Linear Algebra – Lesson 13
PCA and SVD

In this lesson, we will discuss two related techniques in linear algebra – Principal Component Analysis (PCA) and Singular Value Decomposition (SVD).

1. Principal Component Analysis (PCA)
Before we walk through the idea of PCA, let's read through the first paragraph in its Wikipedia page. Try to see what is clear for you and what isn't:

Principal component analysis

From Wikipedia, the free encyclopedia

Principal component analysis (**PCA**) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called **principal components**. If there are n observations with p variables, then the number of distinct principal components is $\min(n-1, p)$. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors (each being a linear combination of the variables and containing *n* observations) are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

2.1 Dimensionality reduction – a simple example
One of the main applications of Principle Component Analysis (PCA) is dimensionality reduction. Dimensionality reduction can be thought of as a problem of coding and decoding.

Throughout this section, you should think of vectors not as arrows, but as points in space (each point lies at the tip of the vector arrow).

You are given multiple vectors (let's say m vectors) of n dimensions ($\bar{x}^1, \bar{x}^2, \ldots \bar{x}^m \in \mathbb{R}^n$), and your task is to represent them in a space of lower dimension p. For example, Figure 1a shows an example where $n = 2$ and $p = 1$ (the points are two-dimensional, and you wish to represent each point using a single number). In this example, you can see that using the x coordinate only (projecting onto the x axis; Figure 1b) gives a pretty good representation. We lose all the information about the vertical axis, but it seems like a fairly good choice. In contrast, looking at Figure 1c, you can see the projecting the points onto the y axis gives a bad result – we lose a lot of information.

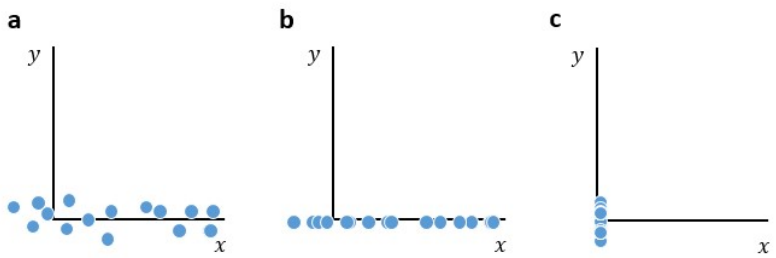
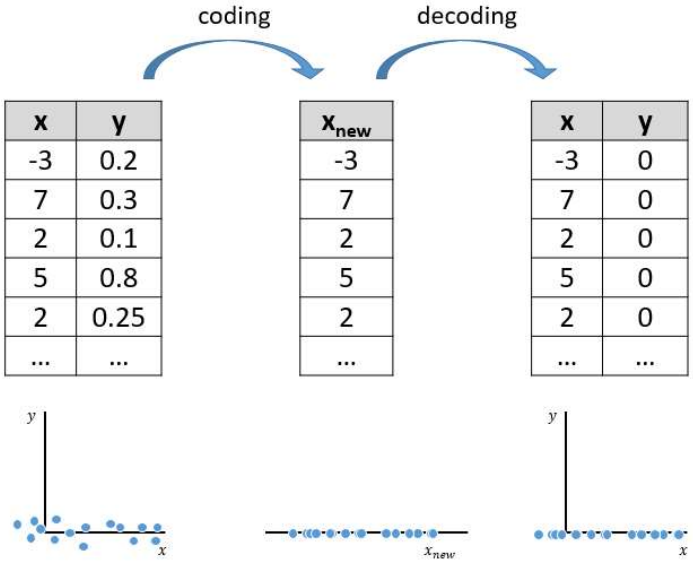


Figure 1.

Questions: In what sense is 1b better than 1c?
Answer: The representation in 1b is better in the sense that it preserves more of the variability between different points. Points that were farther apart before remain far apart even after the transformation.

