Statistical Methods 2

Lecture Notes

Content

Notation

For an $n \times p$ matrix $\boldsymbol{M} = [m_{ij}]$ we let

- $m_{(j)} = (m_{1j}, m_{2j}, \ldots, m_{nj})$ denote the jth column of $\boldsymbol{M}$,
- $m_i = (m_{i1}, m_{i2}, \ldots, m_{ip})$ denote the ith row of $\boldsymbol{M}$,
- $\bar{m}_{(j)} = \frac{1}{n}\sum_{i=1}^{n} m_{ij}$ denote the mean of the elements on the jth column of $\boldsymbol{M}$,
- $\bar{m} = (\bar{m}_{(1)}, \ldots, \bar{m}_{(p)})$,
- $\boldsymbol{M}_{1:q} = [m_{(1)} \ldots m_{(q)}]$ for any $q \in \{1, \ldots, p\}$.

Remark that, with this notation in place, we have

$$\boldsymbol{M} = [m_{(1)} \ldots m_{(p)}] = [m_1 \ldots m_n]^\top.$$

We let $\mathbf{1}_n = (1, \ldots, 1) \in \mathbb{R}^n$, $\boldsymbol{I}_n$ be the $n \times n$ identity matrix and

$$\boldsymbol{C}_n = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top.$$

We remark that the matrix $\boldsymbol{Q} := \boldsymbol{C}_n\boldsymbol{M}$ is such that $\bar{q} = (0, \ldots, 0)$ and, for this reason, the matrix $\boldsymbol{C}_n$ is called the centring matrix.

Below we denote by $O(p)$ the set of $p \times p$ orthogonal matrices, that is $\boldsymbol{M} \in O(p)$ if and only if $\boldsymbol{M}^{-1} = \boldsymbol{M}^\top$.

Notation (end)

Throughout this course we denote by $\boldsymbol{X}^0$ an $n \times p$ matrix containing n observations of a p-dimensional variable, and we let $\boldsymbol{X} = \boldsymbol{C}_n \boldsymbol{X}^0$ be the centred data matrix.

Then, we let

- s_{jk} be the sample covariance between the jth and the kth variables, that is

$$s_{jk} = \frac{1}{n} \sum_{i=1}^{n} (x_{ij}^0 - \bar{x}_{(j)}^0)(x_{ik}^0 - \bar{x}_{(k)}^0) = \frac{x_{(j)}^\top x_{(k)}}{n},$$

- $\boldsymbol{S}$ be the sample covariance matrix, that is

$$\boldsymbol{S} = \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X} = [s_{jk}],$$

- s_j be the sample standard deviation of the jth variable, that is

$$s_j = \sqrt{s_{jj}} = \sqrt{\frac{x_{(j)}^\top x_{(j)}}{n}} = \frac{\|x_{(j)}\|}{\sqrt{n}},$$

- r_{jk} be the sample correlation between the jth and kth variables, that is

$$r_{jk} = \frac{\sum_{i=1}^{n} (x_{ij}^0 - \bar{x}_{(j)}^0)(x_{ik}^0 - \bar{x}_{(k)}^0)}{\sqrt{\sum_{i=1}^{n} (x_{ij}^0 - \bar{x}_{(j)}^0)^2 \sum_{i=1}^{n} (x_{ik}^0 - \bar{x}_{(k)}^0)^2}} = \frac{x_{(j)}^\top x_{(k)}}{\|x_{(j)}\| \, \|x_{(k)}\|},$$

- $\boldsymbol{R}$ be the matrix of sample correlations, that is

$$\boldsymbol{R} = \boldsymbol{D}^{-1/2} \boldsymbol{S} \boldsymbol{D}^{-1/2}$$

where $\boldsymbol{D} = \operatorname{diag}(s_1^2, \ldots, s_p^2)$.

Part I

Multivariate Analysis

Chapter 1: Principal Component Analysis[a]

In this chapter we consider an $n \times p$ data matrix $\boldsymbol{X}^0$.

Principal component analysis (PCA) is a dimensionality reduction technique invented by Karl Pearson in 1901, which is used to identify a small number of uncorrelated variables, known as Principal Components (PCs), from a larger set of variables.

There are two types of dimensionality reduction techniques, namely those based on

- Feature (i.e. variable) selection, where a subsect of the original features is selected (as in LASSO regression, see Chapter 7).

- Feature extraction, where a small number of new features, designed to preserve 'most' of the information present in the dataset, are built from the existing set of features.

PCA falls in the second class of dimensionality reduction techniques.

Informally speaking, in PCA the new features are obtained through an orthogonal projection of the data onto a lower dimensional linear subspace, in such a way that the variance of the projected data points is maximized.

[a]The main reference for this chapter is [2, Chapter 8]; see also [1, Section 12.1].

Preliminaries: Interpretation of the correlation matrix $\boldsymbol{R}$

The correlation r_{kl} measures the degree of linear dependence between the kth and the lth variables.

Letting $z_{(j)} = x_{(j)}/s_j$ for $j \in \{k, l\}$, we have

$$r_{kl} = \frac{1}{n} z_{(k)}^\top z_{(l)}$$

so that the correlation r_{kl} can be interpreted as the covariance between the two transformed variables $z_{(k)}$ and $z_{(l)}$.

Letting $\boldsymbol{Z} = \boldsymbol{X}\boldsymbol{D}^{-1/2}$, with $\boldsymbol{D} = \text{diag}(s_1^2, \ldots, s_p^2)$, it follows that

$$\begin{aligned}
\boldsymbol{R} &= \boldsymbol{D}^{-1/2}\boldsymbol{S}\boldsymbol{D}^{-1/2} \\
&= \frac{(\boldsymbol{X}\boldsymbol{D}^{-1/2})^\top \boldsymbol{X}\boldsymbol{D}^{-1/2}}{n} \\
&= \frac{1}{n}\boldsymbol{Z}^\top\boldsymbol{Z} \\
&=: \boldsymbol{S}_Z.
\end{aligned}$$

In words, the correlation matrix $\boldsymbol{R}$ can be interpreted as the sample covariance matrix associated to the transformed data matrix $\boldsymbol{Z}$.

The linear transformation $\boldsymbol{X} \mapsto \boldsymbol{Z} := \boldsymbol{X}\boldsymbol{D}^{-1/2}$ therefore allows to obtain unit variances (since $\boldsymbol{S}_Z = \boldsymbol{R}$).

Transforming the data into uncorrelated data

The objective of PCA is to find a matrix $\boldsymbol{B} \in O(p)$ such that the linear transformation $\boldsymbol{X} \mapsto \boldsymbol{XB}$ achieves uncorrelatedness, i.e. we want $\boldsymbol{B}$ to be such that

$$\boldsymbol{B} \in O(p), \quad \frac{1}{n}(\boldsymbol{XB})^\top(\boldsymbol{XB}) = \boldsymbol{B}^\top \boldsymbol{SB} = \boldsymbol{L} \quad \text{with } \boldsymbol{L} \text{ diagonal.} \tag{1.1}$$

To find a matrix $\boldsymbol{B}$ such that (1.1) holds recall that, since $\boldsymbol{S}$ is a real and symmetric $p \times p$ matrix,

- the eigenvalues of $\boldsymbol{S}$ are all real (see [2, Theorem A.6.3, page 468]); we denote them by $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ in what follows.
- $\boldsymbol{S} = \boldsymbol{A\Lambda A}^\top$ where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_p)$ and where the matrix $\boldsymbol{A} \in O(p)$ has the standardized eigenvectors $\{v_j\}_{j=1}^p$ of $\boldsymbol{S}$ as columns, that is $\boldsymbol{A} = [v_1 \ldots v_p]$ (spectral decomposition theorem, see [2, Theorem A.6.4, page 469]).

Therefore, the transformation $\boldsymbol{X} \mapsto \boldsymbol{Y} := \boldsymbol{XA}$ is such that

$$\frac{1}{n}\boldsymbol{Y}^\top \boldsymbol{Y} = \boldsymbol{A}^\top \boldsymbol{SA} = \boldsymbol{A}^\top \boldsymbol{A\Lambda A}^\top \boldsymbol{A} = \boldsymbol{\Lambda} \tag{1.2}$$

showing that (1.1) holds for $\boldsymbol{B} = \boldsymbol{A}$ and for $\boldsymbol{L} = \boldsymbol{\Lambda}$.

