

Lecture 3: Principal Component Analysis (PCA)

1 Introduction

Under many circumstances the data we encounter can be viewed as a cloud of n points in a d -dimensional space. For example, consider the expression matrix E , of N_g genes (rows) and N_s samples (columns). Element E_{gs} is the expression level of gene g in sample s .

Sample i is represented by N_g numbers - the components of a vector:

$$\vec{e}^{(i)} = \begin{pmatrix} e_1^{(i)} \\ e_2^{(i)} \\ \vdots \\ e_{N_g}^{(i)} \end{pmatrix} = \begin{pmatrix} E_{1i} \\ E_{2i} \\ \vdots \\ E_{N_g i} \end{pmatrix}$$

The samples form a cloud of N_s points in N_g dimensional space.

Gene g is represented by N_s numbers - the components of a vector in N_s dimensional space:

$$\vec{e}^{(g)} = \begin{pmatrix} e_1^{(g)} \\ e_2^{(g)} \\ \vdots \\ e_{N_s}^{(g)} \end{pmatrix} = \begin{pmatrix} E_{g1} \\ E_{g2} \\ \vdots \\ E_{gN_s} \end{pmatrix}$$

The genes form a cloud of N_g points in N_s dimensional space. Usually $N_s \ll N_g$.

One can think of a variety of questions to ask about such a cloud of points. For example - do these N_s points belong to one cloud? to 2? or more? To answer, we may want to visualize the points. But how does one visualize many points in higher than 3 dimensions? One answer is to **project** the points down to $d = 2$ or 3. **But** how to choose the subspace, or axes, onto which to project? This lecture is devoted to explaining what is meant by "projecting down to a subspace" and to one particular choice of the selected subspace. The presentation will alternate between a general formalism and a simple example, which is inset.

Example - data

Even though normally $N_s \ll N_g$, consider the expression matrix for $N_g = 2$ genes and $N_s = 4$ samples:

patient No:	5	19	27	37
gene 1	1	9	11	3
gene 2	8	2	4	6

Each patient is represented by a 2-component vector, and a point in $d = 2$ dimensions (see Fig 1):

$$\vec{e}^{(5)} = \begin{pmatrix} 1 \\ 8 \end{pmatrix} \quad \vec{e}^{(19)} = \begin{pmatrix} 9 \\ 2 \end{pmatrix} \quad \vec{e}^{(27)} = \begin{pmatrix} 11 \\ 4 \end{pmatrix} \quad \vec{e}^{(37)} = \begin{pmatrix} 3 \\ 6 \end{pmatrix}$$

Example - two projections

Usually we wish to project the N_s data points down from the "high" dimensional space ($d = N_g$) in which they are embedded. In our example (Fig 1) we can project 4 points down from $d = 2$ to $d = 1$. Two simplest projections are shown in Fig 2; onto the x and y axes. On the y axis, the projections of our samples form 4 equidistant points with a maximal spread of $8-2=6$. On the x axis the four points split into two groups of 2 points in each, with a maximal spread of $11-1=10$. The expression levels of gene 1 are more spread out and separate the samples into two groups.

2. PCA

It is evident from the two projections shown for the example, that *what we see depends on the direction onto which we project*.

Question: How does one find the direction whose projections are the **most spread out**?

A better measure of spread is σ^2 , the variance of the projections: for n numbers $x_i, i = 1, 2, \dots, n$, the mean \bar{x} and variance are

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \sigma^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Example: our 4 samples' projections on the e_1 or x axis are 1,9,11,3; the mean is $\bar{e}_1 = 6$; for the projections on e_2 , i.e. 8,2,4,6, the mean is $\bar{e}_2 = 5$. To calculate the variance, construct the *gene-centered expression matrix*

patient No:	5	19	27	37
$e_1^{(k)} - \bar{e}_1$	-5	3	5	-3
$e_2^{(k)} - \bar{e}_2$	3	-3	-1	1

so that

$$\sigma_1^2 = \frac{1}{4} \sum_{k=1}^4 (e_1^{(k)} - \bar{e}_1)^2 = \frac{1}{4} [(-5)^2 + 3^2 + 5^2 + (-3)^2] = 17 \quad (1)$$

Similarly, $\sigma_2^2 = 5$.