This variability, or "spread", is often quantified using the **variance**: $\frac{1}{m-1}\sum_{i=1}^{m}\left(x^i - \bar{x}\right)^2$, where m is the number of samples, x^i is sample number i, and $\bar{x}$ is the mean of all samples (which we will assume to be zero). Notice that in judging how good each new representation is, we assumed that both axes are equally important.

Question: What did we actually do here? Where is this "lower dimension" we are talking about, if everything is still plotted in 2D?
Answer: In fact, panels 1b and 1c already show the "reconstructed" or "decoded" points. We actually chose a new axis on which we represented the points as 1D data ("coding") and then converted it back to the 2D space ("decoding"):



2.2 Dimensionality reduction – a still simple example
Now, look at the next example in Figure 2:

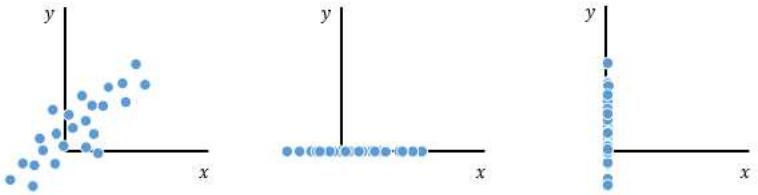


Figure 2.

Here it seems that choosing between projection onto the x axis or the y axis doesn't make much of a difference. Both of them preserve some of the variability, but lose a lot as well. A third option is shown in Figure 3. Now, instead of projecting the data onto one the standard axes, we project it onto a new direction, represented by the red arrow:

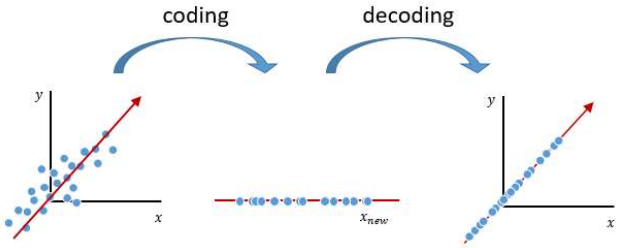
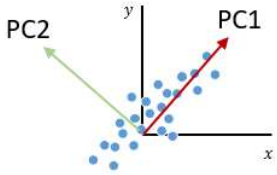


Figure 3.

A better way to think about this is this: we are moving to a new basis, which is still orthogonal, but in which the direction of the first basis vector captures the maximal possible variance in the data (red arrow), and the second basis vector captures the rest of the variance:



These new directions are called the first principal component (PC1) and the second principal component (PC2)[1].

What is special about the new basis that we are looking for?
In the standard basis, the different axes are correlated: if you know the x coordinate of a data-point, you can have a good guess about its y coordinate. In other words, the two axes (or "features") share their variability.
In the new basis we are looking for, the axes are uncorrelated. Knowing the PC1 coordinate of a point, doesn't tell you anything about the PC2 coordinate of this point.

[1] More precisely, these directions are called the principal directions, or principal axes. The principal components are the projections onto these new axes. The first principle component of point $\bar{x}^1$ is its projection onto the unit vector in the direction of the first principal axis.

2.3 Interactive example

Take a look at this interactive example: http://setosa.io/ev/principal-component-analysis/
Move the data points of the left panel and see how this affects the data in the new basis of principal components. Pay attention the two lower panels as well: what is special about the data points distribution in the PC space compared with the standard space?

2.4 Calculating the covariance between two variables

In the next section, we will talk about the covariance matrix. Before we do that, let's make sure we know what covariance and correlation are, and how we can use linear algebra to calculate them.

Intuitively, two variables that change together are correlated. If we measure people's height and show size, we will find that they are correlated – when was is larger, so is the other.

The term "covariance" is strongly related to "correlation". The covariance between two random variables x, y with expected values μ_x and μ_y is defined by:

$$Cov(x,y) = \mathbb{E}\left((x-\mu_x)(y-\mu_y)\right)$$

Where $\mathbb{E}(x)$ is the expectation of x.
Why does this definition make sense?