Definition 1.1 *The vector $y_{(j)} := \boldsymbol{X}a_{(j)}$ is called the jth PC.*

Let $\boldsymbol{G}$ be a $p \times p$ diagonal matrix such that $g_{jj} \in \{1, -1\}$ for all j. Then, $\boldsymbol{G} \in O(p)$ and thus, letting $\tilde{\boldsymbol{A}} = \boldsymbol{AG}$, we have

$$\boldsymbol{S} = \boldsymbol{A\Lambda A}^\top = (\boldsymbol{AG})(\boldsymbol{G}^\top \boldsymbol{\Lambda G})(\boldsymbol{AG})^\top = \tilde{\boldsymbol{A}}\big(\boldsymbol{G}^\top \boldsymbol{\Lambda G}\big)\tilde{\boldsymbol{A}}^\top = \tilde{\boldsymbol{A}}\boldsymbol{\Lambda}\tilde{\boldsymbol{A}}^\top$$

where it is easily checked that $\tilde{\boldsymbol{A}} \in O(p)$.

Therefore, noting that $\tilde{a}_{(j)} = g_{jj}a_{(j)}$, the sign of the PCs is arbitrary (and (1.1) holds for any matrix $\boldsymbol{XB}$ of PCs).

Two key properties of PCA

Remark first that $\sum_{i=1}^n y_{ij} = \sum_{i=1}^n (x_i^\top a_{(j)}) = \big(\sum_{i=1}^n x_i\big)^\top a_{(j)} = 0$ so that, by (1.2), the variance $s^2_{y_{(j)}}$ of jth PC, is given by

$$s^2_{y_{(j)}} = \frac{1}{n}\|y_{(j)}\|^2 = \lambda_j. \tag{1.3}$$

Hence, the eigenvalue λ_j is the variance of the jth PC.

Then, a first key property of PCA is that the total variance of the original variables $\{x_{(j)}\}_{j=1}^p$ and of the principal components $\{y_{(j)}\}_{j=1}^p$ coincides. Indeed, since $\mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{S}\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{S}\boldsymbol{A}\boldsymbol{A}^\top) = \mathrm{tr}(\boldsymbol{S})$, we have

$$\sum_{j=1}^p s_j^2 = \mathrm{tr}(\boldsymbol{S}) = \mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{S}\boldsymbol{A}) = \mathrm{tr}(\boldsymbol{\Lambda}) = \sum_{j=1}^p \lambda_j = \sum_{j=1}^p s^2_{y_{(j)}}.$$

A second key property of PCA is given in the following result.

Theorem 1.1 *The matrix $\boldsymbol{A}$ is such that $a_{(1)} \in \mathrm{argmax}_{v\in\mathcal{A}_1} v^\top \boldsymbol{S} v$ with $\mathcal{A}_1 = \{v \in \mathbb{R}^p : \|v\| = 1\}$ and such that, for all $j \in \{2, \ldots, p\}$,*

$$a_{(j)} \in \operatorname*{argmax}_{v\in\mathcal{A}_j} v^\top \boldsymbol{S} v$$

with $\mathcal{A}_j = \{v \in \mathbb{R}^p : \|v\| = 1,\ a_{(1)}^\top v = \cdots = a_{(j-1)}^\top v = 0\}$.

Informally speaking, Theorem 1.1 says that PCA achieves (1.1) for a matrix $\boldsymbol{B} \in O(p)$ that sequentially maximizes the sample variances $\{b_{(j)}^\top \boldsymbol{S} b_{(j)}\}_{j=1}^p$ of the transformed variables $\{\boldsymbol{X} b_{(j)}\}_{j=1}^p$.

Proof of Theorem 1.1

For $a_{(1)}$ we introduce the Lagrange multiplier $l \in \mathbb{R}$ and solve

$$(u_{1*}, l^*) \in \text{argmax}_{(v,l)\in\mathbb{R}^p\times\mathbb{R}}\, v^\top \boldsymbol{S} v + l(1 - v^\top v^\top).$$

By setting the derivatives w.r.t. v and w.r.t. l equal to zero we obtain

$$\boldsymbol{S}u_{1*} = l^* u_{1*} \quad u_{1*}^\top u_{1*} = 1 \tag{1.4}$$

showing that u_{1*} is an eigenvector of $\boldsymbol{S}$ with corresponding eigenvalue l^*. Pre-multiplying both side of (1.4) by $u_{1*}^\top$, and using the fact that $u_{1*}^\top u_{1*} = 1$, yields $u_{1*}^\top \boldsymbol{S} u_{1*} = l^*$, showing that l^* is the largest eigenvalue of $\boldsymbol{S}$, i.e. that $l^* = \lambda_1$, and thus that $u_{1*} = a_{(1)}$.

To prove the result for $j \in \{2, \ldots, p\}$ we introduce the Lagrange multipliers $\{l_k\}_{k=1}^j$ and solve

$$(u_{j*}, l^*) \in \text{argmax}_{(v,l)\in\mathbb{R}^p\times\mathbb{R}^j}\, v^\top \boldsymbol{S} v + l_j(1 - v^\top v^\top) - \sum_{k=1}^{j-1} l_k v^\top a_{(k)}.$$

By setting the derivatives w.r.t. v and w.r.t. l equal to zero we obtain

$$\boldsymbol{S}u_{j*} = l_j^* u_{j*} + \sum_{k=1}^{j-1} l_k^* a_{(k-1)}, \quad u_{j*}^\top u_{j*} = 0, \quad \max_{k\in\{1,\ldots,j-1\}} \left|u_{j*}^\top a_{(k)}\right| = 0. \tag{1.5}$$

Let $m \in \{1, \ldots, j-1\}$. Then, by pre-multiplying both side of (1.5) by $a_{(m)}^\top$, using the fact that $u_{j*}^\top a_{(m)} = 0$ and recalling that $\boldsymbol{S}a_{(m)} = \lambda_m a_{(m)}$, we obtain

$$a_{(m)}^\top \boldsymbol{S} u_{j*} = l_m^* \Leftrightarrow \lambda_m a_{(m)}^\top u_{j*} = l_m^* \Leftrightarrow l_m^* = 0, \quad \forall m \in \{1, \ldots, j-1\}.$$

Together with (1.5) this shows that u_{j*} is an eigenvector of $\boldsymbol{S}$. Pre-multiplying (1.5) by $u_{j*}^\top$ implies that $u_{j*}^\top \boldsymbol{S} u_{j*} = l_j^*$ and thus, since $u_{j*} \neq a_{(m)}$ for all $m \in \{1, \ldots, j-1\}$, it follows that l_j^* is the jth largest eigenvalue of $\boldsymbol{S}$, i.e. that $l_j^* = \lambda_j$, and thus that $u_{j*} = a_{(j)}$. The proof is complete.

Dimensionality reduction using PCA

Since $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{A}$ and $\boldsymbol{A}^{-1} = \boldsymbol{A}^\top$ it follows that $\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{A}^\top$, and thus

$$x_{(j)} = \boldsymbol{Y} a_j = \sum_{k=1}^{p} y_{(k)} a_{jk}, \quad j = 1, \ldots, p. \tag{1.6}$$

Using (1.3) and the fact that the principal components $\{y_{(j)}\}_{j=1}^p$ are orthogonal, (1.6) implies that $s_j^2 = \sum_{k=1}^p a_{jk}^2 \lambda_j$.

Then, to reduce the dimension of $x_{(j)}$ we can choose a $q < p$ and omit the PCs $y_{(q+1)}, \ldots, y_{(p)}$ in the decomposition (1.6) of $x_{(j)}$, as they are those that contribute the least to the variance of this variable. Let

$$\tilde{x}_{(j)} = \sum_{k=1}^{q} y_{(k)} a_{jk} \tag{1.7}$$

denote the resulting q-dimensional approximation of the variable $x_{(j)}$.

