Topics
- Principal component analysis
- Applications to data compression

This lecture is mostly based on LAA 7.5. Additional reading can be found in ALA 8.8

Principal Component Analysis with Applications to Image Processing and Statistics

We start with a motivating application from satellite imagery analysis. The Landsat satellites are a pair of imaging satellites that record images of terrain and coastline: these satellites cover almost every square mile of the earth's surface every 16 days.

Satellite sensors acquire seven simultaneous images of any given region, with each sensor recording energy from separate wavelength bands: three in the visible light spectrum and four in the infrared and thermal bands.

Each image is digitized and stored as a rectangular array of numbers, with each number indicating the signal intensity at the corresponding pixel. Each of the seven images is one channel of a multichannel or multispectral image.

The seven Landsat images of a given region typically contain a lot of redundant information, as some features will appear across most channels. However, other features, because of their color or temperature, may only appear in one or two channels. A goal of multichannel image processing is to view the data in a way that extracts information better than studying each image separately.

One approach, called Principal Component Analysis (PCA), seeks to find a special linear combination of the data that takes a weighted combination of all seven images into just one or two images. Importantly, we want these one or two composite images, or principal components, to capture as much of the scene variance (features) as possible; in particular, features should be more visible in the composite images than any of the original individual ones.

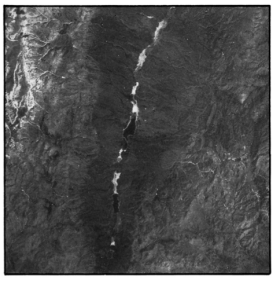
This idea, which we'll explore in detail today, is illustrated with some Landsat imagery taken over Railroad Valley, Nevada.

(a) Spectral band 1: Visible blue.

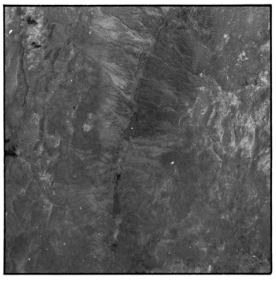
(b) Spectral band 4: Near infrared.

(c) Spectral band 7: Mid-infrared.

(d) Principal component 1: 93.5%.



(e) Principal component 2: 5.3%.



(f) Principal component 3: 1.2%.

Images from three Landsat spectral bands are shown in (a)-(c); the total information in these images is "rearranged" into the three principal components in (d)-(f). The first component, (d), "explains" 93.5% of the scene features (or variance) found in the original data. In this way, we could compress all of the original data to the single image (d) with only a 6.5% loss of scene variance.

PCA can in general be applied to any data that consists of lists of measurements made on a collection of objects or individuals, including data mining, machine learning, image processing, speech recognition, facial recognition, and health informatics. As we'll see next, the way in which these "special combinations" of measurements are computed are via the singular vectors of an observation matrix.

Observation Matrix, Mean, and Covariance

Let $\underline{x}_j \in \mathbb{R}^p$ denote an observation vector obtained from measurement j, and suppose that $j = 1, \ldots, N$ measurements are obtained. The observation matrix $X \in \mathbb{R}^{p \times N}$ is a $p \times N$ matrix with j^{th} column equal to the j^{th} measurement vector $\underline{x}_j$:

$$X = [\underline{x}_1 \ \underline{x}_2 \ \cdots \ \underline{x}_N] \in \mathbb{R}^{p \times N}$$

Example Suppose that $\underline{x}_j \in \mathbb{R}^2$ is a two dimensional data given by the weight and height of the j^{th} student at Penn: $\underline{x}_j = (w_j, h_j) \in \mathbb{R}^2$. Then if measurements are obtained from N students, the observation matrix $X \in \mathbb{R}^{2 \times N}$ has the form:

$$X = \begin{bmatrix} w_1 & w_2 & \cdots & w_N \\ h_1 & h_2 & & h_N \\ \underline{x}_1 & \underline{x}_2 & \cdots & \underline{x}_N \end{bmatrix}.$$

The set of observation vectors can be visualized as a two-dimensional scatter plot:

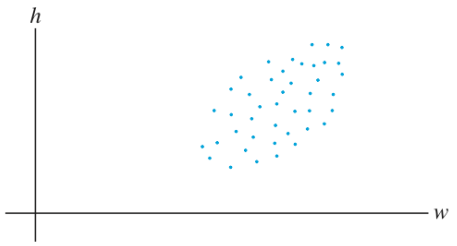


FIGURE 1 A scatter plot of observation vectors $\mathbf{X}_1, \ldots, \mathbf{X}_N$.

