Principal component analysis

Angela Montanari

1 Introduction

Principal component analysis (PCA) is one of the most popular multivariate statistical methods.

It was first introduced by Pearson (1901) and later developed by Hotelling (1933) according to two very different perspectives. The solution more frequently adopted is the one due to Hotelling. We will introduce the method according to Hotelling's formulation and, in the end, we will show its equivalence to Pearson's approach.

The basic idea of principal component analysis is to find a small number of linear combinations of the observed variables which explain most of the variation in the data. The first principal component is the linear combination with maximal variance; we are essentially searching for a dimension along which the observations are maximally separated or spread out. The second principal component is the linear combination with maximal variance in a direction orthogonal to the first principal component, and so on.

2 Some theory

Let's consider a p-dimensional random vector $\mathbf{x}$, with expected value $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and define a linear combination $y_1 = \mathbf{a}_1^T\mathbf{x}$.
We want to find the vector $\mathbf{a}_1$ such that the variance of y_1, $V(y_1) = \mathbf{a}_1^T\boldsymbol{\Sigma}\mathbf{a}_1$ is maximum.

We already know that such a function doesn't admit a finite maximum. In order to optimize it we need to impose a constraint on the norm of the vector $\mathbf{a}_1$. In particular, as a unit norm vector uniquely defines a direction in the multidimensional space, we optimize $V(y_1)$ under the constraint $\mathbf{a}_1^T\mathbf{a}_1 = 1$. More formally we look for a vector $\mathbf{a}_1$ such that:

$$\max_{\mathbf{a}_1} \mathbf{a}_1^T \mathbf{\Sigma} \mathbf{a}_1 \quad \text{under the constraint} \quad \mathbf{a}_1^T \mathbf{a}_1 = 1$$

We can restate the problem in the constrained optimization framework based on Lagrange multipliers by defining the function

$$\phi = \mathbf{a}_1^T \mathbf{\Sigma} \mathbf{a}_1 - \lambda_1 (\mathbf{a}_1^T \mathbf{a}_1 - 1)$$

and by looking for the vector $\mathbf{a}_1$ that maximizes it. This optimization problem can be solved by differentiating ϕ with respect to $\mathbf{a}_1$ and equating to 0 all the partial derivatives:

$$\frac{\partial \phi}{\partial \mathbf{a}_1} = 2\mathbf{\Sigma} \mathbf{a}_1 - 2\lambda_1 \mathbf{a}_1 = 0$$

that is $\mathbf{\Sigma} \mathbf{a}_1 - \lambda_1 \mathbf{a}_1 = 0$ or equivalently

$$\mathbf{\Sigma} \mathbf{a}_1 = \lambda_1 \mathbf{a}_1 \tag{1}$$

One can easily recognize in this identity the relationship between the eigenvalues and the eigenvectors of the covariance matrix $\mathbf{\Sigma}$: λ_1 is an eigenvalue of $\mathbf{\Sigma}$ and $\mathbf{a}_1$ is the corresponding eigenvector. But $\mathbf{\Sigma}$ has p eigenvalues and p corresponding eigenvector. The still open issue is therefore the detection of the couple eigenvalue-eigenvector that is relevant for our purpose.
If we multiply both sides of equation (1) by $\mathbf{a}_1^T$, because of the unit norm constraint we obtain:

$$\mathbf{a}_1^T \mathbf{\Sigma} \mathbf{a}_1 = \lambda_1 \mathbf{a}_1^T \mathbf{a}_1 = \lambda_1.$$

λ_1 exactly coincides with the variance of y_1 i.e. with the quantity we want to optimize. Therefore, in order to derive the linear combination having the largest variance, we simply need to consider the largest eigenvalue of $\mathbf{\Sigma}$ and $\mathbf{a}_1$ will be the corresponding eigenvector.