Remark now that the approximation $\{\tilde{x}_{(j)}\}_{j=1}^p$ of the original set of variable leads to the approximation $\tilde{x}_i = (\tilde{x}_{i1}, \tilde{x}_{2i}, \ldots, \tilde{x}_{ip})$ of observation x_i, which is still a point in $\mathbb{R}^p$.

However, noting that $\tilde{x}_i = \sum_{j=1}^q a_{(j)} y_{ij}$, if follows that $\tilde{x}_i$ is a point in the q-dimensional linear subspace $\text{span}\{a_{(1)}, \ldots, a_{(q)}\} \subset \mathbb{R}^p$ which can therefore be represented with the q-dimensional vector

$$x_i' := (y_{i1}, \ldots, y_{iq}) \in \mathbb{R}^q \tag{1.8}$$

of coordinates (w.r.t. the orthonormal basis $\{a_{(j)}\}_{j=1}^q$).

We call x_i' the q-dimensional reduction of x_i.

Dimension reduction as an orthogonal projection

Since $\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{A}^\top$ it follows that

$$x_i = \boldsymbol{A}y_i = \sum_{j=1}^{p} a_{(j)} y_{ij}, \quad i = 1, \ldots, n. \tag{1.9}$$

The eigenvectors $\{a_{(j)}\}_{j=1}^{p}$ being orthogonal, it follows from (1.9) that $\Pr_{\langle a_{(1)}, \ldots a_{(q)} \rangle}(x_i)$, the orthogonal projection of observation x_i onto the linear subspace $\text{span}\{a_{(1)}, \ldots, a_{(q)}\}$, is given by

$$\Pr_{\langle a_{(1)}, \ldots a_{(q)} \rangle}(x_i) = \sum_{j=1}^{q} a_{(j)} y_{ij}. \tag{1.10}$$

Therefore, the q-dimensional reduction $x_i' \in \mathbb{R}^q$ of $x_i \in \mathbb{R}^p$ defined in (1.8) is equal to the coordinate of $\Pr_{\langle a_{(1)}, \ldots a_{(q)} \rangle}(x_i)$ with respect to the orthonormal basis $\{a_{(j)}\}_{j=1}^{q}$.

In words, as a dimension reduction technique, PCA amounts (i) to projecting (orthogonally) the observations $\{x_i\}_{i=1}^{n}$ onto the linear subspace $\text{span}\{a_{(1)}, \ldots, a_{(q)}\}$ and (ii) to using the resulting coordinates (w.r.t. $\{a_{(j)}\}_{j=1}^{q}$) as a low dimensional representation of the original observations.

Remark: The matrix $\boldsymbol{X}'$ of the reduced data is $\boldsymbol{X}' = \boldsymbol{Y}_{1:q} = \boldsymbol{X}\boldsymbol{A}_{1:q}$.

Remark: By (1.9) observation x_i has coordinates y_i with respect to the orthonormal basis $\{a_{(j)}\}_{j=1}^{p}$.

Dimensionality reduction: The marks dataset[a]

Let $\boldsymbol{X}^0$ be the matrix containing the examination marks of $n = 88$ students in $p = 5$ subjects, namely mechanics, vectors, algebra, analysis and statistics.

The examination was open book for the first two subjects and closed book for the last three subjects.

Figure 1.1 below shows the two-dimensional reduction $\{x_i'\}_{i=1}^n$ of the marks dataset.

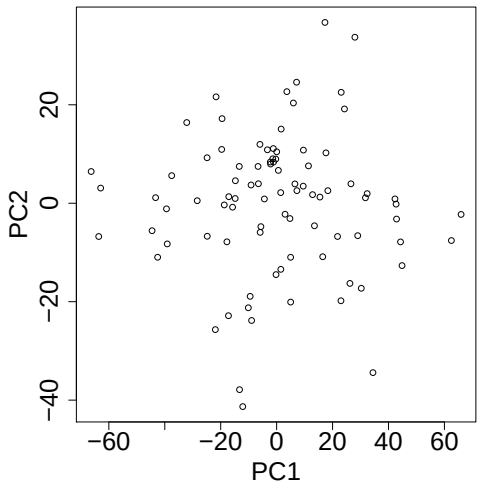

Figure 1.1: Two-dimensional reduction of the marks dataset.

[a]The dataset is from [2] and is available from the `R` package `ggm`.

Choosing the dimension q

On the one hand, we want q small in order to work with low dimensional observations $\{x_i'\}_{i=1}^n$.

On the other hand, q should be large enough so that working with $\{x_i'\}_{i=1}^n$ instead of with the original data points $\{x_i\}_{i=1}^n$ does not result in a loss of information which is 'too important'.

Noting that $\lambda_j / \sum_{k=1}^p \lambda_k$ is the proportion of the total variance accounted for by the jth PC, and thus that $\sum_{j=1}^q \lambda_j / \sum_{j=1}^p \lambda_j$ is the proportion of the total variance accounted for by the first q PCs, the following three approaches are popular in practice for choosing q:

1. Choose an $\eta \in (0, 1]$ and let $q = q_\eta$ where

$$q_\eta = \min \left\{ j : \sum_{k=1}^{j} \lambda_k / \sum_{k=1}^{p} \lambda_k > \eta \right\}. \tag{1.11}$$

2. Graphically display the mapping $j \mapsto \lambda_j$ (Scree plot) and visually identify the PCs having a negligible variance.

3. Let $q = q_{\mathrm{K}}$ where

$$q_{\mathrm{K}} = \max \left\{ j : \lambda_j > \frac{1}{p} \sum_{k=1}^{p} \lambda_k \right\} \quad \text{(Kaiser's criterion)}. \tag{1.12}$$

4. Use Horn's parallel analysis (see below).

Justification of Kaiser's criterion and Horn's parallel analysis

Assume that the x_{ij}'s are realizations of np independent $\mathcal{N}_1(0,1)$ random variables. In this case, as $n \to \infty$ the eigenvalues of the correlation matrix $\boldsymbol{R}$ converge all to one (see Theorem 1.3 below).

This result suggests that if all the PCs are equally informative (i.e. if the original variables are uncorrelated) then all the eigenvalues of $\boldsymbol{R}$ should be around 1. Assuming $\boldsymbol{S} = \boldsymbol{R}$, this suggests that we should take $q < p$ only if $\lambda_j < 1$ for some j and, in particular, that we can consider the jth PC as being little informative if $\lambda_j < 1$. Noting that $\frac{1}{p}\sum_{k=1}^{p} \lambda_k = 1$ if $\boldsymbol{S} = \boldsymbol{R}$, this reasoning leads to the Kaiser's criterion (1.12) for choosing q.

Horn's parallel analysis is an improvement of Kaiser's criterion which takes into account the sampling error which arises from the fact that, in practice, we have a finite number n of observations. More precisely, Horn's parallel analysis amounts to choosing q as follows:

1. Generate M matrices $\{\boldsymbol{X}_m\}_{m=1}^{M}$ of size $n \times p$ and having i.i.d. $\mathcal{N}_1(0,1)$ entries, compute the corresponding correlation matrices $\{\boldsymbol{R}_m\}_{m=1}^{M}$ and, for all m, let $\boldsymbol{S}_m = \boldsymbol{D}^{1/2}\boldsymbol{R}_m\boldsymbol{D}^{1/2}$ with $\boldsymbol{D} = \operatorname{diag}(s_1^2, \ldots, s_p^2)$[a],
2. Compute the eigenvalues $\lambda_1^{(m)} \geq \cdots \geq \lambda_p^{(m)}$ of $\boldsymbol{S}_m$ for $m = 1, \ldots, M$,
3. Compute $\bar{\lambda}_j^{(M)} = \frac{1}{M}\sum_{m=1}^{M} \lambda_j^{(m)}$ for $j = 1, \ldots, p$,
4. Let $q = q_{\mathrm{H}}^{(M)}$ where $q_{\mathrm{H}}^{(M)} = \max\left\{j : \lambda_j > \bar{\lambda}_j^{(M)}\right\}$.

[a] This definition of $\boldsymbol{S}_m$ ensures that $\boldsymbol{S}_m$ and $\boldsymbol{S}$ are on the same scale.

Choosing q: The marks dataset (continued)

The Scree plot for the marks dataset is given in Figure 1.2, which suggests to take $q \in \{1, 2, 4\}$.