Example: The three images (a)-(c) above can be thought of as one image composed of three spectral components, as each image gives information about the same region. We can capture this mathematically by associating a vector $\in \mathbb{R}^3$ to each pixel (each small area of the image) that lists the intensity for that pixel in the three spectral bands. Typically the image is 2000 x 2000 pixels, so there are 4 million pixels in the image. The observation matrix for this data is a matrix with 3 rows and 4 million columns. The data can thus be

visualized as a scatter plot of 4 million points in $\mathbb{R}^3$ (see Figure below for a synthetic example).

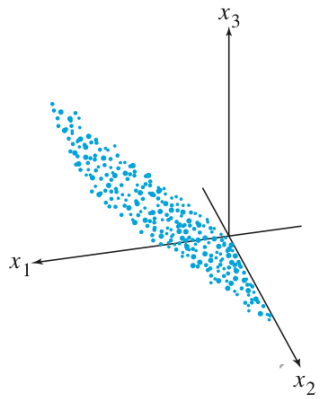


FIGURE 2
A scatter plot of spectral data for a satellite image.

Mean and Covariance

To understand PCA, we need to understand some basic concepts from Statistics. We will define the mean and covariance of a set of observations $x_1, \dots, x_N$. For our purposes, these will simply be things we can compute from the data, but you should be aware that these are well motivated quantities from a statistical perspective. You will learn more about this in ESE 3010, STAT 4300 or ESE 4020.

Let's start with an observation matrix $X \in \mathbb{R}^{p \times N}$, with columns $x_1, \dots, x_N \in \mathbb{R}^p$. The sample mean $\underline{m}$ of the observation vectors $x_1, \dots, x_N$ is given by

$$\underline{m} = \frac{1}{N}(x_1 + \cdots + x_N) = \sum_{j=1}^{N} x_j.$$

Another name for the sample mean is the centroid of the data, which we encountered when we learned about the k-means algorithm.

Since PCA is interested in directions of (maximal) variation in our data, it makes sense to subtract off the mean $\underline{m}$, as it captures the average behaviour of our data set. To that end, define the centered observations to be

$$\hat{x}_j = x_j - \underline{m}, \quad j = 1, \dots, N,$$

and the centered or de-meaned observation matrix

$$\hat{X} = [\hat{x}_1 \ \hat{x}_2 \ \cdots \ \hat{x}_N].$$

For example, Fig. 3 below shows a centered version of the weight/height data illustrated in Fig. 1:

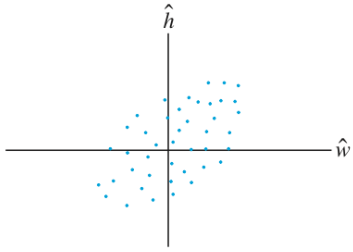


FIGURE 3
Weight–height data in mean-deviation form.

Finally, we define the sample covariance matrix $S \in \mathbb{R}^{p \times p}$ as

$$S = \frac{1}{N} \hat{X} \hat{X}^T.$$

Since any matrix of the for AA^T is positive semidefinite (can you see why?), so is S. Note sometimes $1/N-1$ is used as normalization; this is motivated by statistical considerations beyond the scope of this course (it leads to S being an unbiased estimator of the "true" covariance of the data). We will just use $\frac{1}{N}$.

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

$$\underline{x}_1 = \begin{bmatrix} 1 \\ 9 \\ 7 \end{bmatrix}, \quad \underline{x}_2 = \begin{bmatrix} 4 \\ 2 \\ 13 \end{bmatrix}, \quad \underline{x}_3 = \begin{bmatrix} 7 \\ 8 \\ 1 \end{bmatrix}, \quad \underline{x}_4 = \begin{bmatrix} 8 \\ 4 \\ 5 \end{bmatrix}.$$

The sample mean is $\underline{m} = \frac{1}{4}(\underline{x}_1 + \underline{x}_2 + \underline{x}_3 + \underline{x}_4) = \begin{bmatrix} 5 \\ 4 \\ 5 \end{bmatrix}$.