- If we have a positive covariance, this means that when x tends to be greater than its mean, so does y. When x is smaller than its mean, so is y.
- Let's assume that $\mu_x = 0$ and also $\mu_y = 0$. Then $Cov(x,y)$ will be high if x and y tend to change together, because they will tend to be positive together and negative together. Therefore, the product xy tends to be positive.

Pearson's correlation is simply the covariance of the z-scores. In other words, we can calculate Pearson's correlation between two variable like that[2]:

1. Remove the mean from each variable (so it has a mean of 0)
2. Scale each variable so that its variance is equal to 1
3. Calculate the covariance of the two new variables

So, covariance and correlation are the same thing, up to scaling and zero-meaning the data.

2.4.1 The dot product and the covariance

It turns out that the dot product is very closely related to the covariance.
Remember, we have a set of m measurements in two dimensions (each dimension is a variable).
Let's arrange them in two vectors $\bar{x}$ and $\bar{y}$ (assume that each vector has a mean of zero). Now, using our sample, instead of theoretical expectations, we will have sample means (or empirical means):

$$Cov(x,y) = \frac{1}{m}\sum_{i=1}^{m} x_i y_i$$

[22] This actually tells you that Pearson's correlation is not sensitive to scaling and shifting of the data. Another way to say it, is that Pearson's correlation between two variables is their covariance, scaled by their individual variances: $\rho_{xy} = \frac{Co\ (x,y)}{\sigma_x \sigma_y}$

But this is simply the dot product between the two vectors, divided by m:

$$Cov(x,y) = \frac{1}{n}\bar{x}^T\bar{y}$$

A note about Pearson's correlation and the dot product

Similarly, the dot product is closely related to Pearson's correlation coefficient.

We shift and scale each of the vectors $\bar{x}, \bar{y}$ to have zero mean and unit variance, and denote them by $\tilde{x}, \tilde{y}$. The Pearson correlation between the variable x and y is given by[3]:

$$\frac{1}{n}\tilde{x}^T\tilde{y}$$

2.5 The covariance matrix

Now we are ready to define the covariance matrix.

Our goal then is to represent the data points in a new basis (new "feature" axes) in which there is no correlation between different axes. How should we do this?

The natural way to approach this problem is to come up with a relevant matrix, diagonalize it, and use the eigenvectors as the new basis vectors.

2.5.1 The data matrix

Let's think about a set of m measurements with n features. We can represent it using a matrix $A_{m\times n}$. In general, A would be a rectangular matrix. In the above example, it would be a very long matrix, describing the m samples, each with two measured features (m samples in 2D):

$$A = \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ \dots & \dots \\ \dots & \dots \\ \dots & \dots \\ x_m & y_m \end{pmatrix}$$

The **covariance matrix** tells us about the correlation between the features.

$$Covariance\ matrix = \begin{pmatrix} Var(x) & Cov(x,y) \\ Cov(y,x) & Var(y) \end{pmatrix}$$

We will assume that the mean of each feature is 0. We already know how to calculate each element in this matrix! If we define the columns of A as vectors $\bar{a}_1$ and $\bar{a}_2$, the variance of the first dimension is:

$$\frac{1}{m}\sum_{i=1}^{m} x_i^2 = \frac{1}{m}\begin{pmatrix} x_1 & x_2 & \dots & \dots & x_m \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \dots \\ \dots \\ x_m \end{pmatrix} = \frac{1}{m}\bar{a}_1^T\bar{a}_1$$

And the covariance between the two dimensions is:

[3] See short derivations for the relationship between the dot product and Pearson's correlation coefficient here and here. Spoiler alert: You will find the Pearson's correlation coefficient is just the cosine of the angle between the vectors.

$$\frac{1}{m}\sum_{i=1}^{m} x_1 y_i = \frac{1}{m}(x_1 \quad x_2 \quad \dots \quad \dots \quad x_m)\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ \dots \\ y_m \end{pmatrix} = \frac{1}{m}\bar{a}_1^T \bar{a}_2$$

To calculate the covariance between the features, we can look at $A^T A$:

$$Covariance\ matrix = A^T A = \begin{pmatrix} & \end{pmatrix}\begin{pmatrix} & \\ & \end{pmatrix} = \begin{pmatrix} & \end{pmatrix}$$

The diagonal elements tell us the variance in each dimension. The off-diagonal elements tell us about the covariance between the measurements[4].