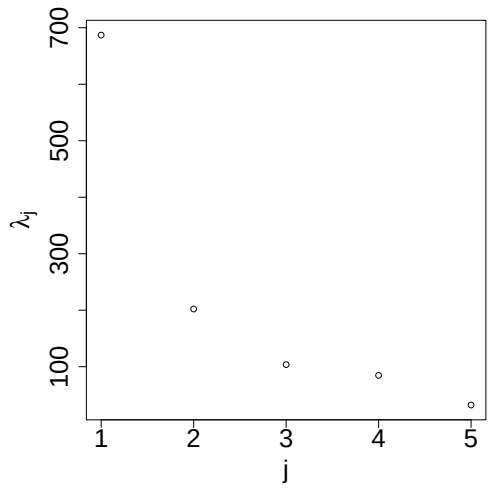

Figure 1.2: Scree plot for the marks dataset.

Using the table below, we see that $q_\eta = 2$ for $\eta = 0.8$ and that $q_\eta = 4$ for $\eta = 0.9$.

k	1	2	3	4	5
$\sum_{j=1}^{k} \lambda_j / \sum_{j=1}^{p} \lambda_j$	0.619	0.801	0.895	0.971	1.000

For this example both the Kaiser criterion and Horn's parallel analysis (with $M = 10\,000$) lead to the choice $q = 1$ (i.e. $q_{\mathrm{K}} = q_{\mathrm{H}}^{(M)} = 1$).

Interpretation of individual PCs

Recall that $y_{(j)} = \boldsymbol{X} a_{(j)}$, i.e. that

$$y_{(j)} = \sum_{k=1}^{p} a_{kj} x_{(k)}. \tag{1.13}$$

Remark: $\{a_{kj}\}_{k=1}^{p}$ are sometimes called the loadings.

We can then interpret the principal component $y_{(j)}$ by looking at the loadings, since a_{kj} measures the "impact" of the variable $x_{(k)}$ in the definition of $y_{(j)}$.

In practice, we often say that $x_{(k)}$ loads $y_{(j)}$ if

$$|a_{kj}| > 0.7 \max\{|a_{lj}|\}_{l=1}^{p}. \tag{1.14}$$

Recall that, by (1.6),

$$x_{(k)} = \sum_{j=1}^{p} a_{kj} y_{(j)}$$

so that the variable $x_{(k)}$ has coordinates $\{a_{kj}\}_{j=1}^{p}$ with respect to $\{y_{(j)}\}_{j=1}^{p}$. Since the PCs are orthogonal, this implies that, for all $J \subset \{1, \ldots, p\}$, the orthogonal projection of $x_{(k)}$ onto the subspace $\text{span}\{y_{(j)},\, j \in J\}$ is given by

$$\Pr_{\langle \{y_{(j)}\}_{j \in J} \rangle}(x_{(k)}) = \sum_{j \in J} a_{kj} y_{(j)} \tag{1.15}$$

and therefore has coordinates $\{a_{kj}\}_{j \in J}$ with respect to the basis $\{y_{(j)},\, j \in J\}$ of the linear subspace $\text{span}\{y_{(j)},\, j \in J\} \subset \mathbb{R}^p$.

Consequently, plotting the projection of the variables $\{x_{(j)}\}_{j=1}^{p}$ onto the linear subspace spanned by $y_{(k)}$ and $y_{(l)}$ allows to visualize which variables load these two PCs.

Interpretation of individual PCs: The marks dataset (continued)

For the marks dataset we obtain the following matrix $\boldsymbol{A}$:

Table 1.1: Matrix $\boldsymbol{A}$ for the marks dataset.

variable	$a_{(1)}$	$a_{(2)}$	$a_{(3)}$	$a_{(4)}$	$a_{(5)}$
mechanics	**-0.505**	**0.749**	-0.300	0.296	-0.079
vectors	-0.368	0.207	0.416	**-0.783**	-0.189
algebra	-0.346	-0.076	0.145	-0.003	**0.924**
analysis	**-0.451**	-0.301	**0.597**	0.518	-0.285
statistics	**-0.535**	**-0.548**	**-0.600**	-0.176	-0.151

Remark: In the table a_{kj} is in bold if (1.14) holds.

Since $a_{j1} < 0$ for all j if follows using (1.13) that $-y_{(1)}$ is a weighted sum of all the variables, and thus that the first PC represents an "average" mark.

Since $a_{j2} > 0$ for $j \in \{1, 2\}$ while $a_{j2} < 0$ for $j \in \{3, 4, 5\}$ it follows that the second PC represents a contrast between the open-book and closed-book examinations (i.e. $|y_{i2}|$ is large if individual i either performs well in closed book exams and poorly in open book exam, or performs poorly on closed book exam and well in open book exams). Notice that PC2 is mainly loaded by the marks obtained in mechanics and statistics

The third PC is essentially a contrast between the analysis and the statistics examinations, while the fourth and the fifth PCs essentially represent the marks obtained in vectors and in algebra, respectively.

Interpretation of individual PCs: The marks dataset (continued)

The following two plots represent the orthogonal projection of the variables $\{x_{(j)}\}_{j=1}^{p}$ onto the space span$\{y_{(1)}, y_{(2)}\}$ (left plot) and onto the space span$\{y_{(2)}, y_{(3)}\}$ (right plot).

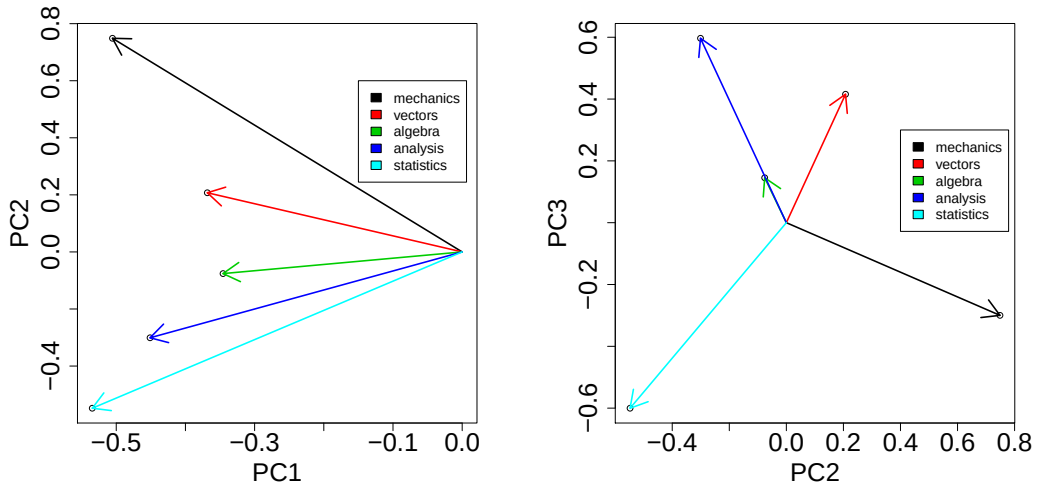

Figure 1.3: Interpretation of the PCs axis for the marks dataset.

As discussed earlier, these two plots allow to visualize which variables load the first three PCs.

For instance, in accordance with the observed matrix $\boldsymbol{A}$ (Table 1.1),

- We see that the variables 'mechanics' and 'statistics' dominate the space span$\{y_{(1)}, y_{(2)}\}$, in the sense that these two variables have the largest coordinates (in absolute value) in both directions. Hence, these two variables load the most the first two PCs.

- In the right plot we see that the variable 'algebra' is almost negligible, and thus does not load $y_{(2)}$ and $y_{(3)}$.