In order to derive the wanted solution equation(1) may be rewritten as:

$$(\mathbf{\Sigma} - \lambda_1 \mathbf{I}) \mathbf{a}_1 = 0.$$

It is an homogeneous linear equation system and it admits a non trivial solution if and only if

$$\det(\mathbf{\Sigma} - \lambda_1 \mathbf{I}) = 0.$$

This means that λ_1 must be a root of the characteristic polynomial (an eigenvalue) and $\mathbf{a}_1$ will be the corresponding eigenvector. We have already proved

that the solution is offered by the largest eigenvalue and by the corresponding eigenvector.

The linear combination y_1 having the eigenvector $\mathbf{a}_1$ as the vector of coefficients is the *first principal component*.

Our search for maximum variance linear combinations can be further pursued.

We look for a second linear combination of $\mathbf{x}$, $y_2 = \mathbf{a}_2^T\mathbf{x}$, under the constraints: $\mathbf{a}_2^T\mathbf{a}_2 = 1$ and $\mathbf{a}_2^T\mathbf{a}_1 = \mathbf{a}_1^T\mathbf{a}_2 = 0$. The first constraint is the usual unit norm constraint, the second one simply means that we look for $\mathbf{a}_2$ in the orthogonal complement of $\mathbf{a}_1$.

The constrained optimization problem now consists in maximizing, with respect to $\mathbf{a}_2$, the following function:

$$\phi = \mathbf{a}_2^T\mathbf{\Sigma}\mathbf{a}_2 - \lambda_2(\mathbf{a}_2^T\mathbf{a}_2 - 1) - \lambda_3\mathbf{a}_2^T\mathbf{a}_1$$

After differentiating with respect to $\mathbf{a}_2$, and equating the derivatives to 0 we obtain:

$$2\mathbf{\Sigma}\mathbf{a}_2 - 2\lambda_2\mathbf{a}_2 - \lambda_3\mathbf{a}_1 = 0 \tag{2}$$

On pre-multiplying both sides of equation (2) by $\mathbf{a}_1^T$ we obtain

$$2\mathbf{a}_1^T\mathbf{\Sigma}\mathbf{a}_2 - 2\lambda_2\mathbf{a}_1^T\mathbf{a}_2 - \lambda_3\mathbf{a}_1^T\mathbf{a}_1 = 0$$

Remembering that $\mathbf{a}_1^T$ is an eigenvector of $\mathbf{\Sigma}$, and hence $\mathbf{a}_1^T\mathbf{\Sigma} = \lambda_1\mathbf{a}_1^T$, that $\mathbf{a}_2^T\mathbf{a}_2 = 1$ and that $\mathbf{a}_1^T\mathbf{a}_2 = 0$ we obtain that $\lambda_3 = 0$.

The problem we need to solve is therefore $\mathbf{\Sigma}\mathbf{a}_2 - \lambda_1\mathbf{a}_2 = 0$ or equivalently

$$\mathbf{\Sigma}\mathbf{a}_2 = \lambda_2\mathbf{a}_2 \tag{3}$$

λ_2 is an eigenvalue of $\mathbf{\Sigma}$ and $\mathbf{a}_2$ is the corresponding eigenvector. As we are looking for the linear combination having the largest variance not accounted for by y_1 we will choose the second largest eigenvalue of $\mathbf{\Sigma}$ and the corresponding eigenvalue. $y_2 = \mathbf{a}_2^T\mathbf{x}$ is the *second principal component*

The above process can be continued for all principal components. We will derive as many principal components as the observed variables. In general, the k-th PC of $\mathbf{x}$ is $y_k = \mathbf{a}_k^T\mathbf{x}$ and $V(\mathbf{a}_k^T\mathbf{x}) = \lambda_k$ where λ_k is the k-th largest eigenvalue of $\mathbf{\Sigma}$, and $\mathbf{a}_k$ is the corresponding eigenvector.

