Northeastern University, Department of Mathematics

MATH G5110: Applied Linear Algebra and Matrix Analysis. (Fall 2020)

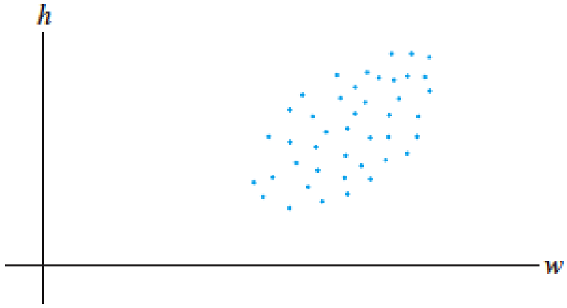
- Instructor: **He Wang** Email: **he.wang@northeastern.edu**

§14 Principle Component Analysis

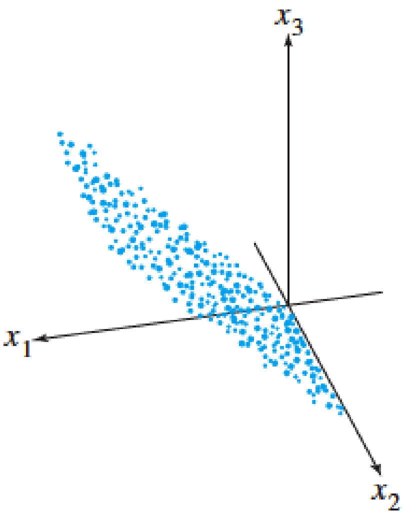
Principal component analysis is an effective way to suppress redundant information and provide in only one or two composite sections most of the information from the initial data.

Principal component analysis is used to analyze multivariate data: a sequence of (observations) vectors in $\mathbb{R}^n$.

Example: The two-dimensional data is given by a set of **weights** and **heights** of n college students. Let $\vec{x}_i$ denote the observation vector in $\mathbb{R}^2$ that lists the weight and height of the i-th student.



Example: Typically, the image is 1000 × 1000 pixels, so there are 1 million pixels in the image. The data for the image form a matrix with 3 rows and 1 million columns. In this case, the multidimensional character of the data refers to the three spectral dimensions.



Mean and Covariance Let $[X_1 \; ... \; X_n]$, be a $p \times n$ matrix of observations, such as described above.

The **sample mean** of the observation vectors

The sample mean of the observation vectors is the point in the center of the scatter plot. The **mean-deviation form** is

The **sample covariance matrix** of the observation vectors is the $p \times p$ matrix S

Denote the coordinates of X by $x_1, ..., x_p$. The diagonal entry s_{ii} in S is the **variance** of x_i. The variance of x_i measures the spread of the values of X_i.

The entry s_{ij} in S for $i \neq j$ is called the **covariance** of x_i and x_j.

If $s_{ij} = 0$, x_i and x_j are called **uncorrelated**.

The **total variance** of the data is the sum of the variances on the diagonal of S, i.e., the trace of S.

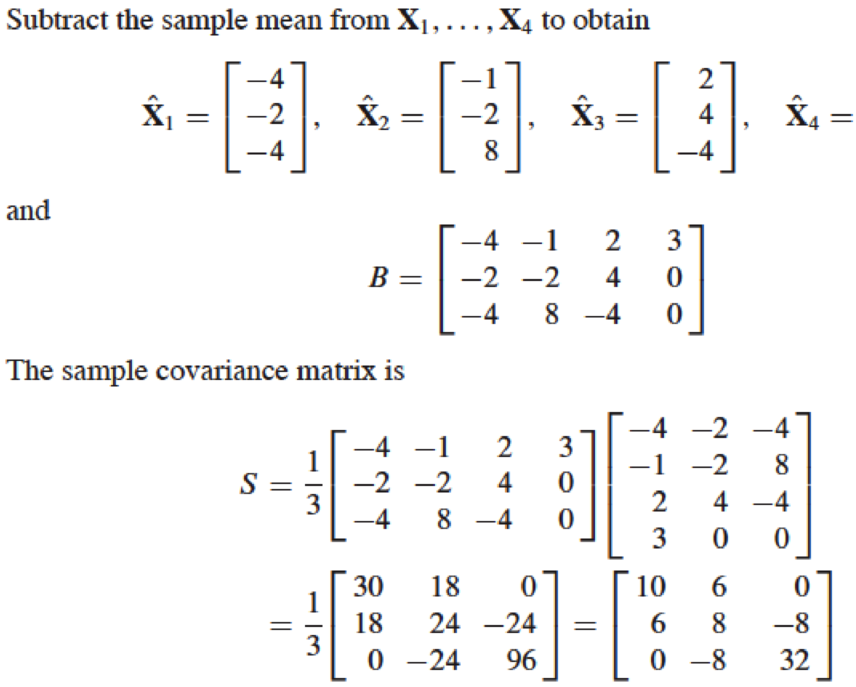
Example 1. Three measurements are made on each of four individuals in a random sample from a population. The observation vectors are

