Principal Component Analysis: application to Statistical Process Control

Gilbert Saporta, Ndeye Niang Keita

Chapter 1

Principal Component Analysis: Application to Statistical Process Control

1.1. Introduction

Principal component analysis (PCA) is an exploratory statistical method for graphical description of the information present in large datasets. In most applications, PCA consists of studying p variables measured on n individuals. When n and p are large, the aim is to synthesize the huge quantity of information into an easy and understandable form.

Unidimensional or bidimensional studies can be performed on variables using graphical tools (histograms, box plots) or numerical summaries (mean, variance, correlation). However, these simple preliminary studies in a multidimensional context are insufficient since they do not take into account the eventual relationships between variables, which is often the most important point.

Principal component analysis is often considered as the basic method of factor analysis, which aims to find linear combinations of the p variables called components used to visualize the observations in a simple way. Because it transforms a large number of correlated variables into a few uncorrelated principal components, PCA is a dimension reduction method. However, PCA can also be used as a multivariate outlier detection method, especially by studying the last principal components. This property is useful in multidimensional quality control.

Chapter written by Gilbert SAPORTA and Ndèye NIANG.

1.2. Data table and related subspaces

1.2.1. *Data and their characteristics*

Data are generally represented in a rectangular table with n rows for the individuals and p columns corresponding to the variables. Choosing individuals and variables to analyze is a crucial phase which has an important influence on PCA results. This choice has to take into account the aim of the study; in particular, the variables have to describe the phenomenon to analyze.

Usually PCA deals with numerical variables. However, ordinal variables such as ranks can be also processed by PCA. Later in this chapter, we present the concept of supplementary variables which afterwards integrates nominal variables.

1.2.1.1. *Data table*

Let $\mathbf{X}$ be the (n, p) matrix of observations:

$$\mathbf{X} = \begin{pmatrix} x_1^1 & \dots & x_1^p \\ \vdots & \vdots & \vdots \\ x_i^1 & x_i^j & x_i^p \\ \vdots & \vdots & \vdots \\ x_n^1 & \dots & x_n^p \end{pmatrix}$$

where x_i^j is the value of individual i for variable j (denoted $\mathbf{x}^j$) which is identified with a vector of n components $(x_1^j, \dots, x_n^j)'$. In a similar way, an individual i is identified to a vector $\boldsymbol{x}_i$ of p components with $\boldsymbol{x}_i = (x_i^1, \dots, x_i^p)'$.

Table 1.1 is an example of such a data matrix. Computations have been carried out with SPAD 5 software, version 5 [1], kindly provided by J.-P. Gauchi.

The data file contains 57 brands of mineral water described by 11 variables defined in Table 1.2. The data come from the bottle labels. Numerical variables are homogenous, they are all active variables (see section 1.4.3). A variable of a different kind such as price would be considered as a supplementary variable. On the other hand, qualitative variables such as country, type and whether still or sparkling (PG) are necessarily supplementary variables.

1. DECISIA (former CISIA-CERESTA), Building Hoche, 13 rue Auger, 93697 Pantin cedex.

Name	Country	Type	PG	CA	MG	NA	K	SUL	NO3	HCO3	CL
Evian	F	M	P	78	24	5	1	10	3.8	357	4.5
Montagne des Pyrénées	F	S	P	48	11	34	1	16	4	183	50
Cristaline-St-Cyr	F	S	P	71	5.5	11.2	3.2	5	1	250	20
Fiée des Lois	F	S	P	89	31	17	2	47	0	360	28
Volcania	F	S	P	4.1	1.7	2.7	0.9	1.1	0.8	25.8	0.9
Saint Diéry	F	M	G	85	80	385	65	25	1.9	1350	285
Luchon	F	M	P	26.5	1	0.8	0.2	8.2	1.8	78.1	2.3
Volvic	F	M	P	9.9	6.1	9.4	5.7	6.9	6.3	65.3	8.4
Alpes/Moulettes	F	S	P	63	10.2	1.4	0.4	51.3	2	173.2	1
Orée du bois	F	M	P	234	70	43	9	635	1	292	62
Arvie	F	M	G	170	92	650	130	31	0	2195	387
Alpes/Roche des Ecrins	F	S	P	63	10.2	1.4	0.4	51.3	2	173.2	10
Ondine	F	S	P	46.1	4.3	6.3	3.5	9	0	163.5	3.5
Thonon	F	M	P	108	14	3	1	13	12	350	9
Aix les Bains	F	M	P	84	23	2	1	27	0.2	341	3
Contrex	F	M	P	486	84	9.1	3.2	1187	2.7	403	8.6
La Bondoire Saint Hippolite	F	S	P	86	3	17	1	7	19	256	21
Dax	F	M	P	125	30.1	126	19.4	365	0	164.7	156
Quézac	F	M	G	241	95	255	49.7	143	1	1685.4	38
Salvetat	F	M	G	253	11	7	3	25	1	820	4
Stamna	GRC	M	P	48.1	9.2	12.6	0.4	9.6	0	173.3	21.3
Iolh	GR	M	P	54.1	31.5	8.2	0.8	15	6.2	267.5	13.5
Avra	GR	M	P	110.8	9.9	8.4	0.7	39.7	35.6	308.8	8
Rouvas	GRC	M	P	25.7	10.7	8	0.4	9.6	3.1	117.2	12.4
Alisea	IT	M	P	12.3	2.6	2.5	0.6	10.1	2.5	41.6	0.9
San Benedetto	IT	M	P	46	28	6.8	1	5.8	6.6	287	2.4
San Pellegrino	IT	M	G	208	55.9	43.6	2.7	549.2	0.45	219.6	74.3
Levissima	IT	M	P	19.8	1.8	1.7	1.8	14.2	1.5	56.5	0.3
Vera	IT	M	P	36	13	2	0.6	18	3.6	154	2.1
San Antonio	IT	M	P	32.5	6.1	4.9	0.7	1.6	4.3	135.5	1
La Française	F	M	P	354	83	653	22	1055	0	225	982
Saint Benoit	F	S	G	46.1	4.3	6.3	3.5	9	0	163.5	3.5
Plancoët	F	M	P	36	19	36	6	43	0	195	38
Saint Alix	F	S	P	8	10	33	4	20	0.5	84	37
Puits Saint Georges/Casino	F	M	G	46	33	430	18.5	10	8	1373	39
St-Georges/Corse	F	S	P	5.2	2.43	14.05	1.15	6	0	30.5	25
Hildon bleue	B	M	P	97	1.7	7.7	1	4	26.4	236	16
Hildon blanche	B	M	G	97	1.7	7.7	1	4	5.5	236	16
Mont Roucous	F	M	P	1.2	0.2	2.8	0.4	3.3	2.3	4.9	3.2
Ogeu	F	S	P	48	11	31	1	16	4	183	44
Highland Spring	B	M	P	35	8.5	6	0.6	6	1	136	7.5
Parot	F	M	G	99	88.1	968	103	18	1	3380.51	88
Vernière	F	M	G	190	72	154	49	158	0	1170	18
Terres de Flein	F	S	P	116	4.2	8	2.5	24.5	1	333	15
Courmayeur	IT	M	P	517	67	1	2	1371	2	168	1
Pyrénées	F	M	G	48	12	31	1	18	4	183	35
Puits Saint Georges/Monoprix	F	M	G	46	34	434	18.5	10	8	1373	39
Prince Noir	F	M	P	528	78	9	3	1342	0	329	9
Montcalm	F	S	P	3	0.6	1.5	0.4	8.7	0.9	5.2	0.6
Chantereine	F	S	P	119	28	7	2	52	0	430	7
18 Carats	F	S	G	118	30	18	7	85	0.5	403	39
Spring Water	B	S	G	117	19	13	2	16	20	405	28
Vals	F	M	G	45.2	21.3	453	32.8	38.9	1	1403	27.2
Vernand	F	M	G	33.5	17.6	192	28.7	14	1	734	6.4
Sidi Harazem	MO	S	P	70	40	120	8	20	4	335	220
Sidi Ali	MO	M	P	12.02	8.7	25.5	2.8	41.7	0.1	103.7	14.2
Montclar	F	S	P	41	3	2	0	2	3	134	3

Table 1.1. *Data table*

1.2.1.2. *Summaries*

1.2.1.2.1. Centroid

Let $\overline{\boldsymbol{x}}$ be the vector of arithmetic means of each of the p variables, defining the centroid:

$$\overline{\boldsymbol{x}} = (\overline{x}^1, \ldots, \overline{x}^p)'$$

Name	Complete water name as labeled on the bottle
Country	Identified by the official car registration letters; sometimes it is necessary to add a letter, for example Crete: GRC (Greece Crete)
Type	M for mineral water, S for spring water
PG	P for still water, G for sparkling water
CA	Calcium ions (mg/litre)
MG	Magnesium ions (mg/litre)
NA	Sodium ions (mg/litre)
K	Potassium ions (mg/litre)
SUL	Sulfate ions (mg/litre)
NO3	Nitrate ions (mg/litre)
HCO3	Carbonate ions (mg/litre)
CL	Chloride ions (mg/litre)

Table 1.2. *Variable description*

where $\overline{x}^j = \sum_{i=1}^n p_i x_i^j$.