If we pre-multiply both sides of equation (3) by $\mathbf{a}_1^T$ we obtain:

$$\mathbf{a}_1^T\mathbf{\Sigma}\mathbf{a}_2 = \lambda_2\mathbf{a}_1^T\mathbf{a}_2$$

that, due to the orthogonality constraint $\mathbf{a}_1^T\mathbf{a}_2 = 0$, becomes:

$$\mathbf{a}_1^T\mathbf{\Sigma}\mathbf{a}_2 = 0. \tag{4}$$

One can easily recognize in $\mathbf{a}_1^T\mathbf{\Sigma}\mathbf{a}_2$ the expression of the covariance between y_1 and y_2. In fact:

$$cov(y_1, y_2) = E[(y_1 - \bar{y_1})(y_2 - \bar{y_2})] = E[(\mathbf{a}_1^T\mathbf{x} - \mathbf{a}_1^T\boldsymbol{\mu})(\mathbf{a}_2^T\mathbf{x} - \mathbf{a}_2^T\boldsymbol{\mu})^T] =$$

$$= \mathbf{a}_1^T E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T]\mathbf{a}_2 = \mathbf{a}_1^T\mathbf{\Sigma}\mathbf{a}_2$$

Equation (4) means that the first and the second principal components are uncorrelated. The same holds for any pair of principal components.

In summary: starting from the p-dimensional random vector $\mathbf{x}$ we have obtained a new p-dimensional random vector $\mathbf{y}$, such that $\mathbf{y} = \mathbf{A}^T\mathbf{x}$ where $\mathbf{A}$ is an orthonormal matrix whose k-th column is the eigenvector of $\mathbf{\Sigma}$ corresponding to the k-th largest eigenvalue. Thus, the PCs are defined by an orthonormal linear transformation of $\mathbf{x}$.

Directly from the previous derivation we have $\mathbf{\Sigma}\mathbf{A} = \mathbf{A}\mathbf{\Lambda}$ where $\mathbf{\Lambda}$ is the diagonal matrix whose k-th diagonal element is λ_k, the k-th eigenvalue of $\mathbf{\Sigma}$, and $\lambda_k = V(\mathbf{a}_k^T\mathbf{x}) = V(y_k)$. Two alternative ways of expressing $\mathbf{\Sigma}\mathbf{A} = \mathbf{A}\mathbf{\Lambda}$ that derive from A being orthogonal are

$$\mathbf{A}^T\mathbf{\Sigma}\mathbf{A} = \mathbf{\Lambda}$$

meaning that the covariance matrix of the principal components is diagonal (i.e. principal components are uncorrelated), and

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T$$

$\mathbf{\Sigma}$ and $\mathbf{\Lambda}$ are two similar matrices; this implies they have the same trace $tr(\mathbf{\Sigma}) = tr(\mathbf{\Lambda})$. In statistical terms this means that the Principal component transformation leaves the total variance unchanged.

Exercise 3

Given a bi-variate random vector x with covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

derive the principal components

Solution

The eigenvalues and the eigenvectors of $\mathbf{\Sigma}$ must be derived:

$$(\mathbf{\Sigma} - \lambda\mathbf{I})\mathbf{a} = \begin{bmatrix} 5-\lambda & 2 \\ 2 & 2-\lambda \end{bmatrix} \mathbf{a} = 0$$

In order to derive the eigenvalues, the following equation needs to be solved:

$$\det \begin{bmatrix} 5-\lambda & 2 \\ 2 & 2-\lambda \end{bmatrix} = 0$$

$$(5-\lambda)(2-\lambda) - 4 = 0$$
$$10 - 5\lambda - 2\lambda + \lambda^2 - 4 = 0$$
$$\lambda^2 - 7\lambda + 6 = 0$$