$$X_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}, \quad X_2 = \begin{bmatrix} 4 \\ 2 \\ 13 \end{bmatrix}, \quad X_3 = \begin{bmatrix} 7 \\ 8 \\ 1 \end{bmatrix}, \quad X_4 = \begin{bmatrix} 8 \\ 4 \\ 5 \end{bmatrix}$$

Compute the sample mean and the covariance matrix.

Sample mean is

$$\mathbf{M} = \frac{1}{4}\left(\begin{bmatrix}1\\2\\1\end{bmatrix} + \begin{bmatrix}4\\2\\13\end{bmatrix} + \begin{bmatrix}7\\8\\1\end{bmatrix} + \begin{bmatrix}8\\4\\5\end{bmatrix}\right) = \frac{1}{4}\begin{bmatrix}20\\16\\20\end{bmatrix} = \begin{bmatrix}5\\4\\5\end{bmatrix}$$



Subtract the sample mean from $\mathbf{X}_1, \ldots, \mathbf{X}_4$ to obtain

$$\hat{\mathbf{X}}_1 = \begin{bmatrix}-4\\-2\\-4\end{bmatrix}, \quad \hat{\mathbf{X}}_2 = \begin{bmatrix}-1\\-2\\8\end{bmatrix}, \quad \hat{\mathbf{X}}_3 = \begin{bmatrix}2\\4\\-4\end{bmatrix}, \quad \hat{\mathbf{X}}_4 = \begin{bmatrix}3\\0\\0\end{bmatrix}$$

and

$$B = \begin{bmatrix}-4 & -1 & 2 & 3\\ -2 & -2 & 4 & 0\\ -4 & 8 & -4 & 0\end{bmatrix}$$

The sample covariance matrix is

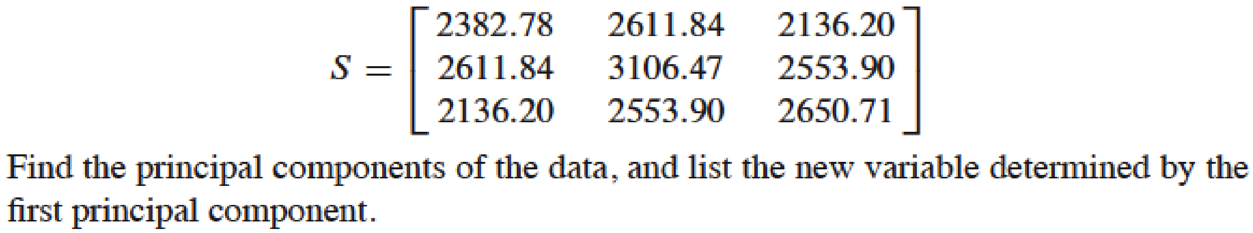
$$S = \frac{1}{3}\begin{bmatrix}-4 & -1 & 2 & 3\\ -2 & -2 & 4 & 0\\ -4 & 8 & -4 & 0\end{bmatrix}\begin{bmatrix}-4 & -2 & -4\\ -1 & -2 & 8\\ 2 & 4 & -4\\ 3 & 0 & 0\end{bmatrix}$$

$$= \frac{1}{3}\begin{bmatrix}30 & 18 & 0\\ 18 & 24 & -24\\ 0 & -24 & 96\end{bmatrix} = \begin{bmatrix}10 & 6 & 0\\ 6 & 8 & -8\\ 0 & -8 & 32\end{bmatrix}$$

Principal Component Analysis For simplicity, assume that the matrix $[X_1 \ldots X_n]$ is already in mean-deviation form.

The goal of principal component analysis is to find an orthogonal $p \times p$ matrix $P = [\vec{u}_1 \ldots \vec{u}_p]$ that determines a change of variable, $X = PY$, with the property that the new variables $\vec{y}_1, \ldots, \vec{y}_p$ are uncorrelated and are arranged in order of decreasing variance.