If the data are collected following a random sampling, the n individuals all have the same importance in the computations of the sample characteristics. The same weight $p_i = 1/n$ is therefore allocated to each observation.

However, it can be useful for some applications to use weight p_i varying from one individual to another as grouped data or a reweighted sample. These weights, which are positive numbers adding to 1, can be viewed as frequencies and are stored in a diagonal matrix of size n:

$$\mathbf{D}_p = \begin{pmatrix} p_1 & & \\ & \ddots & \\ & & p_n \end{pmatrix}.$$

We then have the following matrix expressions: $\overline{\boldsymbol{x}} = \mathbf{X}'\mathbf{D}_p\mathbf{1}_n$ where $\mathbf{1}_n$ represents the vector of $\mathbb{R}^n$ with all its components equal to 1. The centered data matrix associated with $\mathbf{X}$ is then $\mathbf{Y}$ with $y_i^j = x_i^j - \overline{x}^j$ and $\mathbf{Y} = \mathbf{X} - \mathbf{1}_n\overline{\boldsymbol{x}}' = (\mathbf{I}_n - \mathbf{1}_n\mathbf{1}_n'\mathbf{D}_p)\mathbf{X}$, where $\mathbf{I}_n$ is the unity matrix of dimension n.

1.2.1.2.2. Covariance matrix and correlation matrix

Let $s_j^2 = \sum_{i=1}^n p_i(x_i^j - \overline{x}^j)^2$ and $s_{k\ell} = \sum_{i=1}^n p_i(x_i^k - \overline{x}^k)(x_i^\ell - \overline{x}^\ell)$, the variance of variable j and the covariance between variables k and ℓ, respectively. They are stored in the covariance matrix $\mathbf{S} = \mathbf{X}'\mathbf{D}_p\mathbf{X} - \overline{\boldsymbol{x}}\overline{\boldsymbol{x}}' = \mathbf{Y}'\mathbf{D}_p\mathbf{Y}$.

We define the linear correlation coefficient between variables k and ℓ by:

$$r_{k\ell} = \frac{s_{k\ell}}{s_k s_\ell}.$$

If $\mathbf{Z}$ is the standardized data table associated with $\mathbf{X}$, $z_i^j = (x_i^j - \overline{x}^j)/s_j$, we have $\mathbf{Z} = \mathbf{Y}\mathbf{D}_{1/s}$ where $\mathbf{D}_{1/s}$ the diagonal matrix of the inverse of standard deviations:

$$\mathbf{D}_{1/s} = \begin{pmatrix} 1/s_1 & & \\ & \ddots & \\ & & 1/s_p \end{pmatrix}.$$

$\mathbf{R}$ is the correlation matrix containing the linear correlation coefficients between all pairs of variables; we have $\mathbf{R} = \mathbf{D}_{1/s}\mathbf{S}\mathbf{D}_{1/s}$. $\mathbf{R}$ is the covariance matrix of standardized variables. It summarizes linear dependency structure between the p variables.

Tables 1.3 and 1.4 list the numerical summaries associated with the dataset example.

Variable	Mean	Standard deviation	Minimum	Maximum
CA	102.46	118.92	1.20	528.00
MG	25.86	28.05	0.20	95.00
NA	93.85	195.51	0.80	968.00
K	11.09	24.22	0.00	130.00
SUL	135.66	326.31	1.10	1371.00
NO3	3.83	6.61	0.00	35.60
HCO3	442.17	602.94	4.90	3380.51
CL	52.47	141.99	0.30	982.00

Table 1.3. *Simple statistics for continuous variables*

	CA	MG	NA	K	SUL	NO3	HCO3	CL
CA	1.00							
MG	0.70	1.00						
NA	0.12	0.61	1.00					
K	0.13	0.66	0.84	1.00				
SUL	0.91	0.61	0.06	–0.03	1.00			
NO3	–0.06	–0.21	–0.12	–0.17	–0.16	1.00		
HCO3	0.13	0.62	0.86	0.88	–0.07	–0.06	1.00	
CL	0.28	0.48	0.59	0.40	0.32	–0.12	0.19	1.00

Table 1.4. *Correlation matrix*

1.2.2. *The space of statistical units*

The Pearson geometrical approach is based on a data cloud associated with the observations: each unit defined by p coordinates is then considered as an element of a vector space of p dimensions, referred to as the space of statistical units. The centroid $\overline{\boldsymbol{x}}$ defined in section 1.2.1.2 is then the barycenter of the data cloud.

PCA consists of visualizing the most reliable data cloud possible within a space of a few dimensions. **[AQ: Have reworded; please confirm correct]** The analysis is based on distances between points representing the individuals. The method by which

these distances are computed influences the results to a large extent. It is therefore essential to determine it before any. **[AQ: Is some text missing?]**

1.2.2.1. *'The metric'*

In the usual 3D physical space, computing a distance is simple using Pythagoras' formula. However, in statistics, the problem is more complicated: how can distances between individuals described by variables having measurement units as different as euros, kg, km, etc. be calculated? The Pythagoras formula is as arbitrary as any other. The following general formulation therefore has to be used: $\mathbf{M}$ is a positive definite symmetric matrix of size p and the distance between two individual $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is defined by the quadratic form:

$$d^2(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \boldsymbol{x}_j)'\mathbf{M}(\boldsymbol{x}_i - \boldsymbol{x}_j) = d^2(i, j).$$

In theory, the choice of $\mathbf{M}$ depends on the user who is the only one to precisely determine the adequate metric. In practice, however, the usual metrics in PCA are $\mathbf{M} = \mathbf{I}_p$ if the variances are not too different and are expressed in the same measurement unit; otherwise, the metric $\mathbf{M} = \mathbf{D}_{1/s^2}$, the diagonal matrix of the variance inverses, is preferred. This latter metric is the most used (default option in many PCA programs) because as well as suppressing the measurement units, it gives the same importance in the computation of distances to each variable, whatever its variance. Using this metric is equivalent to standardizing the variables, setting them dimensionless and setting them all the same variance of 1. In the example, the variable standard deviations are very different (Figure 1.3), and then the variables will be standardized.

REMARK 1.1.– Every symmetric positive matrix $\mathbf{M}$ can be written as $\mathbf{M} = \mathbf{T}\mathbf{T}'$. We therefore have: $\boldsymbol{x}_i'\mathbf{M}\boldsymbol{x}_j = \boldsymbol{x}_i'\mathbf{T}'\mathbf{T}\boldsymbol{x}_j = (\mathbf{T}\boldsymbol{x}_i)'(\mathbf{T}\boldsymbol{x}_j)$. It is then possible to use $\mathbf{X}$ and the metric $\mathbf{M}$ rather than the identity metric $\mathbf{I}_p$ and the transformed data matrix $\mathbf{X}\mathbf{T}'$. PCA usually consists of standardizing the variables and using the identity metric $\mathbf{I}_p$, referred to as standardized PCA.