Interpretation of the low dimensional reduction of the observations: The marks dataset (continued)

To interpret the two dimensional reduction of the data represented in Figure 1.1 it is useful to merge Figure 1.1 and Figure 1.3 as follows:

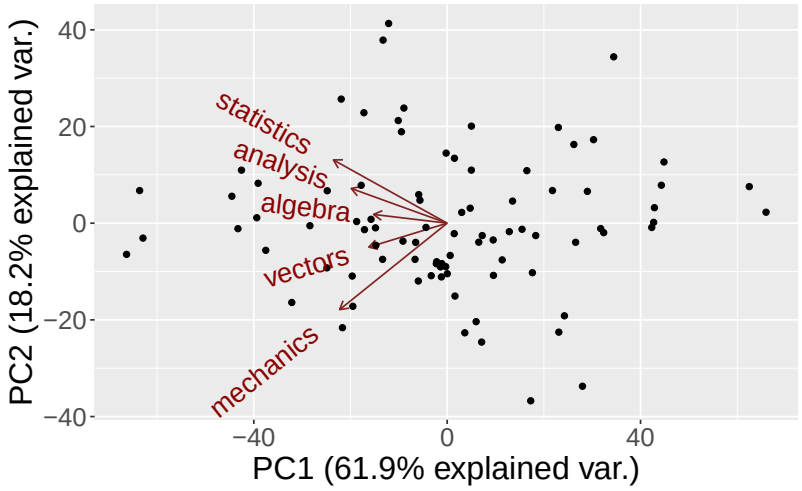

Figure 1.4: Two dimensional reduction of the marks dataset with interpretation of the first two PCs.

Remark: As mentioned above, the sign of the PCs components is arbitrary and in the above plot the y-axis is $-y_{(2)}$ and not $y_{(2)}$ as in Figure 1.1.

From Figure 1.4 we see that the observations

- on the left (resp. right) of the 'vertical origin' are those with a high (resp. low) mark in all subjects (and in particular in statistics and mechanics),
- on the top (resp. bottom) of the 'horizontal origin' are those with a large (resp. low) mark in the closed book exams (notably in statistics and analysis) and a low (resp. high) mark in open book examinations (notably in mechanics).

Non-invariance of PCA to the scaling of the data

Principal component analysis is not scale invariant. More precisely, if $\boldsymbol{K}$ is a $p \times p$ diagonal matrix and if we denote by $\tilde{\boldsymbol{Y}}$ the PCs associated to the matrix $\tilde{\boldsymbol{X}} := \boldsymbol{X}\boldsymbol{K}$ then, in general, $\tilde{\boldsymbol{Y}} \neq \boldsymbol{Y}\boldsymbol{K}$. In words, rescaling the different variables $\{x_{(j)}\}_{j=1}^{p}$ does not amount, in general, to rescaling the PCs $\{y_{(j)}\}_{j=1}^{n}$.

To illustrate this point in Figure 1.5 below we compare the two dimensional reduction of the marks dataset obtained in Figure 1.1 with the one obtained from a rescaled version of this dataset[a].

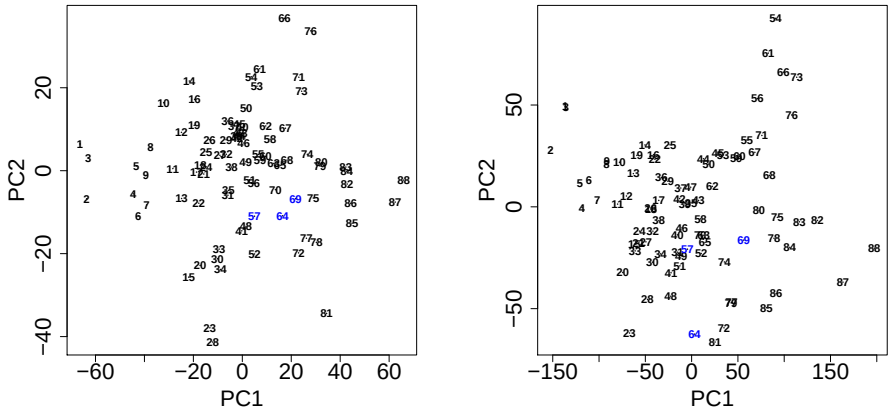

Figure 1.5: Two dimensional reduction of the original (left) and rescaled (right) marks dataset.

We observe in Figure 1.5 that the rescaling of the variables has modified the relative position of the observations. For instance, in the original dataset observation 64 has observations 57 and 69 as nearest neighbours while this is clearly not the case in the rescaled dataset.

$\Rightarrow$ It is therefore important that the original variables are on an appropriate scale before performing PCA.

[a] More precisely, $x_{(1)}$ is multiplied by 2, $x_{(2)}$ by -1, $x_{(3)}$ by 3, $x_{(4)}$ -4 and $x_{(5)}$ by 2.

High Dimensional PCA

Computing the eigenvectors and eigenvalues of $\boldsymbol{S}$ require $\mathcal{O}(p^3)$ operations, making PCA computationally expensive when p is large.

As shown in the following theorem, it is also possible to perform PCA in $\mathcal{O}(n^3)$ operations.

Theorem 1.2 *Assume $p > n$ and let $\tilde{\lambda}_1 \geq \cdots \geq \tilde{\lambda}_n \geq 0$ be the eigenvalues of the matrix $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$, with associated orthonormal eigenvectors $\{\tilde{v}_j\}_{j=1}^n$. Then,*

$$\lambda_j = \tilde{\lambda}_j, \quad a_{(j)} = \frac{\boldsymbol{X}^\top \tilde{v}_j}{(n\tilde{\lambda}_j)^{1/2}}, \quad \forall j \in \{1, \ldots, n\}$$

while $\lambda_j = 0$ for all $j > n$.

This result is useful in applications where p is much larger than n. A typical example where this happens is when the dataset contains a few images, each of them being represented by a vector of potentially several million dimensions (corresponding to three colour values for each pixel).

Remark: Theorem 1.2 shows that $\text{rank}(\boldsymbol{X}\boldsymbol{X}^\top) = \text{rank}(\boldsymbol{X}^\top\boldsymbol{X})$.

Proof of Theorem 1.2

Let $\tilde{\lambda} \geq 0$ be an eigenvalue of the matrix $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$ with associated unitary eigenvector $\tilde{v}$. Then,

$$\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top\tilde{v} = \tilde{\lambda}\tilde{v}$$

and pre-multiplying both side of this equality by $\boldsymbol{X}^\top$ yields

$$\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top\tilde{v} = \tilde{\lambda}\boldsymbol{X}^\top\tilde{v} \Leftrightarrow \boldsymbol{S}\big(\boldsymbol{X}^\top\tilde{v}\big) = \tilde{\lambda}\big(\boldsymbol{X}^\top\tilde{v}\big).$$

Therefore, if $\tilde{\lambda} > 0$ then $\lambda = \tilde{\lambda}$ is a (non-zero) eigenvalue of $\boldsymbol{S}$ with associated unitary eigenvector v defined by

$$v = \frac{\boldsymbol{X}^\top\tilde{v}}{\|\boldsymbol{X}^\top\tilde{v}\|} = \frac{\boldsymbol{X}^\top\tilde{v}}{\big(\tilde{v}^\top(\boldsymbol{X}\boldsymbol{X}^\top\tilde{v})\big)^{1/2}} = \frac{\boldsymbol{X}^\top\tilde{v}}{(n\tilde{\lambda})^{1/2}}.$$

Let $r \leq n$ be the number of non-zero eigenvalues of $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$. Then, the above computations shows that the r non-zero eigenvalues of $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$ allow to obtain r non-zero eigenvalues of $\boldsymbol{S}$. Remark that this shows that $r \leq p$.

Let $r' \leq p$ be the number of non-zero eigenvalues of $\boldsymbol{S}$. Then, to complete the proof it remains to show that $r' = r$.

To this aim let λ be a non-zero eigenvalue of $\boldsymbol{S}$ with associated unitary eigenvector v. Then,

$$\boldsymbol{S}v = \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}v = \lambda v \implies \frac{1}{n}\big(\boldsymbol{X}\boldsymbol{X}^\top\big)\big(\boldsymbol{X}v\big) = \lambda(\boldsymbol{X}v)$$

showing that $\tilde{\lambda} = \lambda$ is a non-zero eigenvalue of $\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^\top$. Hence, $r' = r$ and the proof is complete.

Population principal components

We now assume that the observations $x_1^0, \dots, x_n^0$ are i.i.d. realizations of an $\mathbb{R}^p$-valued random variable X^0 with expectation $\mathbb{E}[X^0] = \mu$ and variance $\mathrm{Var}(X^0) = \boldsymbol{\Sigma}$. Let $X = X^0 - \mu$.