The centered observations $\hat{\underline{x}}_j = \underline{x}_j - \underline{m}$ are then

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

and the centered observation matrix is

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

The sample covariance matrix is

$$S = \frac{1}{4} \hat{X} \hat{X}^T = \begin{bmatrix} 15/2 & 9/2 & 0 \\ 9/2 & 6 & -6 \\ 0 & -6 & 24 \end{bmatrix}.$$

You might be wondering what the entries S_{ij} of the covariance matrix S_{mean}. Let's take a bit of a closer look. We'll consider a small example where the observations $x_j \in \mathbb{R}^2$ are two dimensional, and assume we have $N=3$ observations. Let the first measurement be $a \in \mathbb{R}$ and the second $b \in \mathbb{R}$, so that $x_j = (a_j, b_j) \in \mathbb{R}^2$ and the centered observation is $\hat{x}_j = (\hat{a}_j, \hat{b}_j) \in \mathbb{R}^2$. Our centered observation matrix is then

$$\hat{X} = \begin{bmatrix} \hat{a}_1 & \hat{a}_2 & \hat{a}_3 \\ \hat{b}_1 & \hat{b}_2 & \hat{b}_3 \end{bmatrix} = \begin{bmatrix} \hat{a}^T \\ \hat{b}^T \end{bmatrix},$$

where we defined $\hat{a} = (\hat{a}_1, \hat{a}_2, \hat{a}_3) \in \mathbb{R}^3$ and $\hat{b} = (\hat{b}_1, \hat{b}_2, \hat{b}_3)$ as the vectors in $\mathbb{R}^3$ containing all of the centered first and second measurements, respectively.

Then, we can write our sample covariance matrix as:

$$S = \frac{1}{3} \hat{X} \hat{X}^T = \frac{1}{3} \begin{bmatrix} \hat{a}^T \\ \hat{b}^T \end{bmatrix} \begin{bmatrix} \hat{a} & \hat{b} \end{bmatrix} = \begin{bmatrix} \frac{\|\hat{a}\|^2}{3} & \frac{\hat{a}^T \hat{b}}{3} \\ \frac{\hat{a}^T \hat{b}}{3} & \frac{\|\hat{b}\|^2}{3} \end{bmatrix}.$$

The diagonal entry $S_{11} = \frac{\|\hat{a}\|^2}{3}$ is called the variance of measurement 1.

Expanding it out:

$$S_{11} = \frac{\|\hat{a}\|^2}{3} = \frac{1}{3}\left(\hat{a}_1^2 + \hat{a}_2^2 + \hat{a}_3^2\right)$$
$$= \frac{1}{3}\left((a_1 - m_1)^2 + (a_2 - m_2)^2 + (a_3 - m_3)^2\right)$$

We see that S_{11} captures how much the first measurement a_j deviates from its mean value m_1, on average, i.e., it measures how much a_j varies relative to its mean. Similarly, $S_{22} = \frac{\|\hat{b}\|^2}{3}$ is the variance of measurement 2.

Now let's look at the off diagonal term $S_{12} = S_{21} = \frac{\hat{a}^T \hat{b}}{3}$. Recall from our work on inner products that $\hat{a}^T \cdot \hat{b} = \|\hat{a}\| \cdot \|\hat{b}\| \cos\theta$, where θ is the angle between $\hat{a}$ and $\hat{b}$. We can view

$$\cos\theta = \frac{\hat{a}^T \hat{b}}{\|\hat{a}\| \cdot \|\hat{b}\|}$$

as a measure of how well aligned, or correlated, if $\hat{a}$ and $\hat{b}$ are parallel, $\cos\theta = 1$ or -1, and if $\hat{a}$ and $\hat{b}$ are perpendicular, $\cos\theta = 0$. This lets us interpret $S_{12} = \frac{\hat{a}^T \hat{b}}{3}$, which is proportional to $\cos\theta$, as a measure of how similarly a and b deviate from their means: if $\hat{a}^T \hat{b}$ is positive, this means $\hat{a}$ and $\hat{b}$ tend to move up or down together; if it is negative, they tend to move in opposite directions, and if it is small (or zero), a and b tend to move independently of each other. Since S_{12} captures how the 1st and 2nd measurements vary with each other, it is called their covariance.

Finally, although we worked out these concepts for $\underline{x}_j \in \mathbb{R}^2$ and $j=1,2,3$, these concepts extend naturally to the general setting:

- S_{ii} = variance of measurement i across measurements $j=1,\dots,N$
- S_{kl} = covariance of measurements k and l across measurements $j=1,\dots,N$.