1.2.2.2. *Inertia*

Inertia is a fundamental notion of PCA. The total inertia of a data cloud is the weighted mean of square distances between points and the centroid. It represents the dispersion of the data cloud around the barycenter. Note that

$$I_g = \sum_{i=1}^{n} p_i(\boldsymbol{x}_i - \overline{\boldsymbol{x}})'\mathbf{M}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}) = \sum_{i=1}^{n} p_i||\boldsymbol{x}_i - \overline{\boldsymbol{x}}||^2.$$

It can be shown that the inertia around a particular point, defined by:

$$I_a = \sum_{i=1}^{n} p_i(\boldsymbol{x}_i - \mathbf{a})'\mathbf{M}(\boldsymbol{x}_i - \mathbf{a}),$$

may be written according to Huyghens formula:

$$I_a = I_g + (\overline{\boldsymbol{x}} - \mathbf{a})'\mathbf{M}(\overline{\boldsymbol{x}} - \mathbf{a}) = I_g + ||\overline{\boldsymbol{x}} - \mathbf{a}||^2.$$

It can be also shown that twice the total inertia is equal to the average of all of pairs of square distances between the n individuals. However, the most used equation is:

$$I = \mathrm{tr}(\mathbf{MS}) = \mathrm{tr}(\mathbf{SM}).$$

Then, if $\mathbf{M} = \mathbf{I}_p$, the inertia is equal to the sum of the p variances. In the case of the metric $\mathbf{M} = \mathbf{D}_{1/s^2}$, the inertia equals the trace of the correlation matrix i.e. p, the number of variables. The inertia then does not depend on the variables values but only on their number.

In the following chapter, this last case will be considered. For PCA with a general metric $\mathbf{M}$, see the books of Saporta [SAP 06] or Lebart *et al.* [LEB 06].

1.2.3. *Variables space*

Each variable is defined by n coordinates; it is then considered as a vector of a space of n dimensions referred to as variable space. To compute the 'distances' between variables, we use $\mathbf{D}_p$, the diagonal weight matrix, which has (in case of zero-mean variables) the following properties:

– the scalar product between two variables $\mathbf{x}^k$ and $\mathbf{x}^\ell$ is

$$(\mathbf{x}^k)'\mathbf{D}_p\mathbf{x}^\ell = \sum_{i=1}^{n} p_i x_i^k x_i^\ell,$$

which is the covariance $v_{k\ell}$;

– the square norm of a variable is then equal to its variance

$$||\mathbf{x}^j||^2_{\mathbf{D}_p} = s_j^2$$

and the standard deviation represents the variable 'length';

– by denoting the angle between two variables as $\theta_{k\ell}$, we have

$$\cos\theta_{k\ell} = \frac{<\mathbf{x}^k, \mathbf{x}^\ell>}{||\mathbf{x}^k||\,.\,||\mathbf{x}^\ell||} = \frac{v_{k\ell}}{s_k s_\ell} = r_{k\ell},$$

which is the linear correlation coefficient.

In the variables space, we are interested in angles rather than distances and the variables will be represented as vectors rather than points.

1.3. Principal component analysis

1.3.1. *The method*

Recall that the purpose of PCA is to find synthetic representations of large numerical datasets, in particular by using 2D plots. If the initial spaces of statistical units and variables representation have too many dimensions, it is impossible to visualize the data cloud. We therefore look for spaces with few dimensions best fitting the data cloud, that is, which save the best initial cloud configuration.

The method consists of projecting the data cloud in order to minimize the shrinkage of the distances which are inherent to the projection. This is equivalent to choosing the projection space F which maximizes the criterion:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} p_i p_j d^2(i,j).$$

The subspace we look for is such that the average of the square distances between projections is maximal (the projection reduces distances); in other words, the inertia of projections cloud has to be maximal. It is shown [SAP 06] that the search of the subspace F can be sequential: first we look for the 1D subspace with maximal inertia then we look for the 1D subspace orthogonal to this **[AQ: Changed 'the former' to 'this'; please confirm the meaning has not been changed]** with maximal inertia, and so on.

1.3.2. *Principal factors and principal components*

We begin by looking for a 1D subspace i.e. a straight line defined by a unit vector $\mathbf{u} = (u_1, \ldots, u_p)'$. As explained in the previous section, the vector has to be defined such that the points projected onto its direction have maximal inertia. The projection, or coordinate c_i, of an individual i onto Δ is defined by: $c_i = \sum_{j=1}^{p} \mathbf{x}_i^j u_j$ (Figure 1.1).

The list of the individuals coordinates c_i on Δ forms a new artificial variable $\mathbf{c} = (c_1, \ldots, c_n)' = \sum_{j=1}^{p} \mathbf{x}^j u_j = \mathbf{X}\mathbf{u}$; it is a linear combination of the original variables. The inertia (or variance) of points projected onto Δ is then:

$$\mathrm{Var}(\mathbf{c}) = \sum_{i}^{n} p_i c_i^2 = \mathbf{c}'\mathbf{D}_p\mathbf{c} = \mathbf{u}'\mathbf{X}'\mathbf{D}_p\mathbf{X}\mathbf{u} = \mathbf{u}'\mathbf{S}\mathbf{u}.$$

Recall that the usual case of standardized PCA is considered; the covariance matrix of standardized data then corresponds then to the correlation matrix $\mathbf{R}$. The criterion of maximizing the inertia of projected points onto Δ is then written as:

$$\max_{\mathbf{u}} \mathbf{u}'\mathbf{S}\mathbf{u} = \max_{\mathbf{u}} \mathbf{u}'\mathbf{R}\mathbf{u},$$

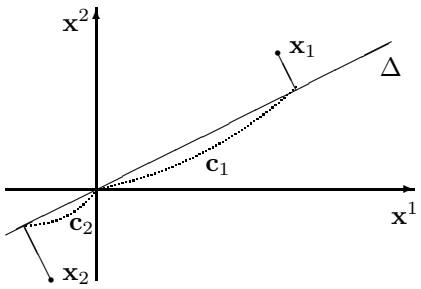

Figure 1.1. *Projection onto direction* Δ

under the constraint $\mathbf{u}'\mathbf{u} = 1$.

The solution of this quadratic maximization problem is $\mathbf{u}_1$, the eigenvector of $\mathbf{R}$ associated with the largest eigenvalue λ_1. We then search the vector $\mathbf{u}_2$ orthogonal to $\mathbf{u}_1$ such that the inertia of points projected onto this direction is maximal. Similarly, it is shown that $\mathbf{u}_2$ is the eigenvector of $\mathbf{R}$ associated with the second largest eigenvalue λ_2. More generally, the subspace of q dimensions which we are looking for is spanned by the first q eigenvectors of the matrix $\mathbf{R}$ associated with the largest eigenvalues.

Vectors $\mathbf{u}_j$ are called *principal factors*. They contain the coefficients to be applied to the original variables in the linear combination $\mathbf{c} = \mathbf{X}\mathbf{u}$.

Principal components are artificial variables defined by principal factors: $\mathbf{c}^j = \mathbf{X}\mathbf{u}_j$; they contain the coordinates of the orthogonal projections of individuals onto the axes defined by the $\mathbf{u}_j$.

In practice, PCA will consist of diagonalizing the $\mathbf{R}$ matrix to obtain the $\mathbf{u}_j$ and computing the principal components $\mathbf{c}^j = \mathbf{X}\mathbf{u}_j$.

See Tables 1.5 and 1.6 for the example results.

Number	Eigenvalues	%	% cumulated
1	3.8168	47.71	47.71
2	2.0681	25.85	73.56
3	0.9728	12.16	85.72
4	0.7962	9.95	95.67
5	0.1792	2.24	97.91
6	0.0924	1.16	99.07
7	0.0741	0.93	100.00
8	0.0004	0.00	100.00

Table 1.5. *Eigenvalues* λ_1, λ_2, *etc.*

	Eigenvectors				
Variables	1	2	3	4	5
CA	–0.28	–0.54	–0.17	–0.20	0.01
MG	–0.47	–0.18	–0.04	–0.17	–0.46
NA	–0.44	0.29	–0.03	0.17	0.62
K	–0.43	0.32	0.01	–0.12	–0.45
SUL	–0.23	–0.60	–0.03	–0.03	0.33
NO3	0.12	0.06	–0.97	0.15	–0.06
HCO3	–0.40	0.35	–0.13	–0.35	0.24
CL	–0.32	–0.07	0.07	0.86	–0.15

Table 1.6. *Eigenvectors*

1.3.3. *Principal factors and principal components properties*

1.3.3.1. *Principal component variance*

The variance of a principal component is equal to the eigenvalue λ: $\mathrm{Var}(\mathbf{c}^j) = \lambda_j$. The variance of $\mathbf{u}$ is defined by $\mathbf{Su} = \mathbf{Ru}$, $\mathbf{u}'\mathbf{u} = 1$ and:

$$\mathrm{Var}(\mathbf{c}) = \mathbf{c}'\mathbf{D}_p\mathbf{c} = \mathbf{u}'\mathbf{X}'\mathbf{D}_p\mathbf{X}\mathbf{u} = \mathbf{u}'\mathbf{S}\mathbf{u} = \mathbf{u}'\mathbf{R}\mathbf{u} = \mathbf{u}'(\lambda\mathbf{u}) = \lambda\mathbf{u}'\mathbf{u} = \lambda.$$

The principal components are therefore linear combinations of original variables with maximal variance.