The solutions of this second degree linear equation are 6 and 1. Let's consider the largest eigenvalue (6) first and derive the corresponding eigenvector:

$$\begin{bmatrix} 5-6 & 2 \\ 2 & 2-6 \end{bmatrix} \mathbf{a}_1 = 0$$

$$\begin{bmatrix} -1 & 2 \\ 2 & -4 \end{bmatrix} \begin{bmatrix} {}_1a_1 \\ {}_2a_1 \end{bmatrix} = 0$$

$$\begin{cases} -{}_1a_1 + 2{}_2a_1 = 0 \\ 2{}_1a_1 - 4{}_2a_1 = 0 \end{cases}$$

and, after solving the first equation thus obtaining ${}_1a_1$ and replacing it int the second equation we obtain

$$\begin{cases} {}_1a_1 = 2{}_2a_1 \\ 4{}_2a_1 - 4{}_2a_1 = 0 \end{cases}$$

The second equation is always true for any value of ${}_2a_1$. So we have one equation only, with two unknowns. In order to obtain a solution we must remember the unit norm constraint. This leads to the following linear equation system:

$$\begin{cases} {}_1a_1 = 2_2a_1 \\ {}_1a_1^2 +_2 a_1^2 = 1 \end{cases}$$

After replacing ${}_1a_1$ in the second equation, with a little algebra we obtain

$$\begin{cases} {}_1a_1 = 2_2a_1 \\ {}_2a_1 = \pm\frac{1}{\sqrt{5}} \end{cases}$$

This means that we have two solutions: $\mathbf{a}_1^T = (\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}})$ and $\mathbf{a}_1^T = (-\frac{2}{\sqrt{5}}, -\frac{1}{\sqrt{5}})$. However these two solutions are equivalent as they identify the same direction in the multidimensional space. The positive one is usually preferred. This result points to an important characteristic of Principal components: they are uniquely defined up to a sign change of all the coefficients.

3 Principal components derived from a correlation matrix

The derivation and properties of PCs considered above are based on the eigenvectors and eigenvalues of the covariance matrix. In practice however it is often common to define principal components as linear combinations of standardized variables. This means eigenvalues and eigenvectors are extracted form the correlation matrix, instead of from the covariance matrix.
It might seem that the PCs for a correlation matrix could be obtained fairly easily from those for the corresponding covariance matrix. However, this is not the case; the eigenvalues and eigenvectors of the correlation matrix have no simple relationship with those of the corresponding covariance matrix and there's no way to directly go from one solution to the other.
Let $\boldsymbol{\Delta}$ denote the diagonal matrix whose diagonal elements are the same as those of $\boldsymbol{\Sigma}$. Then the correlation matrix $\boldsymbol{\rho}$ can be obtained from $\boldsymbol{\Sigma}$ as $\boldsymbol{\rho} = \boldsymbol{\Delta}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{\Delta}^{-1/2}$. From this the identity $\boldsymbol{\Sigma} = \boldsymbol{\Delta}^{1/2}\boldsymbol{\rho}\boldsymbol{\Delta}^{1/2}$ derives
Let's recall that, for any principal component derived from $\boldsymbol{\Sigma}$, the following relationship holds: $\boldsymbol{\Sigma}\mathbf{a} = \lambda\mathbf{a}$. It can equivalently be re-written as

$$\boldsymbol{\Delta}^{1/2}\boldsymbol{\rho}\boldsymbol{\Delta}^{1/2}\mathbf{a} = \lambda\mathbf{a}$$

i.e.

$$\boldsymbol{\rho}\boldsymbol{\Delta}^{1/2}\mathbf{a} = \lambda\boldsymbol{\Delta}^{-1/2}\mathbf{a}$$

If we set $\mathbf{c} = \boldsymbol{\Delta}^{1/2}\mathbf{a}$ and we replace it in the above equation we obtain:

$$\boldsymbol{\rho}\mathbf{c} = \lambda\boldsymbol{\Delta}^{-1}\mathbf{c}$$

Since the elements of Δ are not necessarily all equal, the previous expression does not define an eigenvalue-eigenvector relationship and shows that the principal components derive from $\boldsymbol{\Sigma}$ or from $\boldsymbol{\rho}$ are not the same and that knowledge of one set does not allow simple transformation to the other set.