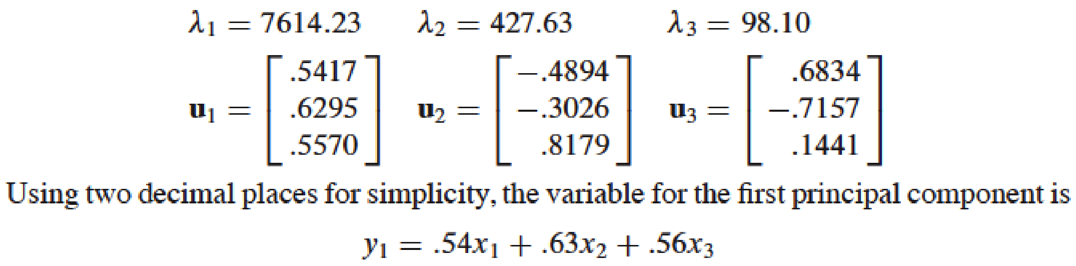
Example: Suppose associated covariance matrix of a multispectral image is given by



$$S = \begin{bmatrix} 2382.78 & 2611.84 & 2136.20 \\ 2611.84 & 3106.47 & 2553.90 \\ 2136.20 & 2553.90 & 2650.71 \end{bmatrix}$$

Find the principal components of the data, and list the new variable determined by the first principal component.

The eigenvalues of S and the associated principal components (the unit eigenvectors) are



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

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

Using two decimal places for simplicity, the variable for the first principal component is

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

Reducing the Dimension of Multivariate Data

Principal component analysis 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, $\vec{y}_1, ..., \vec{y}_p$.

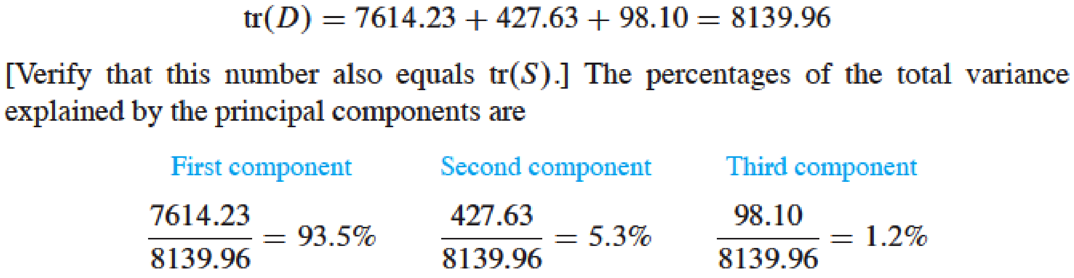
An orthogonal change of variables, $X = PY$ does not change the total variance of the data.

$$\text{Total variance of } x_1, ..., x_p = \text{Total variance of } y_1, ..., y_p = tr(D) {=} \lambda_1 + \cdots + \lambda_p.$$

The variance of y_i is λ_i and the quotient $\lambda_i / tr(S)$ measures the fraction of the total variance that is captured by y_i.

Example Compute the various percentages of variance of in the previous example.

The total variance of the data is



$$\text{tr}(D) = 7614.23 + 427.63 + 98.10 = 8139.96$$

[Verify that this number also equals tr(S).] The percentages of the total variance explained by the principal components are

First component	Second component	Third component
$\dfrac{7614.23}{8139.96} = 93.5\%$	$\dfrac{427.63}{8139.96} = 5.3\%$	$\dfrac{98.10}{8139.96} = 1.2\%$

Example The following table lists the weights and heights of five boys:

Boy	#1	#2	#3	#4	#5
Weight (lb)	120	125	125	135	145
Height (in.)	61	60	64	68	72

1. Find the covariance matrix for the data.
2. Make a principal component analysis of the data to find a single *size index* that explains most of the variation in the data.