1.3.3.2. *A maximal association property*

The variable $\mathbf{c}^1$ has the greatest link to $\mathbf{x}^j$ in the sense of the square correlations sum: $\sum_{j=1}^{p} r^2(\mathbf{c}, \mathbf{x}^j)$ is maximal. It is shown [SAP 06] that:

$$\sum_{j=1}^{p} r^2(\mathbf{c}, \mathbf{x}^j) = \frac{\mathbf{c}'\mathbf{D}_p\mathbf{Z}\mathbf{Z}'\mathbf{D}_p\mathbf{c}}{\mathbf{c}'\mathbf{D}_p\mathbf{c}}$$

where $\mathbf{Z}$ is the standardized data table. The maximum of this ratio is reached when $\mathbf{c}$ is the eigenvector of $\mathbf{Z}\mathbf{Z}'\mathbf{D}_p$ associated with its largest eigenvalue: $\mathbf{Z}\mathbf{Z}'\mathbf{D}_p\mathbf{c} = \lambda\mathbf{c}$.

The principal component $\mathbf{c}$ is then a linear combination of the columns of $\mathbf{Z}$: $\mathbf{c} = \mathbf{Z}\mathbf{u}$ and then $\mathbf{Z}\mathbf{Z}'\mathbf{D}_p\mathbf{c} = \lambda\mathbf{c}$ becomes $\mathbf{Z}\mathbf{Z}'\mathbf{D}_p\mathbf{Z}\mathbf{u} = \lambda\mathbf{Z}\mathbf{u}$. Since we have $\mathbf{Z}'\mathbf{D}_p\mathbf{Z} = \mathbf{R}$ and $\mathbf{Z}\mathbf{R}\mathbf{u} = \lambda\mathbf{Z}\mathbf{u}$ and, if the rank of $\mathbf{Z}$ is p, we obtain $\mathbf{R}\mathbf{u} = \lambda\mathbf{u}$.

1.3.3.3. *Reconstitution formula*

Post-multiplying both members of $\mathbf{X}\mathbf{u}_j = \mathbf{c}^j$ by $\mathbf{u}_j'$ and summing over j, we have

$$\mathbf{X}\sum_{j=1}^{p}\mathbf{u}_j\mathbf{u}_j' = \sum_{j=1}^{p}\mathbf{c}^j\mathbf{u}_j'.$$

It can easily be shown that $\sum_{j=1}^{p}\mathbf{u}_j\mathbf{u}_j' = \mathbf{I}_p$ since the $\mathbf{u}_j$ are orthonormal. We then find $\mathbf{X} = \sum_{j=1}^{p}\mathbf{c}^j\mathbf{u}_j'$. The centered data table may be reconstituted using factors

and principal components. If we only use the first q terms corresponding to the first q largest eigenvalues, we have the best approximation of $\mathbf{X}$ by a matrix of rank q in the least-squares sense (Eckart–Young theorem).

To summarize, it can be said that PCA consists of transforming original correlated variables $\mathbf{x}^j$ into new variables, the principal components $\mathbf{c}^j$, which are uncorrelated linear combinations of the $\mathbf{x}^j$ with maximal variance and with the greatest link to the $\mathbf{x}^j$. PCA is therefore a linear factorial method.

Non-linear extensions of PCA exist: we look for variable transformations, for example, by splines [DEL 88] available in some software (prinqual procedure in SAS). **[AQ: Please define SAS]**

1.4. Interpretation of PCA results

PCA provides graphical representations allowing the visualization of relations between variables and the eventual existence of groups of individuals and groups of variables. PCA results are 2D figures and tables. Their interpretation is the most delicate phase of the analysis and has to be carried out according to a precise scheme to be explained later.

Before beginning the interpretation itself, it is useful to start with a brief preliminary reading of the results in order to roughly verify the dataset contents. It is possible that by examining the first principal plane, we observe some individuals completely outside the rest of the population. This implies either (1) the presence of erroneous data such as typing errors or measurement error which have to be corrected, or (2) individuals totally different from others which must be removed from the analysis to better observe the remaining individuals (they can be reintroduced afterwards as supplementary elements).

After this preliminary study, PCA results can then be examined more closely; we begin with the interpretation phase which consists of several stages.

REMARK 1.2.– Although simultaneous representations of individuals and variables called 'biplot' [GOW 96] exist, we recommend representing the set separately in order to avoid confusion.

1.4.1. *Quality of representations onto principal planes*

PCA allows us to obtain graphical representation of individuals in a space of fewer dimensions than p, but this representation is only a deformed vision of the reality. One of the most crucial points in interpreting the results of PCA consists of appreciating

this deformation (or, in other words, the loss of information due to the dimension reduction) and in determining the number of axes to retain.

The criterion usually employed to measure PCA quality is the percentage of total inertia explained. It is defined:

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{\lambda_1 + \lambda_2 + \ldots + \lambda_k + \ldots + \lambda_p} = \frac{\lambda_1 + \lambda_2 + \ldots + \lambda_k}{I_g}.$$

This is a global measure which has to be completed with other considerations. First, the number of variables must be taken into account: a 10% inertia does not have the same interest for a 20-variable table or for a 100-variable table.

Second, at the individual level, it is necessary to look at the reliability of the representation of each individual, independently of the global inertia percentage. **[AQ: Have reworded; please confirm correct]** It is possible to have a first principal plane with a large total inertia and to find that two individuals, far from each other in the full space, have very close projections (Figure 1.2).

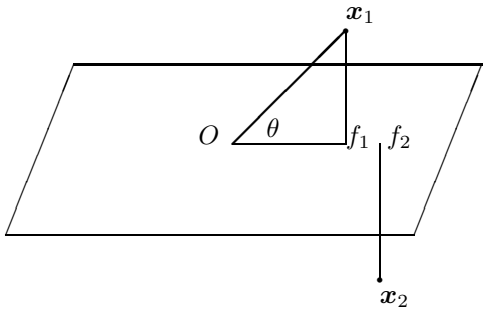

Figure 1.2. *Close projections of distant points*

The most widely used measure of an individual representation quality is the cosine of the angle between the principal plane and the vector $\boldsymbol{x}_i$. If this cosine is large, $\boldsymbol{x}_i$ is close to the plane and we can then examine the position of its projection onto the plane with respect to other points; if the cosine is small, we will be wary of any conclusion.

1.4.2. *Axis selection*

Axis selection is an essential point of PCA but has no rigorous solution. There are theoretical criteria based on statistical tests or eigenvalue confidence intervals but the latter are useful only for non-standardized PCA and the p-dimensional Gaussian case. In the most frequent practical case of correlation matrices, only empirical criteria are applicable. The best known is the Kaiser rule: for standardized data, principal

components corresponding to eigenvalues larger than 1 are retained; this means only components which 'bring' more than original variables are of interest.

It is also usual to employ a scree test, which consists of detecting the existence of a significant decay on the eigenvalues diagram. This is not always easy in practice, however.

In the example we use the Kaiser rule combined with the eigenvalues diagram (see Table 1.7). A break is detected after the second eigenvalue and we retain two axes corresponding to 73.56% explained inertia. The third axis is easily interpreted, but as it is identified with the variable NO3 (correlation –0.96) and is not correlated with other variables, it is not of great interest.

Number	Eigenvalues	%	% cumulated	
1	3.8168	47.71	47.71	**********
2	2.0681	25.85	73.56	**********
3	0.9728	12.16	85.72	**********
4	0.7962	9.95	95.67	**********
5	0.1792	2.24	97.91	****
6	0.0924	1.16	99.07	**
7	0.0741	0.93	100.00	**
8	0.0004	0.00	100.00	*

Table 1.7. *Eigenvalues scree plot*

1.4.3. *Internal interpretation*

PCA results are obtained from variables and individuals called active elements, in contrast to supplementary elements which do not participate directly in the analysis. Active variables and individuals are used to compute principal axes; supplementary variables and individuals are then projected onto these axes.