The decision to deal with standardized or raw data needs therefore to be carefully considered. A drawback of Principal Component Analysis based on covariance matrices is the sensitivity of the PCs to the units of measurement used for each element of $\mathbf{x}$. If there are large differences between the variances of the elements of $\mathbf{x}$, then those variables whose variances are largest will tend to dominate the first few PCs. Let's consider a simple artificial example. Imagine we consider two variables X_1 and X_2 where X_1 can be expressed both in *cm* or in *mm*. The covariance matrices in the two cases are

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 90 & 50 \\ 50 & 90 \end{bmatrix}$$

and

$$\boldsymbol{\Sigma}_2 = \begin{bmatrix} 9000 & 500 \\ 500 & 90 \end{bmatrix}$$

The first principal component of $\boldsymbol{\Sigma}_1$ is $y_1 = 0.707X_1 + 0.707X_2$ and the corresponding eigenvalue, i.e. its variance, is 140. This component explains $140/180 * 100 = 77.78\%$ of the total variance.
The first principal component of $\boldsymbol{\Sigma}_2$ is $y_1 = 0.998X_1 + 0.055X_2$ and its variance is 9027.97. It explains $9027.97/9090 * 100 = 99.32\%$ of the total variance.

A simple change in the measurement unit has transformed a balanced component into a component almost completely dominated by variable X_1. This example has shown that it might be troublesome to use PCs on a covariance matrix when the variables are expressed according to different measurement scales, unless there is a strong conviction that the units of measurements chosen are the only ones that make sense. Even when this condition holds, using the covariance matrix will not provide very informative PCs if the variables have widely differing variances, as the PCs will tend to reproduce those variables having the largest variance. Furthermore, with covariance matrices and non-commensurable variables the PC scores might be difficult to interpret.

Exercise 3
Given the correlation matrix

$$\boldsymbol{\rho}_2 = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

derive the principal components.

4 Sample principal components and dimension reduction

Our developments so far have been related to random vectors. When faced with a real situation we can derive principal components starting from the correlation matrix $\mathbf{R}$ or from the sample covariance matrix $\mathbf{S}$, depending on whether variable standardization has been deemed necessary or not. We will still denote by $\mathbf{a}_k$ the generic vector of coefficients and by l_k the corresponding eigenvalue, representing the variance of X_k. The properties derived for population PCs will hold for sample PCs as well.
Besides deriving uncorrelated linear combinations with maximum variance, a further goal of PCA is to perform "dimension reduction" i.e. to reduce a large number (p) of variables to a much smaller number (m) of PCs whilst retaining as much as possible of the variation in the p original variables.

Many criteria have been proposed in the statistical literature for the determination of m. We will describe them in the following. All of them are heuristic rules rather then formal rigorous procedures. This is because PCA is a completely distribution free method and it is mainly used for descriptive purposes. Of course it is possible to define rigorous inferential procedures and tests for the selection of the number m of components that should be retained, but this would require distributional assumptions while the validity of the method lives independently of them.
Cumulative proportion of explained variance

This criterion suggests to retain as many PCs as are needed in order to explain approximately 80-90% of the total variance.
The proportion of the total variance explained by the k-th PC when dealing with the covariance matrix is

$$\frac{l_k}{\sum_{k=1}^{p} l_k} * 100$$

and

$$\frac{l_k}{p} * 100$$

when dealing with standardized variables.
The criterion suggests to retain m principal components, where m is such that

$$\frac{\sum_{k=1}^{m} l_k}{\sum_{k=1}^{p} l_k} * 100 = \frac{\sum_{k=1}^{m} l_k}{tr(\mathbf{S})} * 100 = 80 - 90\%$$

for the raw data or

$$\frac{\sum_{k=1}^{m} l_k}{p} * 100 = \frac{\sum_{k=1}^{m} l_k}{tr(\mathbf{R})} * 100 = 80 - 90\%$$

for standardized data.
Kaiser's rule

This rule was first proposed by Kaiser as a selection criterion for PCs derived from a correlation matrix. The rules suggests to retain as many principal components as are the eigenvalues of $\mathbf{R}$ larger than 1. The motivation underlying this rule is that we want to retain all the principal components that have a variance larger than the one related to the original variables (that is equal to 1 for standardized data).
As 1 might also be considered the average variance for standardized data, that rule has been modified in order to select PCs derived from the covariance matrix as follows: retain as many PCs as are the ones whose variance is larger than $\bar{l} = \sum_{k=1}^{p} l_k/p$.