2.1 First Principal Component

We can now formulate the following

PROBLEM: find the direction $\hat{\mathbf{u}}$, such that projecting n points in d dimensions, $\vec{e}^{(i)}, i = 1, \dots, n$, onto $\hat{\mathbf{u}}$, gives the largest variance.

RECIPE:

1. Construct the $d \times d$ *Covariance matrix*

$$C_{ij} = \frac{1}{n} \sum_{k=1}^n [e_i^{(k)} - \bar{e}_i][e_j^{(k)} - \bar{e}_j] \quad (2)$$

2. $\hat{\mathbf{u}}$ is the eigenvector of C with the largest eigenvalue;

$$C\hat{\mathbf{u}} = \lambda_1 \hat{\mathbf{u}} \quad (3)$$

3. The value of the (maximal) variance of the projections onto $\hat{\mathbf{u}}$ is the eigenvalue λ_1 .

A proof (that this procedure indeed maximizes variance of the projections) is given below - first I show how this recipe is applied to the example.

For our **example**, these steps take the following form:

1. The 2×2 *Covariance matrix* is

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} = \begin{pmatrix} 17 & -8 \\ -8 & 5 \end{pmatrix} \quad (4)$$

The matrix elements were calculated using in eq. (2) the entries from the gene-centered expression matrix given above, e.g.

$$C_{11} = \frac{1}{4} \sum_{k=1}^4 [e_1^{(k)} - \bar{e}_1][e_1^{(k)} - \bar{e}_1] = \frac{1}{4} [(-5)^2 + 3^2 + 5^2 + (-3)^2] = \sigma_1^2 = 17 \quad (5)$$

$$C_{12} = C_{21} = \frac{1}{4} \sum_{k=1}^4 [e_1^{(k)} - \bar{e}_1][e_2^{(k)} - \bar{e}_2] = \frac{1}{4} [(-5)(-3) + 3(-3) + 5(-1) + (-3)1] = -8 \quad (6)$$

and similarly one finds that $C_{22} = \sigma_2^2 = 5$.

2. Look for $\hat{\mathbf{u}}$, the normalized eigenvector with the largest eigenvalue:

$$C\hat{\mathbf{u}} = \begin{pmatrix} 17 & -8 \\ -8 & 5 \end{pmatrix} \begin{pmatrix} \hat{u}_1 \\ \hat{u}_2 \end{pmatrix} = \lambda_1 \begin{pmatrix} \hat{u}_1 \\ \hat{u}_2 \end{pmatrix} \quad (7)$$

It is easy to check that the (normalized) vector

$$\begin{pmatrix} \hat{u}_1 \\ \hat{u}_2 \end{pmatrix} = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 \\ -1 \end{pmatrix} \quad (8)$$

satisfies (7) with the eigenvalue $\lambda_1 = 21$, which is the largest (the other, smaller one is $\lambda_2 = 1$).

3. To show that the variance of the projections is indeed λ_1 , we first calculate for our 4 sample vectors their projections onto $\hat{\mathbf{u}}$:

$$e_u^{(k)} = \bar{\mathbf{e}}^{(k)} \cdot \hat{\mathbf{u}} = \sum_{i=1}^d e_i^{(k)} \hat{u}_i \quad (9)$$

which yields for sample 5: $e_u^{(5)} = 1 \cdot 2/\sqrt{5} + 8 \cdot (-1)/\sqrt{5} = -6/\sqrt{5}$, and for the others $e_u^{(19)} = 16/\sqrt{5}$, $e_u^{(27)} = 18/\sqrt{5}$ and $e_u^{(37)} = 0$. The mean of these 4 numbers is $\bar{e}_u = 7/\sqrt{5}$ and their variance is given by

$$\sigma_u^2 = \frac{1}{4} \left[\left(\frac{-6}{\sqrt{5}} - \frac{7}{\sqrt{5}} \right)^2 + \left(\frac{16}{\sqrt{5}} - \frac{7}{\sqrt{5}} \right)^2 + \left(\frac{18}{\sqrt{5}} - \frac{7}{\sqrt{5}} \right)^2 + \left(0 - \frac{7}{\sqrt{5}} \right)^2 \right] = 21 = \lambda_1 \quad (10)$$