Active variables (numerical) are those with interesting intercorrelations: they are the main variables of the study. Supplementary variables provide useful information for characterizing the individuals but are not directly used in the analysis. We are only interested by the correlations of supplementary variables and active variables via principal components, and not by the correlations between supplementary variables. Internal interpretation consists of analyzing the results using active variables and individuals. Supplementary elements study is carried out in the external interpretation phase.

1.4.3.1. *Variables*

PCA yields principal components, which are new artificial variables defined by linear combinations of original variables. We must be able to interpret these principal components (according to original variables). This is done simply through the computations of linear correlation coefficients $r(\mathbf{c}, \mathbf{x}^j)$ between principal components and original variables. The largest coefficients (in absolute value) close to 1 are

those of interest (see Table 1.8). In standard PCA, we use standardized data and the computation of $r(\mathbf{c}, \mathbf{x}^j)$ is particularly simple. It may be shown that $r(\mathbf{c}, \mathbf{x}^j) = \sqrt{\lambda}\mathbf{u}_j$.

	Coordinates				
Variables	1	2	3	4	5
CA	–0.55	–0.78	–0.17	–0.18	0.01
MG	–0.91	–0.25	–0.04	–0.15	–0.20
NA	–0.86	0.41	–0.03	0.15	0.26
K	–0.84	0.46	0.01	–0.11	–0.19
SUL	–0.45	–0.87	–0.03	–0.03	0.14
NO3	0.23	0.09	0.96	0.13	–0.03
HCO3	–0.78	0.50	–0.13	–0.31	0.10
CL	–0.62	–0.10	0.07	0.77	–0.06

Table 1.8. *Variable-factor correlations or variables coordinates*

Usually the correlation between the variables for a couple of principal components is synthesized on a graph called the 'correlation display' on which each variable $\mathbf{x}^j$ is positioned by abscissa $r(\mathbf{c}^1, \mathbf{x}^j)$ and ordinate $r(\mathbf{c}^2, \mathbf{x}^j)$. Analyzing the correlation display allows detection of possible groups of similar variables or opposite groups of variables having different behavior, giving a sense of principal axes.

In the example (Figure 1.3), axis 1 is negatively correlated to all the variables (except NO3 which is not significant). Observations with the largest negative coordinate on the horizontal axis correspond to water with the most important mineral concentrations. Along the vertical axis, waters with high calcium and sulfate concentration are in opposition with waters with high potassium and carbonate concentration.

REMARK 1.3.– 'Size effect' When all original variables are positively correlated with each other, the first principal component defines a 'size effect'. It is known that a symmetric matrix with all terms positive has a first eigenvector with all its component having the same sign, which can be chosen to be positive. Then the first principal component is positively correlated to all original variables and individuals are ranked along axis 1 according to the increase of all variables (on average). The second principal component distinguishes individuals with similar 'size' and is referred to as the 'shape factor'.

1.4.3.2. *Observations*

Interpreting observations consists of examining their coordinates and especially their resulting graphical representations referred to as principal planes (Figure 1.4). The aim is to see how the observations are scattered, which observations are similar and which observations differ from the others. In case of non-anonymous observations, they can then help to interpret principal axes ; for example, we will look for opposite individuals along an axis.

Conversely, using results of variables analysis allows observations interpretation. When, for example, the first component is highly correlated with an original variable,

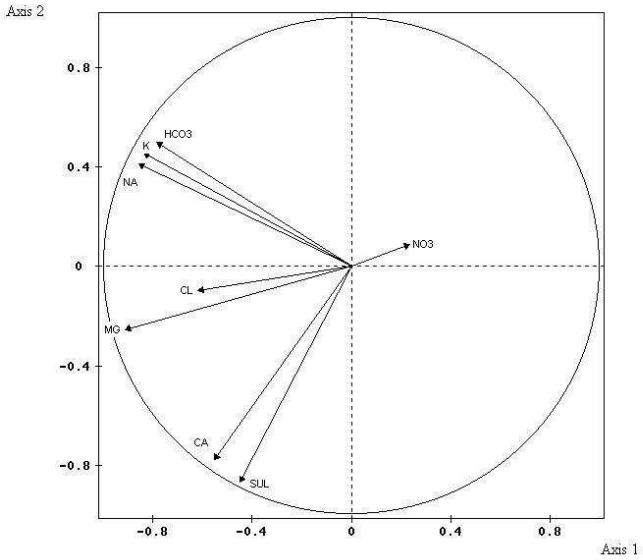

Figure 1.3. *Variable representation onto plane 1, 2*

it means that individuals with large positive coordinates along axis 1 are characterized by this variable of value much larger than average (the origin of the axes represents the centroid of data cloud).

In observation study, it is also very useful to look at the individual contributions of each axis for help in interpreting axes. **[AQ: Have reworded; please confirm correct]** Contribution is defined by $\frac{p_i(c_i^k)^2}{\lambda_k}$, where c_i^k represents the value for individual i of the kth component $\mathbf{c}^k$ and $\lambda_k = \sum_{i=1}^{n} p_i(c_i^k)^2$.

The important contributions are those that exceed observation weight. However, it is necessary to be careful when an individual has an excessive contribution which can produce instability. Removing it can highly modify the analysis results. The analysis should be made without this point, which can be added as a supplementary element. Observations such as ARVIE and PAROT (Figure 1.4) are examples of such points.

It should be noted that, for equal weight, contributions do not provide more information than coordinates. Table 1.9 lists coordinates, contributions and square cosines of angles with principal axes which allow the evaluation of the quality of the representation (see section 1.4.1).

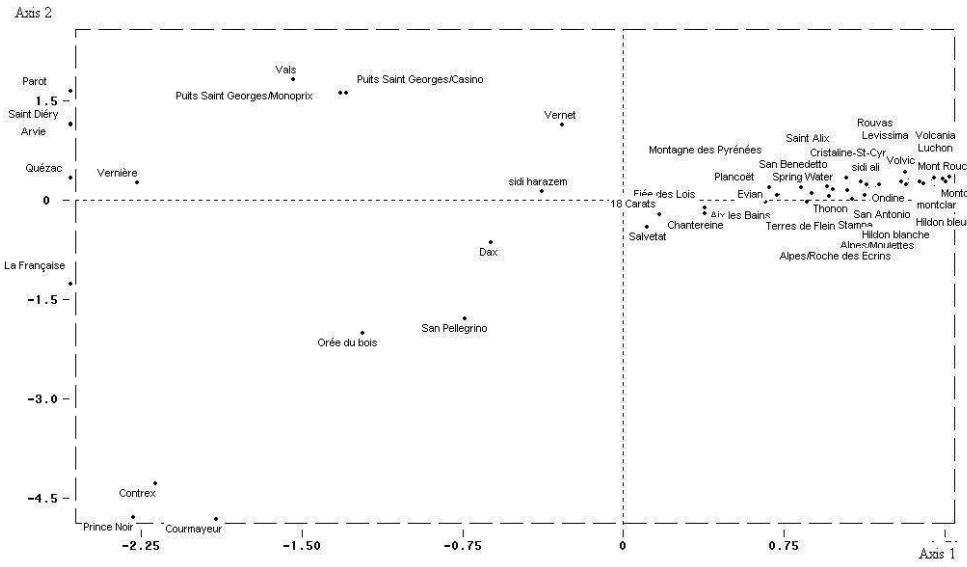

Figure 1.4. *Observations representation on plane 1, 2*

1.4.4. *External interpretation: supplementary variables and individuals*

Recall that supplementary elements do not participate in the computations of principal axes, but are very useful afterwards in consolidating and enriching the interpretation.

The case of numerical supplementary variables has to be distinguished from the categorical variables. The former are positioned on the correlation display after having simply computed the correlation coefficient between each supplementary variable and the principal components. The interpretation is made in the same way as for active variables through the detection of significant correlations.

For supplementary categorical variables, we generally represent each category by its barycenter in the observation space. Some software (especially SPAD, for example) provide helps with interpretation by giving test values which measure the distance of the point representing a category of the origin.

More precisely, the test value measures this distance in number of standard deviations of a normal distribution. They allow an extremal position of a subgroup of observations to be displayed. A category will be considered as significant for an axis if its associated test value is larger in absolute value than 2 (with a 5% risk).

