1] 5 4 5
> B = sweep(d,1,m)
> B
     [,1] [,2] [,3] [,4]
[1,]   -4   -1    2    3
[2,]   -2   -2    4    0
[3,]   -4    8   -4    0
> Sx = (1/(ncol(d)-1))*B%*%t(B)
> Sx
     [,1] [,2] [,3]
[1,]   10    6    0
[2,]    6    8   -8
[3,]    0   -8   32
> library(psych)
> tr(Sx)
[1] 50
```

```
> l = eigen(Sx)$values
> l
[1] 34.551325 13.842964  1.605711
> P = eigen(Sx)$vectors
> P
            [,1]      [,2]       [,3]
[1,] -0.07404999 0.8192675 -0.5686100
[2,] -0.30300421 0.5247359  0.7955128
[3,]  0.95010791 0.2311989  0.2093848
```

```
> Y = t(P)%*%B
> Y
           [,1]       [,2]      [,3]       [,4]
[1,] -2.8982233  8.2809217 -5.160548 -0.222150
[2,] -5.2513377 -0.0191478  2.812683  2.457802
[3,] -0.1541247  0.6526629  1.207292 -1.705830
```

```
> Sy = (1/(ncol(d)-1))*Y%*%t(Y)
> round(Sy, 3)
       [,1]   [,2]  [,3]
[1,] 34.551  0.000 0.000
[2,]  0.000 13.843 0.000
[3,]  0.000  0.000 1.606
> tr(Sy)
[1] 50
```

Covariance of the Transformed Data

- $Y = P^T X$
- $S_Y = \frac{1}{N-1} YY^T$
- $S_Y = \frac{1}{N-1} P^T X (P^T X)^T$
- $S_Y = \frac{1}{N-1} P^T X X^T P = P^T \frac{XX^T}{N-1} P$
- $\boldsymbol{S_Y = P^T S_X P}$

- Recall if A is a square diagonalizable matrix, it can be written as $A = PDP^{-1}$ where D is a diagonal eigenvalue matrix and P is the eigenvector matrix.
- $A = PDP^{-1}$
- Right multiply above by P and left multiply by P^{-1}:
- $P^{-1}AP = D$
- We'll prove next week that if A is symmetric its eigenvectors are orthogonal AND if a matrix is orthogonal then its transpose and inverse are the same.
- Since S_x is symmetric, then P is orthogonal and $P^{-1} = P^T$.
- Therefore $P^{-1}AP = D$ can be written as $P^T S_x P = S_y$
- Therefore, S_y is a diagonal matrix **=> New dimensions are uncorrelated.**
- **The variances on the main diagonal on S_y are the eigenvalues of S_x.**

Reducing the Dimension of Data

- It can be shown that an orthogonal change of variables, $Y = P^T X$ does not change the total variance of the data.

$$\begin{Bmatrix}\text{total variance} \\ \text{of } x_1, \ldots, x_p\end{Bmatrix} = \begin{Bmatrix}\text{total variance} \\ \text{of } y_1, \ldots, y_p\end{Bmatrix} = \lambda_1 + \cdots + \lambda_p$$

- Therefore, the quotient $\frac{\lambda_j}{tr(S_Y)}$ measures the fraction of the total variance that is “explained” or “captured” by y_j .
- Therefore, PCA is potentially valuable for applications in which most of the variation, or dynamic range, in the data is due to variations in only a few of the new variables, $y_1, \ldots, y_p$

Example

- Compute the various percentages of variance of the Railroad Valley multispectral data that were displayed in the previous example.

```
> s = matrix(c( 2382.78, 2611.84, 2136.20, 2611.84, 3106.47, 2553.90, 2136.20, 2553.90, 2650.7
1), nrow = 3)
> s
        [,1]    [,2]    [,3]
[1,] 2382.78 2611.84 2136.20
[2,] 2611.84 3106.47 2553.90
[3,] 2136.20 2553.90 2650.71
```

```
> var = eigen(s)$values/sum(eigen(s)$values)
> var
[1] 0.93541370 0.05253405 0.01205225
```

```
> eigen(s)
$values
[1] 7614.23008  427.62511   98.10481

$vectors
           [,1]       [,2]       [,3]
[1,] -0.5417295 -0.4893606  0.6834145
[2,] -0.6294758 -0.3026230 -0.7156672
[3,] -0.5570363  0.8178909  0.1441008
```

First component

$$\frac{7614.23}{8139.96} = 93.5\%$$

Second component

$$\frac{427.63}{8139.96} = 5.3\%$$

Third component

$$\frac{98.10}{8139.96} = 1.2\%$$

Interpretation of Results

- The results show that 93.5% of the information collected for the Railroad Valley region is displayed in photograph 1, with 5.3% in 2 and only 1.2% remaining for 3.
- Therefore, data have practically no variance in the third (new) coordinate. (i.e. the values of y_3 are all close to zero.)
- Geometrically, the data points locations can be determined fairly accurately by knowing only the values of y_1 and y_2 .
- In fact, y_2 also has relatively small variance, which means that the points lie approximately along a line, and the data are essentially one-dimensional. (Think of a popsicle stick.)

References

- Linear Algebra and Its Application by David Lay