Scree plot

The third criterion, that ca be applied to PCS extracted both from S or R is a graphical criterion. It is based on what is called a "scree plot", that is a plot of the l_k against k ($k = 1, \ldots, p$). A typical pattern is shown in Fig.2

After a sharp decline the curve tends to become flat. The flat portion corresponds to noise components, unable to capture the leading variability; the rule therefore suggests that m should correspond to the value of k at which the elbow of the scree plot occurs (3 in the picture)

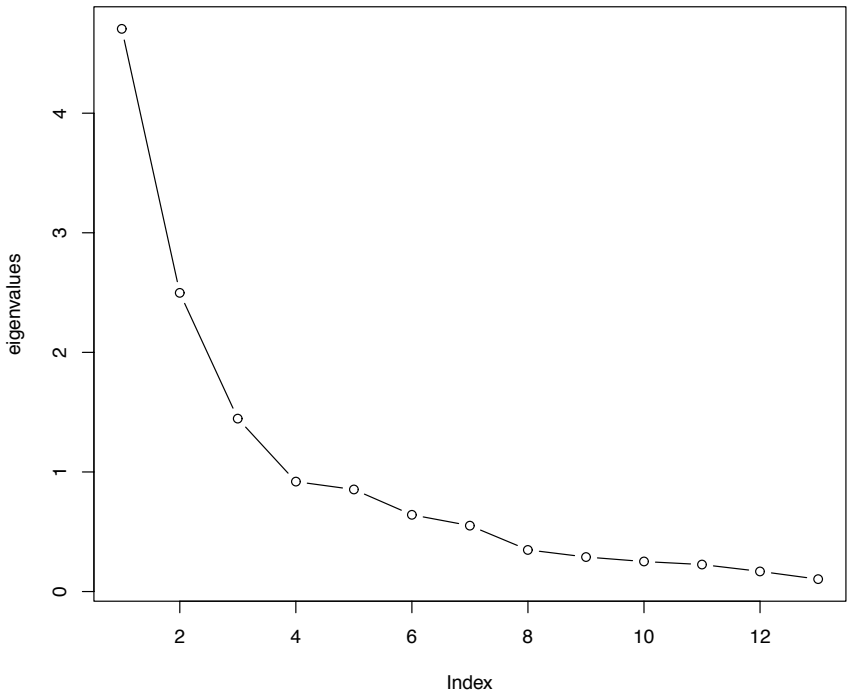

Figure 1: Scree plot

5 An empirical example...with still some theory

A very popular example of an empirical application of PCA is the study of size and shape relations for painted turtles, due to Jolicoeur and Mosimann (1960). They measured carapace length, width and height (in mm) for 24 male turtles. The data were reported in Table 1. The average vector for the three variables is $\bar{\mathbf{x}}^T = (113.375, 88.29167, 40.70833)$ and the variances are 132.9844, 47.9566, 10.78993 respectively. The variances, however expressed in the same measurement unit, are very different. This suggests that a kind of normalization is advisable in order to prevent the first PC from being completely dominated by X_1. They resorted to a log transform.

As the following covariance matrix obtained from the log transformed data shows, this transformation effectively homogenizes variances but its use cannot be suggested as a general rule because it requires that all the data are positive and no zero value is present in the data set, and this is not always the case. Anyway, for reasons that will be made clear in the following, it is very useful in studies of size and shape (allometry) as is the one at hand.

$$\mathbf{S} = 10^{-3} \begin{bmatrix} 11.072 & 8.019 & 8.160 \\ 8.019 & 6.417 & 6.005 \\ 8.160 & 6.005 & 6.773 \end{bmatrix}$$

The total variance is 0.024261488.