Individuals			Coordinates					Contributions					Square cosines				
Identifier	P. Rel.	DISTO	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Evian	1.75	0.71	0.72	0.06	0.06	−0.21	−0.18	0.2	0.0	0.0	0.1	0.3	0.73	0.01	0.01	0.06	0.04
Montagne des Pyrénées	1.75	1.08	0.95	0.19	0.15	0.33	0.01	0.4	0.0	0.0	0.2	0.0	0.84	0.04	0.02	0.10	0.00
Cristaline–St–Cyr	1.75	1.38	0.98	0.16	0.54	0.00	0.07	0.4	0.0	0.5	0.0	0.0	0.69	0.02	0.21	0.00	0.00
Fiée des Lois	1.75	0.80	0.38	−0.11	0.60	−0.21	−0.22	0.1	0.0	0.7	0.1	0.5	0.18	0.02	0.45	0.05	0.06
Volcania	1.75	2.81	1.45	0.33	0.71	0.15	0.07	1.0	0.1	0.9	0.1	0.0	0.75	0.04	0.18	0.01	0.00
Saint Diéry	1.75	16.07	−3.54	1.48	0.13	0.56	−0.95	5.8	1.9	0.0	0.7	8.9	0.78	0.14	0.00	0.02	0.06
Luchon	1.75	2.36	1.40	0.25	0.52	0.12	0.11	0.9	0.1	0.5	0.0	0.1	0.83	0.03	0.12	0.01	0.00
Volvic	1.75	2.12	1.32	0.41	−0.12	0.24	−0.11	0.8	0.1	0.0	0.1	0.1	0.82	0.08	0.01	0.03	0.01
Alpes/Moulettes	1.75	1.31	1.07	0.01	0.40	−0.06	0.04	0.5	0.0	0.3	0.0	0.0	0.87	0.00	0.12	0.00	0.00
Orée du bois	1.75	6.37	−1.22	−2.02	0.16	−0.49	−0.37	0.7	3.5	0.0	0.5	1.4	0.23	0.64	0.00	0.04	0.02
Arvie	1.75	52.52	−6.51	2.67	0.10	0.35	−1.25	19.5	6.1	0.0	0.3	15.4	0.81	0.14	0.00	0.00	0.03
Alpes/Roche des Ecrins	1.75	1.31	1.07	0.01	0.40	−0.06	0.04	0.5	0.0	0.3	0.0	0.0	0.87	0.00	0.12	0.00	0.00
Ondine	1.75	1.93	1.14	0.22	0.74	−0.03	0.06	0.6	0.0	1.0	0.0	0.0	0.67	0.03	0.28	0.00	0.00
Thonon	1.75	2.35	0.96	0.05	−1.17	0.01	−0.10	0.4	0.0	2.5	0.0	0.1	0.39	0.00	0.58	0.00	0.00
Aix les Bains	1.75	0.99	0.66	−0.03	0.59	−0.30	−0.12	0.2	0.0	0.6	0.2	0.1	0.44	0.00	0.35	0.09	0.02
Contrex	1.75	25.50	−2.18	−4.29	−0.57	−1.40	0.07	2.2	15.6	0.6	4.3	0.0	0.19	0.72	0.01	0.08	0.00
La Bondoire Saint Hippol	1.75	6.57	1.33	0.26	−2.13	0.41	0.00	0.8	0.1	8.2	0.4	0.0	0.27	0.01	0.69	0.03	0.00
Dax	1.75	1.78	−0.62	−0.64	0.61	0.61	−0.07	0.2	0.3	0.7	0.8	0.0	0.22	0.23	0.21	0.21	0.00
Quézac	1.75	15.10	−3.37	0.36	−0.18	−1.56	−0.78	5.2	0.1	0.1	5.4	6.0	0.75	0.01	0.00	0.16	0.04
Salvetat	1.75	3.00	0.11	−0.41	0.14	−0.77	0.25	0.0	0.1	0.0	1.3	0.6	0.00	0.06	0.01	0.20	0.02
Stamna	1.75	1.66	1.04	0.15	0.73	0.06	0.04	0.5	0.0	1.0	0.0	0.0	0.66	0.01	0.32	0.00	0.00
Iolh	1.75	1.00	0.73	0.09	−0.24	−0.05	−0.35	0.2	0.0	0.1	0.0	1.2	0.53	0.01	0.06	0.00	0.12
Avra	1.75	24.03	1.45	0.22	−4.63	0.58	−0.22	1.0	0.0	38.7	0.7	0.5	0.09	0.00	0.89	0.01	0.00
Rouvas	1.75	1.63	1.20	0.23	0.32	0.14	−0.04	0.7	0.0	0.2	0.0	0.0	0.88	0.03	0.06	0.01	0.00
Alisea	1.75	2.43	1.44	0.30	0.45	0.16	0.06	0.9	0.1	0.4	0.1	0.0	0.85	0.04	0.08	0.01	0.00
San Benedetto	1.75	1.13	0.83	0.18	−0.29	−0.09	−0.30	0.3	0.0	0.2	0.0	0.9	0.61	0.03	0.08	0.01	0.08
San Pellegrino	1.75	4.15	−0.74	−1.79	0.33	−0.21	−0.15	0.3	2.7	0.2	0.1	0.2	0.13	0.77	0.03	0.01	0.01
Levissima	1.75	2.40	1.38	0.27	0.58	0.11	0.07	0.9	0.1	0.6	0.0	0.0	0.80	0.03	0.14	0.01	0.00
Vera	1.75	1.42	1.15	0.18	0.21	0.03	−0.07	0.6	0.0	0.1	0.0	0.0	0.93	0.02	0.03	0.00	0.00
San Antonio	1.75	1.80	1.30	0.27	0.13	0.10	0.02	0.8	0.1	0.0	0.0	0.0	0.94	0.04	0.01	0.01	0.00
La Française	1.75	68.27	−5.64	−2.83	0.45	5.28	0.55	14.6	6.8	0.4	61.3	3.0	0.47	0.12	0.00	0.41	0.00
Saint Benoit	1.75	1.93	1.14	0.22	0.74	−0.03	0.06	0.6	0.0	1.0	0.0	0.0	0.67	0.03	0.28	0.00	0.00
Plancoët	1.75	1.10	0.68	0.19	0.73	0.11	−0.12	0.2	0.0	1.0	0.0	0.2	0.42	0.03	0.49	0.01	0.01
Saint Alix	1.75	1.88	1.04	0.33	0.74	0.28	−0.02	0.5	0.1	1.0	0.2	0.0	0.58	0.06	0.29	0.04	0.00
Puits Saint Georges/Casi	1.75	6.28	−1.29	1.62	−0.79	−0.20	1.03	0.8	2.2	1.1	0.1	10.3	0.27	0.42	0.10	0.01	0.17
St–Georges/Corse	1.75	2.70	1.33	0.32	0.84	0.28	0.08	0.8	0.1	1.3	0.2	0.1	0.66	0.04	0.26	0.03	0.00
Hildon bleue	1.75	13.11	1.51	0.27	−3.22	0.54	−0.08	1.0	0.1	18.8	0.6	0.1	0.17	0.01	0.79	0.02	0.00
Hildon blanche	1.75	1.52	1.13	0.07	−0.15	0.07	0.12	0.6	0.0	0.0	0.0	0.1	0.84	0.00	0.02	0.00	0.01
Mont Roucous	1.75	2.84	1.53	0.35	0.50	0.23	0.08	1.1	0.1	0.5	0.1	0.1	0.82	0.04	0.09	0.02	0.00
Ogeu	1.75	1.09	0.97	0.19	0.15	0.29	0.01	0.4	0.0	0.0	0.2	0.0	0.87	0.03	0.02	0.08	0.00
Highland spring	1.75	1.79	1.17	0.21	0.61	0.04	0.02	0.6	0.0	0.7	0.0	0.0	0.77	0.02	0.21	0.00	0.00
Parot	1.75	63.44	−6.61	3.99	−0.40	−1.58	1.09	20.1	13.5	0.3	5.5	11.6	0.69	0.25	0.00	0.04	0.02
Vernière	1.75	7.65	−2.27	0.26	0.19	−1.27	−0.88	2.4	0.1	0.1	3.6	7.6	0.67	0.01	0.00	0.21	0.10
Terres de Flein	1.75	1.33	0.86	−0.03	0.45	−0.14	0.16	0.3	0.0	0.4	0.0	0.2	0.55	0.00	0.16	0.02	0.02
Courmayeur	1.75	29.42	−1.90	−4.83	−0.46	−1.30	0.46	1.7	19.8	0.4	3.7	2.1	0.12	0.79	0.01	0.06	0.01
Pyrénées	1.75	1.06	0.98	0.19	0.14	0.23	0.00	0.4	0.0	0.0	0.1	0.0	0.90	0.03	0.02	0.05	0.00
Puits Saint Georges/Mono	1.75	6.37	−1.32	1.62	−0.80	−0.20	1.02	0.8	2.2	1.1	0.1	10.2	0.27	0.41	0.10	0.01	0.16
Prince Noir	1.75	30.69	−2.29	−4.80	−0.23	−1.46	0.33	2.4	19.6	0.1	4.7	1.1	0.17	0.75	0.00	0.07	0.00
Montcalm	1.75	2.94	1.49	0.31	0.71	0.17	0.09	1.0	0.1	0.9	0.1	0.1	0.76	0.03	0.17	0.01	0.00
Chantereine	1.75	0.87	0.38	−0.20	0.54	−0.42	−0.15	0.1	0.0	0.5	0.4	0.2	0.17	0.05	0.33	0.20	0.02
18 Carats	1.75	0.51	0.17	−0.22	0.48	−0.23	−0.25	0.0	0.0	0.4	0.1	0.6	0.06	0.09	0.45	0.10	0.13
Spring Water	1.75	6.53	0.88	0.10	−2.37	0.23	−0.24	0.4	0.0	10.1	0.1	0.6	0.12	0.00	0.86	0.01	0.01
Vals	1.75	7.28	−1.54	1.82	0.24	−0.43	1.15	1.1	2.8	0.1	0.4	12.9	0.33	0.46	0.01	0.03	0.18
Vernand	1.75	1.87	−0.29	1.13	0.44	−0.33	0.18	0.0	1.1	0.3	0.2	0.3	0.04	0.68	0.10	0.06	0.02
sidi harazem	1.75	1.91	−0.38	0.13	0.12	1.10	−0.44	0.1	0.0	0.0	2.7	1.9	0.08	0.01	0.01	0.64	0.10
sidi ali	1.75	1.98	1.11	0.27	0.78	0.12	0.06	0.6	0.1	1.1	0.0	0.0	0.62	0.04	0.30	0.01	0.00
montclar	1.75	1.93	1.32	0.23	0.31	0.08	0.09	0.8	0.0	0.2	0.0	0.1	0.91	0.03	0.05	0.00	0.00