1. First arrange the data in mean-deviation form. The sample mean vector is easily seen to be $\mathbf{M} = \begin{bmatrix} 130 \\ 65 \end{bmatrix}$. Subtract $\mathbf{M}$ from the observation vectors (the columns in the table) and obtain

$$B = \begin{bmatrix} -10 & -5 & -5 & 5 & 15 \\ -4 & -5 & -1 & 3 & 7 \end{bmatrix}$$

Then the sample covariance matrix is

$$S = \frac{1}{5-1}\begin{bmatrix} -10 & -5 & -5 & 5 & 15 \\ -4 & -5 & -1 & 3 & 7 \end{bmatrix}\begin{bmatrix} -10 & -4 \\ -5 & -5 \\ -5 & -1 \\ 5 & 3 \\ 15 & 7 \end{bmatrix}$$

$$= \frac{1}{4}\begin{bmatrix} 400 & 190 \\ 190 & 100 \end{bmatrix} = \begin{bmatrix} 100.0 & 47.5 \\ 47.5 & 25.0 \end{bmatrix}$$

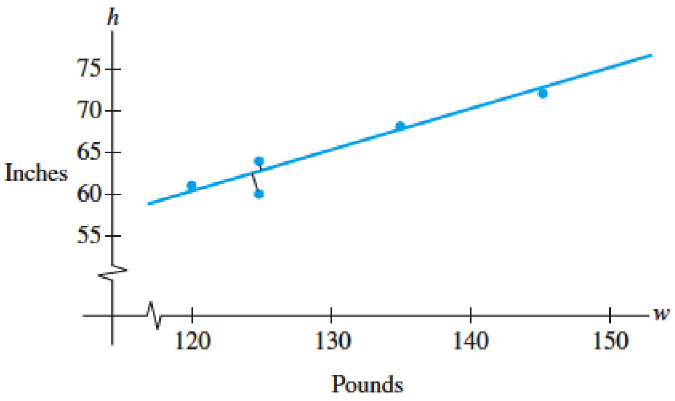
2. The eigenvalues of S are (to two decimal places)

$$\lambda_1 = 123.02 \quad \text{and} \quad \lambda_2 = 1.98$$

The unit eigenvector corresponding to λ_1 is $\mathbf{u} = \begin{bmatrix} .900 \\ .436 \end{bmatrix}$. (Since S is 2×2, the computations can be done by hand if a matrix program is not available.) For the *size index*, set

$$y = .900\hat{w} + .436\hat{h}$$

where $\hat{w}$ and $\hat{h}$ are weight and height, respectively, in mean-deviation form. The variance of this index over the data set is 123.02. Because the total variance is $\operatorname{tr}(S) = 100 + 25 = 125$, the size index accounts for practically all (98.4%) of the variance of the data.



Lab1.

Data from an old census in the Madison, Wisconsin area provided information on five socioeconomic variables for a collection of 14 neighborhoods. The variables are

- total population (thousands)
- median school years
- total employment (thousands)
- health services employment (hundreds)
- median home values ($10,000s)

The data is summarized by the mean vector and the covariance matrix of the variables. Namely let $\vec{x}_1, \ldots, \vec{x}_{14} \in \mathbb{R}^5$ be the data samples from the 14 neighborhoods, then the sample mean is

$$\vec{m} = \frac{1}{14} \sum_{i=1}^{14} \vec{x}_i$$

and the sample covariance is

$$S = \frac{1}{13} \sum_{i=1}^{14} (\vec{x}_i - \vec{m})\ (\vec{x}_i - \vec{m})^T$$

The data are not provided, but we do have the mean and sample covariance, which are

$$(\vec{m})^T = (4.32 \quad 14.01 \quad 1.95 \quad 2.17 \quad 2.45)$$

$$S = \begin{pmatrix} 4.308 & 1.683 & 1.803 & 2.155 & -0.253 \\ 1.683 & 1.768 & 0.588 & 0.177 & 0.176 \\ 1.803 & 0.588 & 0.801 & 1.065 & -0.158 \\ 2.155 & 0.177 & 1.065 & 1.97 & -0.357 \\ -0.253 & 0.176 & -0.158 & -0.357 & 0.504 \end{pmatrix}$$