2.2 Higher Principal Components

The eigenvector $\hat{\mathbf{u}}^{(2)}$, that corresponds to λ_2 , the second largest eigenvalue of the covariance matrix is the *second Principal Component*, the next is the third and so on. Since these are eigenvectors of a symmetric matrix, they are orthogonal to each other. Representing the data points by their projections onto the first two gives a projection onto the *plane of largest variance* and similarly the first three give the $d = 3$ dimensional space in which the projected data exhibit the largest variation. The total variance of the data,

$$\sigma^2 = \sum_k (\bar{\mathbf{e}}^{(k)} - \bar{\mathbf{e}})^2$$

is also given by the sum of the variances of the projections on the principal components:

$$\sigma^2 = \sum_{\alpha=1}^d \sigma_{\alpha}^2 = \sum_{\alpha} \lambda_{\alpha}$$

and one often gives, together with the projection onto the first 2 or 3 principal components, the fraction of the total variance that has been captured, e.g. for a $d = 3$ representation

$$\sigma^{\text{captured}} = \sum_{\alpha=1}^3 \sigma_{\alpha}^2 / \sigma^2$$

2.3 Warning

PCA is not more than what was presented; it should be born in mind that for many important questions PCA does not give the answer. There may be situations in which the data break into two distinct clouds, but in order to see this one has to project onto a subspace in which the *variation is not maximal* (see fig with two pitas)

2.4 Proof of Recipe

Let us calculate the variance of $k = 1, 2 \dots n$ vectors $\bar{\mathbf{e}}^{(k)}$ onto a direction $\hat{\mathbf{u}}$. The n projections are $e_u^{(k)} = \bar{\mathbf{e}}^{(k)} \cdot \hat{\mathbf{u}}$. The average of these n projections is the projection of the average vector:

$$\bar{e}_u = \frac{1}{n} \sum_{k=1}^n e_u^{(k)} = \frac{1}{n} \sum_{k=1}^n \bar{\mathbf{e}}^{(k)} \cdot \hat{\mathbf{u}} = \left(\frac{1}{n} \sum_{k=1}^n \bar{\mathbf{e}}^{(k)} \right) \cdot \hat{\mathbf{u}} = \bar{\mathbf{e}} \cdot \hat{\mathbf{u}}$$

and the variance of the projections is

$$\begin{aligned}
\sigma_u^2 &= \frac{1}{n} \sum_{k=1}^n \left[e_u^{(k)} - \bar{e}_u \right]^2 = \frac{1}{n} \sum_{k=1}^n \left[\hat{\mathbf{u}} \cdot \left(\vec{\mathbf{e}}^{(k)} - \bar{\mathbf{e}} \right) \right]^2 \\
&= \sum_{i=1}^d \sum_{j=1}^d \hat{u}_i \hat{u}_j \left[\frac{1}{n} \sum_{k=1}^n \left(e_i^{(k)} - \bar{e}_i \right) \left(e_j^{(k)} - \bar{e}_j \right) \right] \\
&= \sum_{i=1}^d \sum_{j=1}^d \hat{u}_i \hat{u}_j C_{ij}
\end{aligned} \tag{11}$$

where in the last equality we recognized the object in square brackets as the covariance matrix C_{ij} . This result is written in concise vector notation as

$$\sigma_u^2 = \hat{\mathbf{u}}^T C \hat{\mathbf{u}} \tag{12}$$

expressing the variance of the projections onto any, general direction $\hat{\mathbf{u}}$ in terms of the vector $\hat{\mathbf{u}}$ and the covariance matrix of the data, C . Our task is to find the direction $\hat{\mathbf{u}}$ which maximizes this variance, under the constraint $\hat{\mathbf{u}}^2 = 1$, namely that $\hat{\mathbf{u}}$ is a vector of unit norm. Hence we have to maximize

$$F = \hat{\mathbf{u}}^T C \hat{\mathbf{u}} - \lambda(\hat{\mathbf{u}}^2 - 1) \tag{13}$$

where λ is a Lagrange multiplier. Taking the derivative with respect to $\hat{u}_i, i = 1, \dots, d$, we get d equations that have the form

$$C \hat{\mathbf{u}} = \lambda \hat{\mathbf{u}} \tag{14}$$

which identifies the extremal $\hat{\mathbf{u}}$ as an eigenvalue of C , the Lagrange multiplier λ as the corresponding eigenvalue and, with (12), the variance of the projections onto $\hat{\mathbf{u}}$ is given by $\sigma_u^2 = \lambda$, as claimed above.