Remark: $\bar{x}^0$ and $\boldsymbol{S}$ are estimators for μ and $\boldsymbol{\Sigma}$, respectively.

Our objective is to find a matrix $\tilde{\boldsymbol{B}}$ such that

$$\tilde{\boldsymbol{B}} \in O(p), \quad \mathrm{Var}(\tilde{\boldsymbol{B}}^\top X) = \tilde{\boldsymbol{L}} \quad \text{with } \tilde{\boldsymbol{L}} \text{ diagonal.} \tag{1.16}$$

Similarly to what we did for finding a matrix $\boldsymbol{B}$ such that (1.1) holds, let

- $\boldsymbol{L} = \mathrm{diag}(l_1, \dots, l_p)$, with $l_1 \geq \cdots \geq l_p$ the eigenvalues of $\boldsymbol{\Sigma}$,
- $\boldsymbol{\Gamma} = [\gamma_{(1)} \dots \gamma_{(p)}]$, with $\{\gamma_{(j)}\}_{j=1}^p$ a set of orthonormal eigenvectors of $\boldsymbol{\Sigma}$ (associated to the eigenvalues $\{l_j\}_{j=1}^p$).

Then, $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{L}\boldsymbol{\Gamma}^\top$ and thus the random variable $Y := \boldsymbol{\Gamma}^\top X$ is such that $\mathrm{Var}(Y) = \boldsymbol{L}$, showing that (1.16) holds for $\tilde{B} = \boldsymbol{\Gamma}$ and $\tilde{\boldsymbol{L}} = \boldsymbol{L}$.

Remark: The jth component of Y, that is the random variable $Y_j = \gamma_{(j)}^\top X$, is called the jth population PC.

As we will see below, we can interpret the observations $\{y_i\}_{i=1}^n$ as "approximate" realizations of Y.

Asymptotic behaviour of eigenvalues and eigenvectors: A result for Gaussian random variables

We assume that $X \sim \mathcal{N}_p(0, \mathbf{\Sigma})$ and below we consider the population version of the matrices $\boldsymbol{A}$ and $\mathbf{\Lambda}$ (i.e. in the definition of these matrices we replace the data points $\{x_i\}_{i=1}^n$ by a collection $\{X_i\}_{i=1}^n$ of i.i.d. $\mathcal{N}_p(0, \mathbf{\Sigma})$ random variables).

Under the above assumption on X we have the following result:

Theorem 1.3 *[2, Theorem 8.3.3, page 230] Assume that $l_1 > \cdots > l_p > 0$ and let λ and ℓ be the vectors containing the diagonal elements of $\mathbf{\Lambda}$ and $\boldsymbol{L}$, respectively. Then,*

$$\sqrt{n}(\lambda - \ell) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_p(0, 2\boldsymbol{L}^2)$$

and, for all $j \in \{1, \ldots, p\}$,

$$\sqrt{n}(a_{(j)} - \gamma_{(j)}) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}_p(0, 2\boldsymbol{V}_j), \quad \boldsymbol{V}_j = l_j \sum_{k\neq j} \frac{l_k}{(l_k - l_j)^2} \boldsymbol{\gamma}_{(k)} \boldsymbol{\gamma}_{(j)}^\top$$

where the elements of λ are asymptotically independent of those in $\boldsymbol{A}$.

In words, under the assumptions of the theorem the eigenvalues and eigenvectors of $\boldsymbol{S}$ are consistent estimators of those of $\mathbf{\Sigma}$.

Recalling that $y_i = \boldsymbol{A}^\top x_i$ while $Y = \mathbf{\Gamma}^\top X$, Theorem 1.3 implies that (under some assumptions) y_i can be interpreted as an approximate realization of Y (approximate in the sense that, in the definition of y_i, instead of having $\mathbf{\Gamma}$ we have a consistent estimator of this matrix).

Application 1 of PCA: Data visualization

An important application of PCA is to visualize $p > 2$ dimensional observations in a meaningful way, with the aim of detecting which observations are similar to others and to identify potential clusters.

Below we illustrate this use of PCA using the mtcars dataset[a].

This dataset provides information on $n = 32$ models of car. For each car in the dataset there are 9 continuous features and two categorical features, expressed in varying US units.

It is important to realize that PCA should be applied with care when the dataset contains categorical variables, since the numerical value associated to a given category (i) is arbitrary and (ii) impacts the outcome of PCA (recall that PCA is not scale invariant). For this reason, in what follows we keep only the continuous features.

The resulting $p = 9$ variables are not all expressed in the same unit. In such situations, a standard practise that we adopt below is to standardize the variables before applying PCA, that is to perform PCA using the correlation matrix $\boldsymbol{R}$ and not the covariance matrix $\boldsymbol{S}$.

[a]This dataset is built in `R` and what follows is taken from `https://www.datacamp.com/community/tutorials/pca-analysis-r`

Data visualization: The mtcars dataset (continued)

Figure 1.6 below presents the two-dimensional reduction of the mtcars dataset

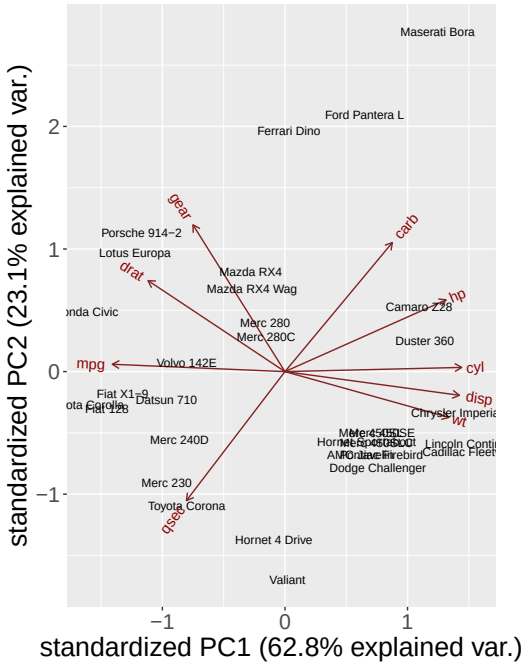

Figure 1.6: Two-dimensional reduction of the mtcars dataset.

As indicated in the plot, the first two PCs explain 89.5% of the variance of the observations, which suggests that most information present in the dataset is preserved in the above 2 dimension reduction of the observations.

From Figure 1.6 we see for instance that the *Maserati Bora*, the *Ferrari Dino* and the *Ford Pantera L* form a cluster (located at the top of the plot), which makes sense since they are all sports cars. The position of these cars on the plot indicates that they have a high value for the variables 'gear' (number of forward gears), 'carb' (number of carburettors), 'hp' (gross horsepower), 'cyl' (number of cylinders), 'disp' (displacement) and 'wt' (weight).

Data visualization: The mtcars dataset (continued)

In order to see if cars with different origins (Europe, Japan and US) tend to be different we can reproduce the above plot making apparent the origin of each car.

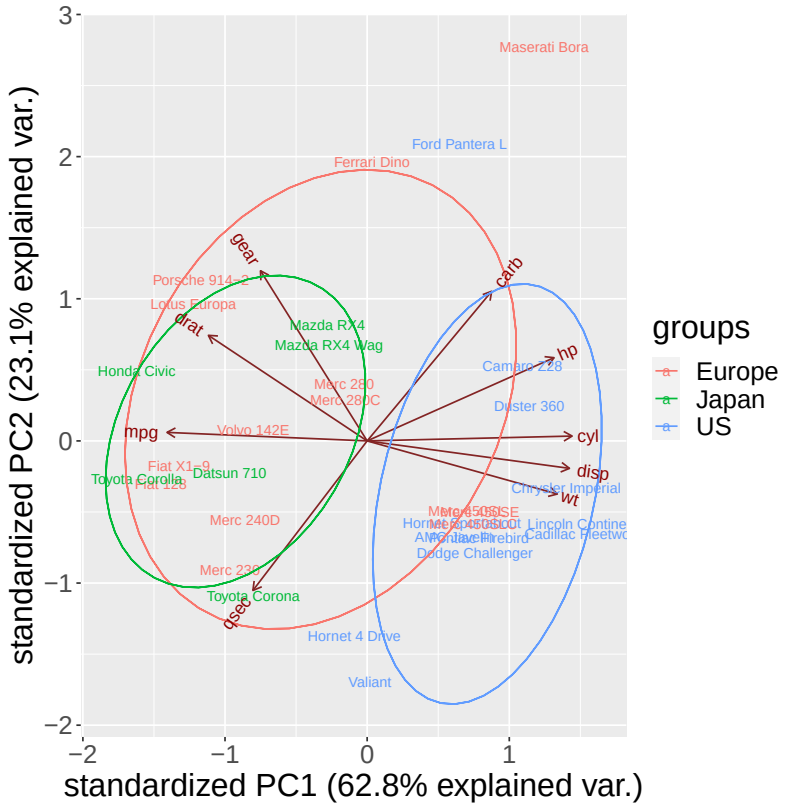

Figure 1.7: Visual clustering of the mtcars dataset.