ONLINE NOTES: Please include examples of correlated, anticorrelated, and uncorrelated vectors, e.g., Fig 3.8 from UMLS.

Principal Component Analysis

To make our notation a little bit cleaner, we'll assume that our observations $\underline{x}_j \in \mathbb{R}^p$ and observation matrix $X \in \mathbb{R}^{p\times N}$ have already been centered so that $X = \tilde{X}$ and $\underline{m} = \underline{0}$.

The goal of PCA is to find a new orthogonal basis for $\mathbb{R}^p$ defined by the orthogonal $p\times p$ matrix $P = [\underline{u}_1 \cdots \underline{u}_p]$, for $\underline{u}_1, \dots, \underline{u}_p$ an orthonormal basis of $\mathbb{R}^p$:

$$\underline{x} = P\underline{y} = y_1\underline{u}_1 + y_2\underline{u}_2 + \cdots + y_p\underline{u}_p$$

with the property that the new coordinates $y_1, \dots, y_p$ are uncorrelated (i.e., have covariance 0) and arranged in decreasing order of variance. The matrix P is orthogonal, and hence $\underline{y} = P^{-1}\underline{x} = P^T\underline{x}$ provides a decomposition of the measurement $\underline{x}$ along the directions $\underline{u}_1, \dots, \underline{u}_p$, where most of the variance in $\underline{x}$ can be found in the direction $\underline{u}_1$, 2nd most in direction $\underline{u}_2$, etc.

What does the covariance matrix of the new variables $\underline{y}$ look like. Note that if $\underline{y}_j = P^T\underline{x}_j$, then $Y = P^TX$ is the observation matrix in our new basis, so that

$$S_y = YY^T = P^TXX^TP = P^TS_xP.$$

Footnote: we add subscripts X and Y to S_X and S_Y to highlight which observations we used

If the change of basis $\underline{x} = P\underline{y}$ is such that the y_i are uncorrelated, the S_y, the covariance matrix of the observations $\underline{y}$, should be diagonal. Our goal is therefore to pick an orthonormal basis $\underline{u}_1, \dots, \underline{u}_p$ so that for $P = [\underline{u}_1 \cdots \underline{u}_p]$,

$$S_y = P^TS_xP$$

is diagonal. But we already know how to do this! S_X is a symmetric matrix, and thus admits a spectral decomposition $S_X = Q\Lambda Q^T$, where $Q = [\underline{u}_1 \cdots \underline{u}_p]$ is an orthogonal matrix composed of the orthonormal eigenvectors of S_X, and $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$ a diagonal matrix of its eigenvalues. Hence, setting $P = Q$, we have

$$S_y = Q^TS_xQ = Q^TQ\Lambda Q^TQ = \Lambda$$

which is diagonal, as we wanted!

The orthonormal eigenvectors $\underline{u}_1, \dots, \underline{u}_p$ of the covariance matrix S_X are called the **principal components** of the data in the observation matrix X: the **first principle component** is the eigenvector $\underline{u}_1$ corresponding to the largest eigenvalue λ_1, etc.

The first principle component $\underline{u}_1$ determines the new variable Y_1 by projecting the original observation $\underline{x}$ along $\underline{u}_1$. In particular, since $\underline{u}_1$ is the first column of P, $\underline{u}_1^T$ is the first row of P^T, and hence

$$Y_1 = \underline{u}_1^T \underline{x},$$

which is a projection of $\underline{x}$ along the unit vector $\underline{u}_1$. In a similar fashion, $\underline{u}_2$ determines Y_2, and so on. The direction $\underline{u}_1$ is aligned with the direction of maximal variance in the data, so we expect "most of" each observation $\underline{x}_j$ to be aligned with $\underline{u}_1$. We'll see how to use this observation to **compress multivariate data via dimensionality reduction** next, but first, let's take a look at a simple example.