2.5.2 The covariance matrix

The matrix $A^T A$ is the sample covariance matrix (or empirical covariance matrix). It is:

- Real and symmetric
- It is diagonalizable, with orthogonal eigenvectors
- Positive definite (its eigenvalues are real and non-negative)

2.5.3 Diagonalizing the covariance matrix

Since the covariance matrix is a real symmetric matrix, we should be able to find its eigenbasis. The eigenvectors will be orthogonal to each other. Now, each data point is represented in a new basis, with n new features which we call $s_1, s_2, \dots, s_n$. In this new basis, the covariance matrix is diagonal:

$$\begin{pmatrix} Var(s_1) & 0 & \dots & 0 \\ 0 & Var(s_2) & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & Var(s_p) \end{pmatrix}$$

This means that there is no covariance between the dimensions – this is exactly what we wanted.

2.6 Dimensionality reduction

The eigenvalues $\lambda_1, \dots, \lambda_p$ are the variances of the features in the new basis. We typically order them from the largest to the smallest:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

Now, to perform dimensionality reduction, we can eliminate the features of the smallest variance.

[4] The covariance matrix is closely related to the correlation matrix. The correlation matrix is the covariance matrix, with each element scaled by the variance of the two relevant features. In the correlation matrix, all the diagonal elements are 1.

2.5 Singular Value Decomposition – SVD

The next section is based on Gilbert Strang's lecture on the SVD.

PCA can be thought of as a result of a more general process called Singular Value Decomposition (SVD).

Let A be an $n \times n$ diagonalizable matrix. Then:

$$A = P\Lambda P^{-1}$$

And we know that the columns of P are the eigenvectors of A: $A\bar{v}_i = \lambda\bar{v}_i$.

If A is also symmetric, we saw that it is diagonalizable by an orthogonal matrix Q:

$$A = Q\Lambda Q^T$$

These are two example of what we call "factorization" – the matrix A is broken into a factor of three matrices. But what if A is not even square? It no longer makes sense to look for "the eigenbasis", because A maps vectors from one space to another space. Yet it turns out that even for $m \times n$ matrices, there exists a very useful factorization, the SVD:

$$A = U\Sigma V^T$$

The matrix U is orthogonal, and its columns are called the left singular vectors.

The matrix V is orthogonal, and its columns are called the right left singular vectors.

The matrix Σ is diagonal, and the values on the diagonal ($\sigma_1 > \sigma_2 > \cdots > \sigma_r > 0$) are called the singular values.

Remember that for orthogonal matrices, $V^TV = I$. So this is equivalent to:

$$AV = U\Sigma$$

Our goal is to find these V, U and Σ.

To understand this factorization more, let's write it explicitly:

$$A\begin{pmatrix} | & | & | \\ \bar{v}_1 & \dots & \bar{v}_r \\ | & | & | \end{pmatrix} = \begin{pmatrix} | & | & | \\ \bar{u}_1 & \dots & \bar{u}_r \\ | & | & | \end{pmatrix}\begin{pmatrix} \sigma_1 & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \sigma_r \end{pmatrix}$$

In other words:

$$A\bar{v}_1 = \sigma_1\bar{u}_1$$
$$A\bar{v}_2 = \sigma_2\bar{u}_2$$
$$\dots$$
$$A\bar{v}_r = \sigma_r\bar{u}_r$$

In the $n \times n$ case, in the $m \times n$ case, we can't even hope to get the same results, because we're moving from $\mathbb{R}^n$ to $\mathbb{R}^m$. However, what we see here is that we can find a set of orthogonal vectors $\bar{v}_1, \bar{v}_2, \dots, \bar{v}_r \in \mathbb{R}^n$ that is mapped to a scaled version of a set of orthogonal vectors $\bar{u}_1, \bar{u}_2, \dots, \bar{u}_r \in \mathbb{R}^m$.