From Figure 1.7 we see that US cars form a cluster on the right of the plot (characterized by a high value of the variables 'hp', 'cyl', 'disp' and 'wt'), while the Japanese cars form a cluster on the left of the plot (characterized by a high value of the variables 'mpg').

Data visualization: The mtcars dataset

Finally, to assess if the other PCs provide any meaningful information about the data we reproduce the latter plot but this time for the third and fourth PC.

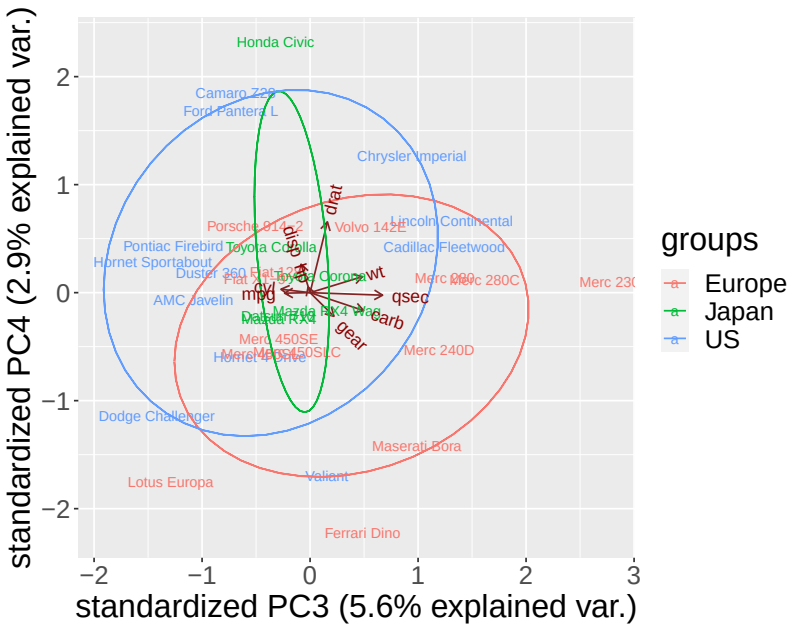

Figure 1.8: Third and fourth PCs for the mtcars dataset.

We do not see any interesting pattern is this plot, which is not surprising since the variance explained by $y_{(3)}$ and $y_{(4)}$ is quite low (i.e. most of the variance in the data is explained by $y_{(1)}$ and $y_{(2)}$, as shown above).

Application 2 of PCA: Dimension reduction in regression

We consider n data points $\{(z_i, x_i^0)\}_{i=1}^n$ in $\mathbb{R}^d \times \mathbb{R}^p$ that we model using

$$Z_i \sim f\big(z_i; g(\alpha + \beta^\top x_i^0)\big), \quad i = 1, \dots, n \tag{1.17}$$

where

- $g : \mathbb{R} \to \Theta \subseteq \mathbb{R}^m$ is a known function,
- $\{f(\cdot;\theta),\, \theta \in \Theta\}$ is a collection of p.d.f. on $\mathbb{R}^d$,
- $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$ are parameters to estimate.

We first show how the parameters α and β can be estimated from the principal components $\{y_{(j)}\}_{j=1}^p$.

To this aim recall that $y_i = \boldsymbol{A}^\top x_i$ and let $h : \mathbb{R}^{1+p} \to \mathbb{R}^{1+p}$ be such that $h(\alpha, \beta) = (\alpha + \beta^\top \bar{x}^0, \boldsymbol{A}^\top \beta)$ for all $(\alpha, \beta) \in \mathbb{R}^{1+p}$.

Remark that because the matrix $\boldsymbol{A}$ is invertible the function h is one-to-one. Therefore, noting that if $(a, \gamma) = h(\alpha, \beta)$ then

$$\alpha + \beta^\top x_i^0 = (\alpha + \beta^\top \bar{x}^0) + \beta^\top x_i = a + \gamma^\top \boldsymbol{A}^\top x_i = a + \gamma^\top y_i,$$

it follows that the model (1.17) is equivalent to the model

$$Z_i \sim f\big(z_i; g(a + \gamma^\top y_i)\big), \quad i = 1, \dots, n. \tag{1.18}$$

Using the invariance property of the maximum likelihood estimator (MLE), this implies that if $(\hat{a}, \hat{\gamma})$ is the MLE of (a, γ) in the model (1.18) then $h^{-1}\big(\hat{a}, \hat{\gamma}\big) = (\hat{a} - \hat{\gamma}^\top \boldsymbol{A}^\top \bar{x}^0, \boldsymbol{A}\hat{\gamma})$ is the MLE of (α, β) in the model (1.17).

Remark: By using the orthogonal PCs instead of the original correlated variables $\{x_{(j)}^0\}_{j=1}^p$, this approach for computing the MLE (α, β) may improve the numerical accuracy of the estimates.

Principal component regression

We now consider the problem of estimating α and β using only the first $q \leq p$ principal components.

To this aim, for all $u \in \mathbb{R}^p$ let $y_n^{(q)}(u) = \boldsymbol{A}_{1:q}^\top u$ be the (coordinate of the) projection of u onto the space $\text{span}\{a_{(1)}, \ldots, a_{(q)}\}$, and let $(\tilde{a}, \tilde{\gamma})$ be an estimate (e.g. the MLE) of (a, γ) in the model

$$Z_i \sim f\Big(z_i; g\big(a + \gamma^\top y_n^{(q)}(x_i)\big)\Big), \quad i = 1, \ldots, n \tag{1.19}$$

where $a \in \mathbb{R}$ and $\gamma \in \mathbb{R}^q$.

Remark: We have $(y_{i1}, \ldots, y_{iq}) = y_n^{(q)}(x_i)$ for $i = 1, \ldots, n$.

Then, given the estimate $(\tilde{a}, \tilde{\gamma})$ of (a, γ), it is reasonable to estimate β using a $\tilde{\beta} \in \mathbb{R}^p$ such that

$$\tilde{\gamma}^\top y_n^{(q)}(x_i) = \tilde{\beta}^\top x_i, \quad \forall i \in \{1, \ldots, n\} \Leftrightarrow \boldsymbol{Y}_{1:q}\tilde{\gamma} = \boldsymbol{X}\tilde{\beta}. \tag{1.20}$$

Since $\boldsymbol{Y}_{1:q} = \boldsymbol{X}\boldsymbol{A}_{1:q}$, (1.20) is equivalent to $\boldsymbol{X}\boldsymbol{A}_{1:q}\tilde{\gamma} = \boldsymbol{X}\tilde{\beta}$ and an estimate $\tilde{\beta}$ of β such that (1.20) holds is therefore given by

$$\tilde{\beta} = \boldsymbol{A}_{1:q}\tilde{\gamma}. \tag{1.21}$$

Then, it is natural to estimate α using $\tilde{\alpha} \in \mathbb{R}$ such that

$$\tilde{\alpha} + \tilde{\beta}^\top x_i^0 = \tilde{a} + \tilde{\gamma}^\top y_n^{(q)}(x_i), \quad \forall i \in \{1, \ldots, n\},$$

that is to estimate α using

$$\tilde{\alpha} = \tilde{a} - \tilde{\beta}^\top \bar{x}^0. \tag{1.22}$$

This technique for estimating (α, β) is called principal component regression (PCR).