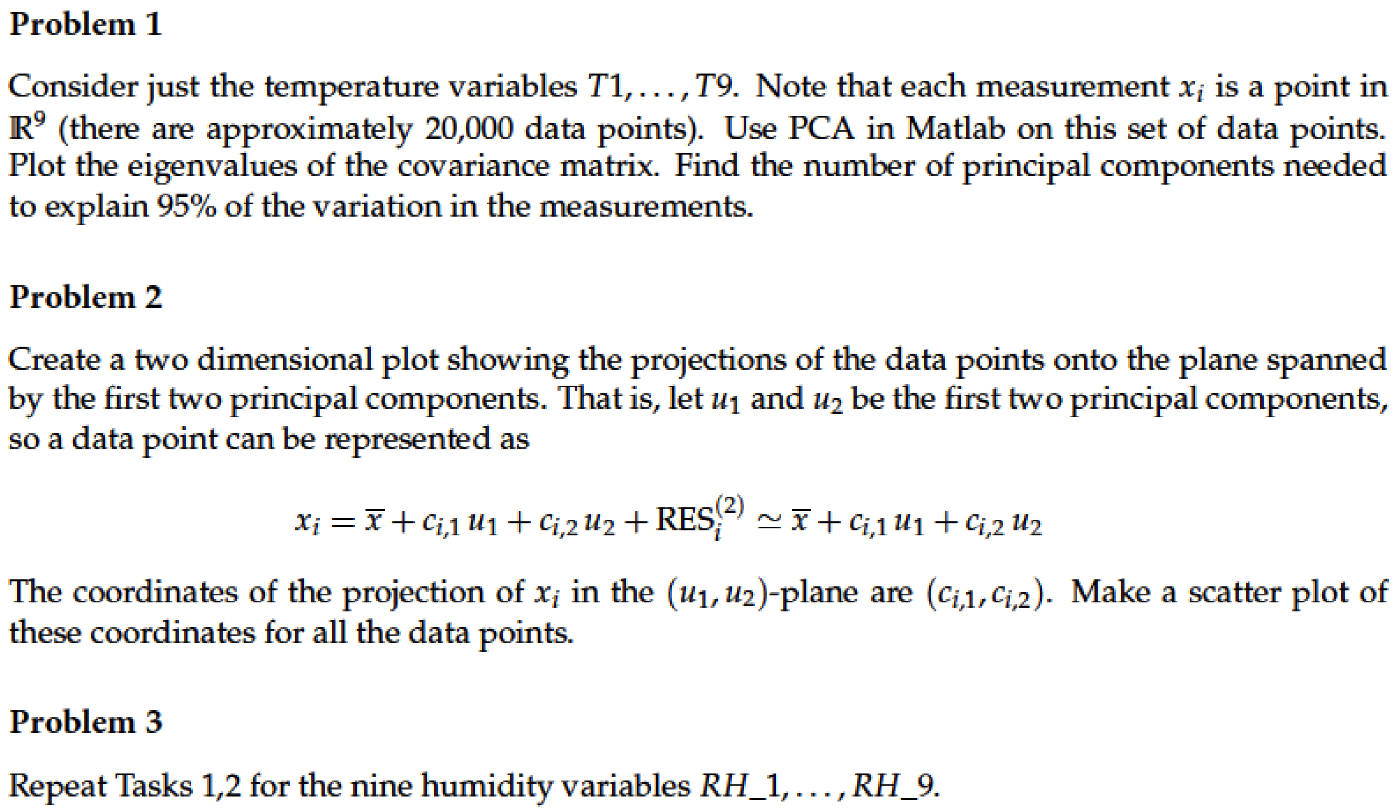
(1) Find the principal components for this data. Plot the eigenvalues of the covariance matrix in decreasing order. How many components are needed to explain 95% of the total sample variance?

(2) Let $\vec{z}_1, ..., \vec{z}_{14} \in \mathbb{R}^5$ be the projections of the data points onto the first two principal components. Compute the sample mean and the sample covariance of the points $\vec{z}_1, ..., \vec{z}_{14} \in \mathbb{R}^5$.

Lab2.

The file energydata_complete.csv contains measurement data of temperature and humidity: you can read about the meaning of the variables here

https://archive.ics.uci.edu/ml/datasets/Appliances+energy+prediction



Problem 1

Consider just the temperature variables $T1, \ldots, T9$. Note that each measurement x_i is a point in $\mathbb{R}^9$ (there are approximately 20,000 data points). Use PCA in Matlab on this set of data points. Plot the eigenvalues of the covariance matrix. Find the number of principal components needed to explain 95% of the variation in the measurements.

Problem 2

Create a two dimensional plot showing the projections of the data points onto the plane spanned by the first two principal components. That is, let u_1 and u_2 be the first two principal components, so a data point can be represented as

$$x_i = \overline{x} + c_{i,1}\, u_1 + c_{i,2}\, u_2 + \mathrm{RES}_i^{(2)} \simeq \overline{x} + c_{i,1}\, u_1 + c_{i,2}\, u_2$$

The coordinates of the projection of x_i in the (u_1, u_2)-plane are $(c_{i,1}, c_{i,2})$. Make a scatter plot of these coordinates for all the data points.

Problem 3

Repeat Tasks 1,2 for the nine humidity variables $RH\_1, \ldots, RH\_9$.