Table 1.9. *Coordinates, contributions and square cosine of individuals*

In the example, the barycenter of sparkling waters (and consequently, that of still waters) is more than 3 standard deviations from the origin $(-3, 5)$. Sparkling waters are significantly very far from the origin.

It is easy to plot supplementary individuals onto the principal axes. Since we have the formulae allowing principal components computations, we simply have to compute linear combinations of these supplementary points characteristics. **[AQ: Please indicate where Table 1.10 should be cross-referenced]**

Categories			Test values					Coordinates				
Label	EFF.	P.ABS	1	2	3	4	5	1	2	3	4	5
1. Country												
France	40	40.00	–1.9	0.7	2.1	–0.5	0.7	–0.33	0.09	0.18	–0.04	0.03
Britain	4	4.00	1.2	0.2	–2.7	0.5	–0.2	1.17	0.16	–1.28	0.22	–0.05
Greece	2	2.00	0.8	0.2	–3.5	0.4	–1.0	1.09	0.15	–2.44	0.26	–0.29
Greece-Crete	2	2.00	0.8	0.2	0.8	0.2	0.0	1.12	0.19	0.52	0.10	0.00
Italy	7	7.00	0.7	–1.5	0.4	–0.5	0.1	0.49	–0.77	0.13	–0.17	0.01
Morocco	2	2.00	0.3	0.2	0.6	1.0	–0.6	0.36	0.20	0.45	0.61	–0.19
2 . Type												
Mineral	38	38.00	–2.5	–0.5	–1.2	–0.7	0.4	–0.46	–0.07	–0.11	–0.06	0.01
Spring	19	19.00	2.5	0.5	1.2	0.7	–0.4	0.92	0.13	0.22	0.11	–0.03
3 . PG												
Sparkling	16	16.00	–3.5	2.7	–0.5	–1.8	0.3	–1.44	0.82	–0.11	–0.34	0.03
Still	41	41.00	3.5	–2.7	0.5	1.8	–0.3	0.56	–0.32	0.04	0.13	–0.01

Table 1.10. *Coordinates and test values of the categories* ***[AQ: What do EFF. and P.ABS. represent?]***

1.5. Application to statistical process control

1.5.1. *Introduction*

Online statistical process control is essentially based on control charts for measurements, drawing the evolution of a product or process characteristics. A control chart is a tool which allows a shift of a location (mean) or a dispersion (standard deviation, range) parameter regarding fixed standard or nominal values to be detected through successive small samples $(\boldsymbol{x}_i, i = 1, 2, \ldots, n)$.

Several types of control charts exist [MON 85, NIA 94], all based on the assumption that the distribution of $\boldsymbol{x}_i$ is $\mathcal{N}(\mu_0, \sigma_0)$. Standard or nominal values μ_0 and σ_0 are assumed known or fixed. If this is not the case, they are replaced by unbiased estimations.

Here, we are only interested in classical Shewhart control charts for the detection of process mean shifts. In Shewhart control charts, at each instant i we use only $\overline{x}_i$, the mean value of observations available, which is compared to lower (LCL) and upper (UCL) control limits:

$$LCL = \mu_0 - 3\sigma_0/\sqrt{n} \quad \text{and} \quad UCL = \mu_0 + 3\sigma_0/\sqrt{n}.$$

This control chart can be seen as a graphical representation of a succession of statistical tests $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ for a set of samples; the standard deviation σ_0 is assumed known. The critical region corresponds to the control chart from the control area. This equivalence to hypothesis tests will facilitate extension to several variables.

In most of the cases there are not one but several characteristics to simultaneously control. The usual practice consists of using as many charts as characteristics. This

method has the major drawback that it does not take into account the correlations between variables representing these p characteristics. That then leads to undesired situations of false alarms (Figure 1.5). The univariate charts may signal an out-of-control situation while the multivariate process is under control (region B and C) or, more severe, a non-detection of a multivariate process shift (region A) may occur. **[AQ: Have reworded; please confirm correct]**

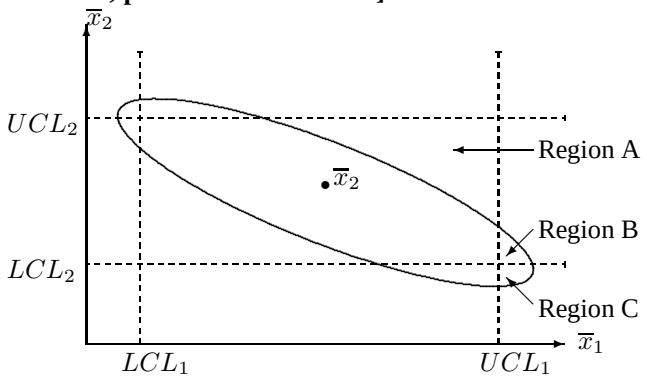

Figure 1.5. *Multivariate control chart*

A global approach through multivariate charts is therefore the only adequate approach (see [NIA 02]). Principal component analysis, which provides artificial but uncorrelated variables, is, in some sense, a first solution to the problem of correlated characteristics. We will see later that once a shift has been detected, adequate univariate charts may help to determine which variables are responsible of this shift, or assignable causes.

1.5.2. *Control charts and PCA*

1.5.2.1. *PCA and outliers*

Multivariate control charts are based on a transformation of a $\mathbb{R}^p$ vector in a scalar through a quadratic function. They can be seen as methods for multidimensional outlier detection. These methods consist of finding a sub-order in $\mathbb{R}^p$ generally based on a multivariate distance measure (a detailed study can be found in [BAR 84]). This measure is then used in a statistical test, to decide if an observation is an outlier when the standardized statistic has an abnormally large or small value, under a model hypothesis (in quality control, normality assumption is often made).