The eigenvalues, i.e. the variances of the principal components are

$l_1 = 0.023303$ $l_2 = 0.000598$ $l_3 = 0.00036$

The first principal component accounts for the 96.0508 per cent of the total variance and so it is enough, alone, to reproduce most of the observed variability. It is also evident that its eigenvalue is larger than the average eigenvalue. The modified version of Kaiser's rule still suggests one PC is enough.

The corresponding eigenvector is $\mathbf{a}_1^T = (0.683102, 0.510220, 0.522539)$. The first principal component is a linear combination of the log transformed carapace measurements; it separates large turtles, with a large, long, thick carapace, and small turtles, with a narrow, short, thin carapace. We might say it separates turtles according to their size. Moreover, since

$$y_1 = 0.683 \ln(length) + 0.510 \ln(width) + 0.523 \ln(height) =$$

$$= \ln[(lenght)^{0.683}(width)^{0.510}(height)^{0.523}]$$

the first PC may be viewed as the $\ln(volume)$ of a box whose dimensions have somehow been adjusted in order to account for the round shape of the carapace.
The second eigenvector is $\mathbf{a}_2^T = (-0.159479, -0.594012, 0.78849)$. Even if the second PC is not important in recovering a reduced dimension representation of our data we ca try to give it an empirical meaning. It is a linear contrast i.e. some variables have a positive coefficient and some have a negative one. It separates turtles having a thick but small carapace, from turtles having a thin and large carapace. It can be thought of as a measure of shape.

We have interpreted the first PCs by examining the size and the sign of the different coefficients. Some authors suggest to use the correlations between each principal component and all the observed variables instead.
In order to derive a simple expression for these correlations let's restate the problem in terms of random variables/vectors. Let's assume, without loss of generality, that we deal with mean centered variables. Then:

$$cov(y_1, \mathbf{x}) = E(y_1 \mathbf{x}^T) = E(\mathbf{a}_1^T \mathbf{x}\mathbf{x}^T) = \mathbf{a}_1^T E(\mathbf{x}\mathbf{x}^T) = \mathbf{a}_1^T \mathbf{\Sigma} = \lambda_1 \mathbf{a}_1^T$$

and

$$corr(y_1, \mathbf{x}) = \lambda_1^{-1/2} cov(y_1, \mathbf{x}) \mathbf{\Delta}^{-1/2} = \lambda_1^{-1/2} \lambda_1 \mathbf{a}_1^T \mathbf{\Delta}^{-1/2} = \lambda_1^{1/2} \mathbf{a}_1^T \mathbf{\Delta}^{-1/2}$$

The corresponding sample expression is

$$corr(y_1, \mathbf{x}) = l_1^{1/2} \mathbf{a}_1^T \mathbf{D}^{-1/2}$$

In our example

$$corr(y_1, \mathbf{x}) = (0.99, 0.97, 0.97)$$

that is the first principal component is highly positively correlated with each of the observed variables. All variables contribute more or less in the same way to the first principal component, confirming the interpretation given above.

PCA is often also used as a method of data display. The scores on the retained principal components, i.e. the coordinates of the n statistical units in the m dimensional space spanned by the first m eigenvectors can be easily computed as

$$\mathbf{Y} = \mathbf{X}\mathbf{A}_m$$

where $\mathbf{Y}$ is the $n \times m$ matrix of the PC scores, $\mathbf{X}$ is the original $n \times p$ data matrix and $\mathbf{A}_m$ is the $p \times m$ matrix whose columns are the first m eigenvectors of $\mathbf{S}$ (the same obviously holds if we replace $\mathbf{X}$ by $\mathbf{Z}$ and $\mathbf{S}$ by $\mathbf{R}$).
If $m = 2$ the scores can be plotted on a Cartesian plane. Examples will be given in the labs.