Lecture 4. Multi-Dimensional Scaling (MDS)

1. Introduction

PCA identifies a subspace, onto which we project the data to capture most of its variation. PCA does not distort the information it presents, but suppresses information which is not presented. Two points may appear to be close in the 2 or 3 dimensional subspace onto which we project, even though in the full high dimensional space they are not near each other. MDS does the opposite: it looks for a low dimensional representation of the data which preserves (as much as possible) the relative distances of the points in the full space. The cost is an overall distortion of the positions (non-linear transformation).

The basic quantity one focuses on is the *distance matrix* \mathbf{D} . It's i, j element is the distance between the two points $\bar{\mathbf{x}}^{(i)}$ and $\bar{\mathbf{x}}^{(j)}$;

$$D_{ij} = |\bar{\mathbf{x}}^{(i)} - \bar{\mathbf{x}}^{(j)}| = \sqrt{\sum_{k=1}^d [x_k^{(i)} - x_k^{(j)}]^2} \quad (15)$$

For N points in d dimensions D is an $N \times N$ symmetric matrix.

In our example of 4 samples in a $d = 2$ dimensional gene-space, D_{12} is the distance between the first two points,

$$D_{12} = |\bar{\mathbf{e}}^{(5)} - \bar{\mathbf{e}}^{(19)}| = \sqrt{(1-9)^2 + (8-2)^2} = 10 \quad (16)$$

and the full distance matrix is given, [for the points ordered (5),(19),(27),(32)] by

$$D = \begin{pmatrix} 0 & 10 & 10.77 & 2.82 \\ 10 & 0 & 2.82 & 7.21 \\ 10.77 & 2.82 & 0 & 8.24 \\ 2.82 & 7.21 & 8.24 & 0 \end{pmatrix} \quad (17)$$

2. Multi Dimensional Scaling

The problem: given N points $\mathbf{x}^{(i)}$ with $i = 1, 2, \dots, N$ in some d dimensional space, find N points $\mathbf{y}^{(i)}$ with $i = 1, 2, \dots, N$ in some $\tilde{d} = 2, 3$ dimensional space, such that their distance matrix $\tilde{\mathbf{D}}$ is as close as possible to the "real" distance matrix \mathbf{D} . The distance matrix $\tilde{\mathbf{D}}$ is defined by

$$\tilde{D}_{ij} = |\bar{\mathbf{y}}^{(i)} - \bar{\mathbf{y}}^{(j)}| \quad (18)$$

How does one define the distance between two distance matrices, $\tilde{\mathbf{D}}$ and \mathbf{D} ? There are $N(N-1)/2$ independent matrix elements; usually we cannot satisfy equality of all, i.e. $\tilde{D}_{ij} = D_{ij}$ for all pairs i, j . Hence one defines a cost function that measures the deviation from this perfect solution. The obvious choice is

$$J_0 = \sum_{i < j} (D_{ij} - \tilde{D}_{ij})^2 \quad (19)$$

This choice is however problematic since it measures absolute error. Say for some points \mathbf{x} we found a good solution \mathbf{y} ; if we scale all coordinates \mathbf{x} by a factor $\lambda > 1$, we would like the similarly scaled positions $\lambda\mathbf{y}$ to be a good solution. The new error is, however, λ^2 times larger. Hence one defines scaled cost functions. These choices are

$$J_{ee} = \frac{\sum_{i < j} (D_{ij} - \tilde{D}_{ij})^2}{\sum_{i < j} D_{ij}^2} \quad (20)$$

$$J_{ff} = \sum_{i < j} \left(\frac{D_{ij} - \tilde{D}_{ij}}{D_{ij}} \right)^2 \quad (21)$$

$$J_{ef} = \frac{1}{\sum_{i < j} D_{ij}} \sum_{i < j} \frac{(D_{ij} - \tilde{D}_{ij})^2}{D_{ij}} \quad (22)$$

Remember that the coordinates of the data points $\mathbf{x}^{(i)}$ are given, and we should minimize the cost J over the unknown coordinates $y_k^{(i)}$. There are $\tilde{d}N$ unknowns ($3N$ or $2N$) and we minimize J over them by gradient descent. The solution one obtains is not unique - depends on the initial point.

J_{ee} emphasizes large errors - a difference of 10 between 1000 and 1010 has the same contribution to the cost and the difference between 10 and 20. J_{ff} measures fractional error.