In the multidimensional case, an outlier may be the result of an extreme value for only one of the p principal components or the result of small systematic errors in several directions. This latter type of outlier corresponds to the problem of orientation (correlation) and not of location (mean) or dispersion (variance). Using principal

component analysis, not as a dimension reduction method but rather as an outlier detection method, facilitates the search for extreme directions.

To best summarize the data structure, not only the first components should be retained but also the last components considered as the residuals of the analysis. Jolliffe [JOL 86] has shown that the first components allow the detection of outliers which inflate the variances and covariances. These outliers are also extreme on original variables, so they can be directly detected. The first components do not yield supplementary information.

On the other hand, outliers not visible on original variables (those that perturb the correlation between variables) will be detected on the last principal components. Several methods of outliers detection based on PCA have been proposed by many authors, specially Hawkins [HAW 74], Gnanadesikan [GNA 77], Jolliffe [JOL 86] and Barnett and Lewis [BAR 84].

Proposed techniques consist of applying formal statistical tests to principal components individually or conjointly. These tests are based on residual statistics computed using the last q principal components. The most widely used residual statistics are:

$$R_{1i}^2 = \sum_{k=p-q+1}^{p} (c_i^k)^2 = \sum_{k=p-q+1}^{p} \lambda_k (y_i^k)^2,$$

where $\mathbf{y}_k = \frac{\mathbf{c}^k}{\sqrt{\lambda_k}}$, R_{1i}^2 is a weighted sum of the standardized principal components which give more importance to principal components with large variances and

$$R_{2i}^2 = \sum_{k=p-q+1}^{p} (c_i^k)^2/\lambda_k = \sum_{k=p-q+1}^{p} (y_i^k)^2.$$

The distributions of these statistics are easily obtained. If the observations are normally distributed with known mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$, $\mathbf{y}_k$ have an exact Gaussian distribution $\mathcal{N}(0,1)$. If there are no outliers in the data, the residual statistics R_{1i}^2 and R_{2i}^2 have χ_q^2 distribution. When $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown, it is also possible (using their estimations) to obtain approximate distributions of these residual statistics and then to perform statistical tests.

1.5.2.2. *Control charts associated with principal components*

In quality control, PCA is used as a method for detecting shifts considered as outliers. The last principal components may be as interesting as the first components, since the type or the direction of the shifts are *a priori* unknown.

Recall that principal components are defined as linear combinations of original variables, which best summarize the data structure. They take into account the

correlations between variables and then, even taken individually, they help to detect shifts (unlike original variables).

Note that principal components charts should not be used instead of but in conjunction with multivariate charts. The problem of false alarms and non-detection of out-of-control (noted on Figure 1.5) is attenuated but not completely suppressed for uncorrelated principal components. The 3-sigma control limits for standardized principal components are then $\pm 3/\sqrt{n}$. The presence of outliers can be tested with a control chart on the R_{2i}^2, whose upper control limit corresponds to the fractile $1-\alpha$ of a χ_q^2.

For residual statistics, the presence of outliers can be tested with a control chart defined by:

$$UCL = \chi^2_{q,1-\alpha}, \qquad LCL = 0 \quad \text{and} \quad Stat = R_{2i}^2.$$

EXAMPLE.– We have simulated 30 samples of 5 observations from a multinormal distribution $\mathcal{N}_3(0, R)$; the three variables are assumed to have zero mean and variances equal to 0.5, 0.3 and 0.1, respectively. The correlation matrix is:

$$R = \begin{pmatrix} 1 & & \\ 0.9 & 1 & \\ 0.1 & 0.3 & 1 \end{pmatrix}.$$

We have then simulated a mean shift for the last five samples which consists of increasing the first variable mean and diminishing the second variable mean by half of their standard deviation. This situation is detected by the adequate multidimensional control chart [NIA 94] as well as the last principal component control chart. In Figure 1.6, note that the last five control points are clearly detected while the phenomenon is not visible on the first two principal components.

When the number of characteristics is small, it is possible to find a simple interpretation for the principal components based on a small number of variables among the p original variables. Control charts based on principal components not only allow the detection of shifts but also help with the detection of assignable causes.

On another hand, if the number of variables is very large, the proposed methods require many control charts for the first and the last components, which may be unpractical. **[AQ: Changed 'annoying' to 'unpractical'; please confirm ok]** We may only use the first q components, as in the dimension reduction approach of PCA, but then it is necessary (1) to test the quality of the representation of the p original variables by the q components and (2) to use methods based on residuals for outliers or shifts detection.

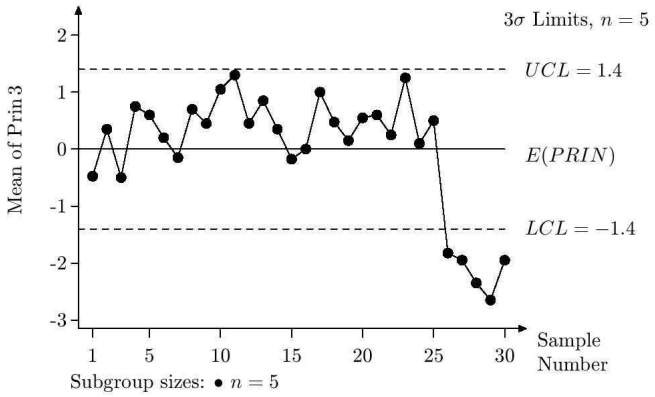

Figure 1.6. *Control chart for the 3rd principal component (Prin 3)*

Furthermore, even if we find q principal components summarizing at best the information present in the p original variables, these q components depend on a large number of variables or on all original variables. To simplify the principal components interpretation, other methods of *projection pursuit* have been proposed [NIA 94]. The work done by Caussinus *et al.* [CAU 03] (see Chapter 3) is useful in improving these methods.

1.6. Conclusion

PCA is a very efficient method for representing correlated data. It is widely used in market study, opinion surveys and in the industrial sector more and more.

We have presented principal components analysis essentially as a linear method for dimension reduction, in which we are generally interested in the first principal components. Through its application to statistical process control, we have seen that PCA can be also used as a multidimensional outlier detection technique, based on the last components.

Non-linear extensions of PCA exist and will be used more frequently [DEL 88, SCH 99].

1.7. Bibliography

[BAR 84] BARNETT V., LEWIS T., *Outliers in Statistical Data*, Wiley, New York, 1984.

[CAU 03] CAUSSINUS H., FEKRI M., HAKAM S., RUIZ-GAZEN A., "A monitoring display of multivariate outliers", *Computational Statistics and Data Analysis*, vol. 44, num. 1–2, p. 237–252, 2003.

[DEL 88] DE LEEUW J., VAN RIJCKEVORSEL J. L. A., *Component and Correspondence Analysis: Dimension Reduction by Functional Approximation*, Wiley, New York, 1988.

[GNA 77] GNANADESIKAN R., *Methods for Statistical Data Analysis of Multivariate Observations*, Wiley, New York, 1977.

[GOW 96] GOWER J. C., HAND D. J., *Biplots*, Chapman & Hall, London, 1996.

[HAW 74] HAWKINS D. M., "The detection of errors in multivariate data using principal component", *Journal of the American Statistical Association*, vol. 69, num. 346, 1974.

[JOL 86] JOLLIFFE I. T., *Principal Component Analysis*, Springer-Verlag, New York, 1986.

[LEB 06] LEBART L., MORINEAU A., PIRON M., *Statistique Exploratoire Multidimensionnelle*, Dunod, Paris, 4th edition, 2006.

[MON 85] MONTGOMERY D. C., *Introduction to Statistical Quality Control*, Wiley, New York, 1985.

[NIA 94] NIANG N. N., Méthodes multidimensionnelles pour la maîtrise statistique des procédés, PhD thesis, University of Paris Dauphine, France, 1994.

[NIA 02] NIANG N. N., "Multidimensional methods for statistical process control: some contributions of robust statistics", LAURO C., ANTOCH J., ESPOSITO V., SAPORTA G., Eds., *Multivariate Total Quality Control*, Heidelberg, Physica-Verlag, p. 136–162, 2002.

[SAP 06] SAPORTA G., *Probabilités, Analyse de Données et Statistique*, Technip, Paris, 2006.

[SCH 99] SCHÖLKOPF B., SMOLA A., MULLER K., "Kernel principal component analysis", SCHÖLKOPF B., BURGES C., SMOLA A., Eds., *Advances in Kernel Methods – Support Vector Learning*, MIT Press, p. 327–352, 1999.