Example The initial data for our Landsat imaging example was composed of 4 million observations $\underline{x}_j \in \mathbb{R}^3$. The associated covariance matrix (after centering the data), is:

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

The eigenvalue/vectors of S_X are:

$$\lambda_1 = 7614.23,\ \underline{u}_1 = \begin{bmatrix} .5471 \\ .6295 \\ .5510 \end{bmatrix},\ \lambda_2 = 427.63,\ \underline{u}_2 = \begin{bmatrix} -.4894 \\ -.3026 \\ .8179 \end{bmatrix},\ \lambda_3 = 98.10,\ \underline{u}_3 = \begin{bmatrix} .6834 \\ -.7157 \\ .1441 \end{bmatrix}$$

Up to 2 decimal places, the first principal component is:

$$Y_1 = .54 x_1 + .63 x_2 + .56 x_3.$$

This equation was used to create image (d) by taking a weighted combination the three spectral band measurements (each suitably converted to gray scale so that they can be combined).
The covariance of the transformed observations $\underline{Y} = (Y_1, Y_2, Y_3)$ is the diagonal matrix

$$\Lambda = \operatorname{diag}(7614.23, 427.63, 98.10)$$

How can we use the fact that $\lambda_1 = 7614.23$ is much larger than $\lambda_2 = 427.63$ and $\lambda_3 = 98.10$?

Compression via Dimensionality Reduction

PCA is very useful for applications in which most of the variation of the data lies in a low dimensional subspace, i.e., can be explained by only a few of the new variables $Y_1, \dots, Y_p$.

The total variance in the original data can be measured by summing together the variances of the original observation variables, i.e., by computing the sum of the diagonal entries of S_X:

$$TVar(X) = s_{11} + s_{22} + \cdots + s_{pp}.$$

Our first observation is that variance is preserved by the change of variables $Y = P^T X$. We leave showing this as an exercise, but the intuition is that since the change of basis matrix P is orthogonal it only rotates/flips vectors, and does not affect length or angles. This means that

$$TVar(Y) = \lambda_1 + \cdots + \lambda_p = s_{11} + \cdots + s_{pp} = TVar(X),$$

where λ_j is the variance in the y_j coordinate. Thus the ratio $\lambda_j / TVar(Y)$ measures the fraction of the total variance "explained" by Y_j.

This suggests that if we are interested in compressing the original data, a strategy could be to discard the directions Y_j for which $\frac{\lambda_j}{TVar(Y)}$ is very small, as these directions do not capture much of the variation in the data.

Example The total variance of the Landsat data is

$$TVar(X) = TVar(Y) = \lambda_1 + \lambda_2 + \lambda_3 = 7614.23 + 427.63 + 98.10 = 8139.96.$$

The percentages of total variance explained by the principal components are:

$$\frac{7614.23}{8139.96} = .935, \quad \frac{427.63}{8139.96} = .053, \quad \frac{98.10}{8139.96} = .012.$$

Thus, 93.5% of the information collected by Landsat for this specific image is captured in photograph (d) — if that is sufficiently accurate for our purposes, we could compress the three images (a)-(c) into (d), and only store (d), reducing our memory requirements by 1/3!

Computing Principal Components and the SVD

The singular value decomposition is the main tool for performing PCA in practice. Suppose $\hat{X}$ is our centered observation matrix, and then define

$$A = \frac{\hat{X}^T}{\sqrt{N}} \in \mathbb{R}^{N \times p}.$$

We assume that $\text{rank}(A) = p$, i.e., our p-dimensional measurements $x_1, \dots, x_N \in \mathbb{R}^p$ span $\mathbb{R}^p$.

Then $A^T A = S_X$ is the covariance matrix of our data. Now, let $A = U \Sigma V^T$ be a singular value decomposition for A: since $\text{rank}(A) = p$, $\Sigma \in \mathbb{R}^{p \times p}$, $V \in \mathbb{R}^{p \times p}$, and $U \in \mathbb{R}^{N \times p}$. Then

$$S_X = A^T A = V \Sigma U^T U \Sigma V^T = V \Sigma^2 V^T \quad (= Q \Lambda Q^T)$$

i.e., $V \in \mathbb{R}^{p \times p}$ orthogonally diagonalizes the symmetric matrix S_X, and defines a spectral factorization of S_X. This allows us to immediately conclude that:

1) The right singular vectors $v_1, \dots, v_p$, which are the orthonormal columns of V, are the eigenvectors of S_X, and hence are the principal components of our data;
2) The square of the singular values of A, σ_i^2, are the p eigenvalues of S_X.

In practice, computing a SVD of A is both faster and more accurate than computing an eigendecomposition of S_X, and is particularly true when the observation vector dimension p is large.