Prediction using the PCR estimator

Assume now that we want to use the model fitted by PCR to predict the value of Z associated to an $u \in \mathbb{R}^p$, that is we want to predict Z using the distribution

$$f\Big(z; g\big(\tilde{\alpha} + \tilde{\beta}^\top u\big)\Big).$$

To this aim remark that

$$\tilde{\gamma}^\top y_n^{(q)}(u - \bar{x}^0) = \tilde{\gamma}^\top \boldsymbol{A}_{1:q}^\top (u - \bar{x}^0) = \tilde{\beta}^\top (u - \bar{x}^0)$$

which, recalling that $\tilde{\alpha} = \tilde{a} - \tilde{\beta}^\top \bar{x}^0$, shows that

$$f\Big(z; g\big(\tilde{\alpha} + \tilde{\beta}^\top u\big)\Big) = f\Big(z; g\big(\tilde{a} + \tilde{\gamma}^\top y_n^{(q)}(u - \bar{x}^0)\big)\Big).$$

In words, given an estimate $(\tilde{a}, \tilde{\gamma})$ of the model (1.23), for every $u \in \mathbb{R}^p$ it is equivalent

- to predict Z from u using the model (1.17) with the corresponding PCR estimated parameters value $(\tilde{\alpha}, \tilde{\beta})$ defined in (1.21)-(1.22).
- to predict Z from from $y_n^{(q)}(u - \bar{x}^0)$ using the model (1.23) with the parameters $(\tilde{a}, \tilde{\gamma})$.

Principal component regression with rescaled variables

As mentioned earlier, the PCs are not scale invariant and, when the variables are not all expressed in the same scale, PCA is typically performed using the correlation matrix $\boldsymbol{R}$ rather than the covariance matrix $\boldsymbol{S}$.

In this case, the PCs are given by $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{D}^{-1/2}\boldsymbol{A}$ and thus, in the context of PCR, using $\boldsymbol{R}$ instead of $\boldsymbol{S}$ to compute the PCs amounts to replacing the function $y_n^{(q)}$ defined above by the function $\check{y}_n^{(q)}$ defined by

$$\check{y}_n^{(q)}(u) = \boldsymbol{A}_{1:q}^\top \boldsymbol{D}^{1/2} u, \quad u \in \mathbb{R}^p.$$

Then, as per above, given an estimate $(\check{a}, \check{\gamma})$ of the model

$$Z_i \sim f\Big(z_i; g\big(a + \gamma^\top \check{y}_n^{(q)}(x_i)\big)\Big), \quad i = 1, \ldots, n \tag{1.23}$$

it is reasonable to estimate (α, β) using a $(\check{\alpha}, \check{\beta}) \in \mathbb{R}^{1+p}$ such that

$$\check{a} + \check{\gamma}^\top \check{y}_n^{(q)}(x_i) = \check{\alpha} + \check{\beta}^\top x_i^0, \quad \forall i \in \{1, \ldots, n\}$$

and thus a natural estimate of (α, β) is given by

$$\check{\beta} = \boldsymbol{D}^{-1/2}\boldsymbol{A}_{1:q}\check{\gamma}, \quad \check{\alpha} = \check{a} - \check{\beta}^\top \bar{x}^0.$$

Remark: We have $(\check{\alpha}, \check{\beta}) = (\tilde{\alpha}, \tilde{\beta})$ when $\boldsymbol{D} = \boldsymbol{I}_p$.

Principal component regression: The MNIST dataset[a]

The MNIST dataset contains 70 000 images, each of them having 28 by 28 pixels. Hence, this dataset contains $p = 28 \times 28 = 784$ features.

Each feature is an integer between 0 and 255. To avoid numerical problems that may arise when the covariates can have a large value we let $\boldsymbol{X}^0$ be the matrix of features divided by 255 (so that $x_{ij}^0 \in [0, 1]$).

Each image represents an integer in the set $\{0, \ldots, 9\}$ and our goal is to build a model to predict the number $Z(u)$ in an image associated to an $u \in \mathbb{R}^p$.

In what follows we consider a very simple approach[b] for performing this task, which consists in the following three steps:

1. In a first step we fit 10 logistic regression models $\mathcal{M}_0, \ldots, \mathcal{M}_9$, where $\mathcal{M}_v$ is used to model the probability that the number in the image is $v \in \{0, \ldots, 9\}$ when we observe $u \in \mathbb{R}^p$. Let $\{\widehat{\mathcal{M}}_v\}_{v=0}^9$ denotes the 10 fitted models.
2. In a second step, for a given $u \in \mathbb{R}^p$, for all $v \in \{0, \ldots, 9\}$ we compute $\hat{p}_v(u)$, the probability that $Z(u) = v$ under the model $\widehat{\mathcal{M}}_v$.
3. For a given $u \in \mathbb{R}^p$ we let $\text{argmax}_{v \in \{0,\ldots,9\}} \hat{p}_v(u)$ be our prediction for $Z(u)$.

The dataset is split into a training set of $n = 60\,000$ examples, used to estimate the models $\{\mathcal{M}_v\}_{v=0}^9$, and a test set of 10 000 examples, used to assess the out-of-sample performance of the approach described above.

[a] The dataset can be downloaded in `R` using the package `mnist`.
[b] This approach is taken from: `https://rpubs.com/kstahl/MNIST-1`.

Principal component regression: The MNIST dataset (continued)

Since the variables $\{x_{(j)}\}_{j=1}^{p}$ are on a similar scale PCA is performed using the covariance matrix $\boldsymbol{S}$.

As we can observe in Figure 1.9, in this example we can take $q \ll p$ without losing much in term of total variance.

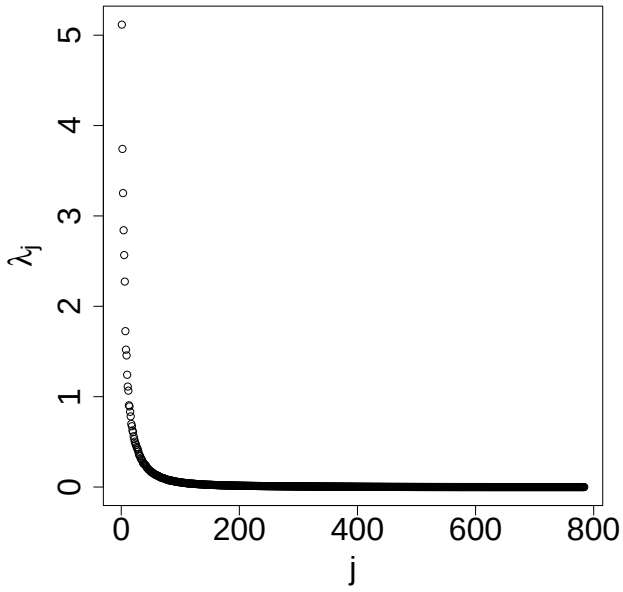

Figure 1.9: Scree plot for the MNIST dataset.

Below, the 10 logistic regression models are estimated using PCR with $q = q_\eta$, with q_η as defined in (1.11) and for different values of $\eta \in (0, 1)$.

Principal component regression: The MNIST dataset

The results of this experiment are summarized in the following table.

η	q_η	Accuracy	Reg. time	Total time
0.8	44	0.9022	15.521	17.855
0.85	59	0.9088	25.492	28.491
0.9	87	0.9133	53.868	58.157
0.95	154	0.9180	162.445	169.656
0.99	331	0.9204	897.522	912.808

In the table, the column "Accuracy" gives the proportion of correct predictions in the test set, the column "Reg. time" the time (in second) needed to perform the 10 logistic regressions and "Total time" gives the total time (in second) of the procedure (which includes the PCA and the computation of the out-of-sample predictions).

We observe that with only 44 PCs the accuracy of the model is close to that obtained with 331 PCs.

Remark: The algorithm used to estimate the model did not converge when all the $p = 784$ features are used.

References

[1] Bishop, C. M. (2006). Pattern recognition. *Machine learning*, 128(9).

[2] Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate analysis*. Probability and mathematical statistics. Academic Press Inc.