As this bi-dimensional plot is an approximation of the structure in the original p-dimensional space it may happen that not all of the n statistical units are well represented in the bi-dimensional projection. A way to check the adequacy of the approximation for each statistical unit is an inspection of the scores on the last (discarded) PCs. If a unit has high scores on the last PCs this means that it is largely displaced when projected on the PC plane and therefore it is misrepresented. One could consider, for each statistical unit the sum of the squared scores on the last $p - m$ PCs. The position on the PC plane of those units having a large value of such a sum should be carefully considered.
The very last PCs having a very small variance might also be useful in detecting almost constant linear relationships among the columns of $\mathbf{X}$. They indeed represent a useful tool for the diagnosis of multicollinearity in multiple linear regression.

6 Historical remarks

According to the formulation of the method due to Hotelling, the transformation of a random vector $\mathbf{x}$ into PCs corresponds to an orthogonal rotation of the original reference system such that the variance of the projections along the new axes is maximum.
Let's consider the bi-dimension situation reported in Fig.3 where, for simplicity reasons, the point cloud is assumed to be mean centered. Let's consider the point P whose coordinated are $(x_{j1} - \bar{x}_1, x_{j2} - \bar{x}_2)$ in the original reference system and $(y_{j1} - \bar{y}_1, y_{j2} - \bar{y}_2)$ in the new rotated reference system. P' is the projection of P on the axis y_1. The squared length of the segment $\overline{OP}$ is fixed and does not depend on the chosen reference system. $\overline{OP'}$ on the contrary varies as the projection direction varies.

As explained in the preceding sections, Hotelling's approach suggests to

look for the projection direction along which

$$\sum_{j=1}^{n} \overline{OP'_j}^2 = \sum_{j=1}^{n} (y_{j1} - \bar{y}_1)^2 = Dev(y_1)$$

is maximum.

Applying Pythagoras' theorem to the triangle OPP' we have

$$\overline{OP_j}^2 = \overline{P_jP'_j}^2 + \overline{OP'_j}^2$$

and summing over all the points

$$\sum_{j=1}^{n} \overline{OP_j}^2 = \sum_{j=1}^{n} \overline{P_jP'_j}^2 + \sum_{j=1}^{n} \overline{OP'_j}^2$$

As already said $\sum_{j=1}^{n} \overline{OP_j}^2$ is fixed for any given sample; maximizing $\sum_{j=1}^{n} \overline{OP'_j}^2$ is therefore equivalent to minimize $\sum_{j=1}^{n} \overline{P_jP'_j}^2$.
This is precisely Pearson's solution. In his 1901 paper "On lines and planes of closest fit to a system of points" he wanted to find the linear combination that minimizes the sum of squared perpendicular deviations of the points from it. (Note that the least squares line minimizes the sum of the vertical distances instead) We have just seen that Pearson's and Hotelling's approaches, however pursuing different goals, lead to the same solutiion.

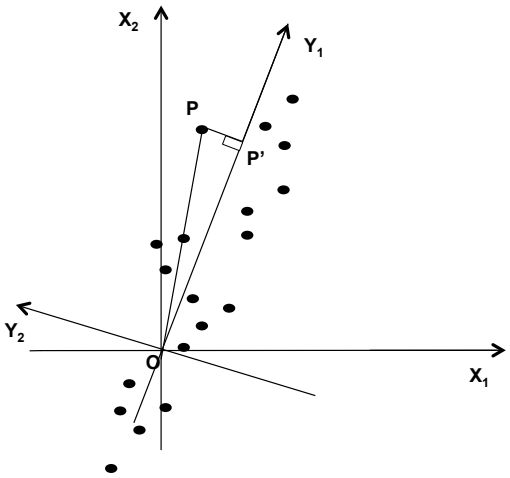

Figure 2: Plot