Finding V and Σ

To find the matrices V and Σ, we'll do a trick. Instead of looking at A, we will look at A^TA:

$$A = U\Sigma\mathrm{V}^{\mathrm{T}}$$

Then:

$$A^TA = V\Sigma^{\mathrm{T}}U^TU\Sigma\mathrm{V}^{\mathrm{T}}$$

But since $U^TU = I$, we get:

$$A^T A = V\Sigma^2 \mathrm{V}^{\mathrm{T}}$$

But wait! This tells us that the matrix $A^T A$ is diagonalized by the orthogonal matrix V, and its eigenvalues are $\sigma_1^2, \sigma_2^2, \dots, \sigma_r^2$[5].

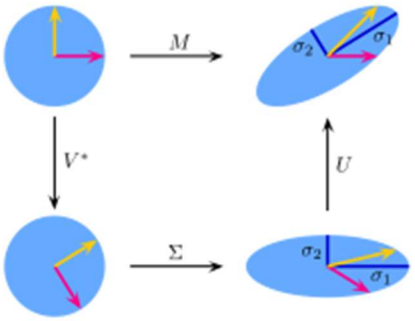
In other words, the columns of V are the eigenvectors of $A^T A$ and the $\sigma^{2\prime}s$ are the eigenvalues of $A^T A$. This is the tight relation between PCA and SVD. If you calculated the PCA, you actually found the relevant matrices of the SVD, and vice versa.

By convention, the $\sigma' s$ are ordered by importance: $\sigma_1 > \sigma_2 > \cdots > \sigma_r > 0$

Similarly, you can show that $AA^T = U\Lambda U^T$

<u>What does it all mean?</u>

SVD has many applications, especially in data science. We can also interpret it geometrically. It tells us that every linear transformation $A_{m\times n}$ can be broken into a series of rotation (with possible reflections), a scaling, and another rotation (with possible reflections). This is illustrated in the following figure from Wikipedia, for the matrix M, which is broken into $U\Sigma V^*$ (here they used V^* to denote the complex conjugate):



$$M = U \cdot \Sigma \cdot V^*$$

<u>A note about choosing the $\bar{u}' s$</u>

You may be tempted to think that we can just diagonalize $A^T A$ to find U. But in fact, a small problem might occur if we have some eigenvalue with a multiplicity greater than 1. In this case, the eigenspace has dimension greater than 1, and we have freedom in choosing exactly where the eigenvectors will point to. We want to make sure that the orthogonal $\bar{v}' s$ are mapped to orthogonal $\bar{u}' s$. To do that, we can define:

$$\bar{u}_1 = \frac{A\bar{v}_1}{\sigma_1}$$
$$\bar{u}_2 = \frac{A\bar{v}_2}{\sigma_2}$$

And so on.

We can easily show that they are orthogonal, and that they are also eigenvectors of AA^T.

Question: Show that $\bar{u}_1$ and $\bar{u}_2$ are orthogonal.

[5] You may have noticed that $A^T A$ is of size $n \times n$, but we only looked at r singular values. The reason is that indeed, the number of singular values may be smaller than n and m. In this case, the rest of the singular values are set to 0. Please refer to the original lecture for a more detailed explanation on how to account for the dimensions here.

Additional resources

A great series of videos on SVD by Steve Brunton
Short SVD overview in Gilbert Strang's Vision of Linear Algebra videos
Short PDF summary of SVD from Gilbert Strang's 18.06SC course
An easy to read PCA tutorial
Yet another well documented tutorial on PCA
PCA vs Ordinary Least Squares
PCA and SVD
PCA vs ICA
PCA and Lagrange multipliers
A step-by-step presentation of PCA