Analyse en composantes principales

Gilbert Saporta, Ndèye Niang

Chapitre 1

Analyse en composantes principales Application à la maîtrise statistique des procédés

1.1. Introduction

L'Analyse en composantes principales est une méthode statistique exploratoire permettant une description essentiellement graphique de l'information contenue dans des grands tableaux de données. Dans la plupart des applications, il s'agit d'étudier p variables mesurées sur un ensemble de n individus. Lorsque n et p sont grands on cherche à synthétiser la masse d'informations sous une forme exploitable et compréhensible. Grâce aux outils de la statistique descriptive, il est possible d'étudier une à une ou deux à deux les variables à travers notamment des résumés graphiques ou numériques (moyenne, variance, corrélation). Mais ces études préalables simples, si elles sont indispensables dans toute étude statistique, sont insuffisantes ici car elles laissent de côté les liaisons éventuelles entre les variables qui sont souvent l'aspect le plus important.

L'analyse en composantes principales notée ACP par la suite, est souvent considérée comme la méthode de base de l'analyse factorielle des données dont l'objectif est de déterminer des fonctions des p variables ou facteurs qui serviront à visualiser les observations de façon simplifiée. En ramenant un grand nombre de variables souvent corrélées entre elles, à un petit nombre de composantes principales (les premières) non

Chapitre rédigé par Gilbert SAPORTA et Ndèye NIANG.

corrélées, l'ACP est une méthode de réduction de la dimension. Nous verrons cependant que l'ACP peut aussi être utilisée comme une méthode de détection de valeurs aberrantes multidimensionnelles notamment par l'exploitation des dernières composantes. Cette propriété se révèle utile en contrôle de qualité multidimensionnel.

1.2. Tableau de données et espaces associés

1.2.1. *Les données et leurs caractéristiques*

Les données sont généralement sous la forme d'un tableau rectangulaire à n lignes représentant les individus et à p colonnes correspondant aux variables. Le choix des individus et des variables est une phase essentielle qui influence dans une large mesure les résultats d'une ACP. Ce choix doit être fait en fonction des buts de l'étude ; les variables doivent notamment décrire le plus possible les phénomènes que l'on cherche à mettre en évidence.

Le plus souvent, l'ACP traite des variables numériques obtenues à l'issue de mesures. Mais elle permet aussi le traitement de variables ordinales comme des notes ou des rangs. Nous verrons dans la suite la notion de variables supplémentaires qui permet d'intégrer, *a posteriori*, des variables qualitatives nominales.

1.2.1.1. *Le tableau de données*

On note X la matrice de dimension (n, p) contenant les observations :

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

où x_i^j est la valeur de l'individu i pour la variable j que l'on notera $\mathbf{x}^j$ et qui sera identifiée au vecteur de n composantes $(x_1^j, \dots, x_n^j)'$. De même l'individu i sera identifié au vecteur $\mathbf{x}_i$ à p composantes avec $\mathbf{x}_i = (x_i^1, \dots, x_i^p)'$.

Le tableau 1.1 est un exemple d'un tel tableau de données ; il servira à l'illustration de nos explications. Les traitements seront effectués grâce au logiciel SPAD, version 5[1].

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

NOM	PAYS	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	GB	M	P	97	1.7	7.7	1	4	26.4	236	16
Hildon blanche	GB	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	GB	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	GB	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
Vernet	F	M	G	33.5	17.6	192	28.7	14	1	734	6.4
sidi harazem	MA	S	P	70	40	120	8	20	4	335	220
sidi ali	MA	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

Tableau 1.1. *Exemple de données*

Le fichier comprend 57 marques d'eau en bouteilles décrites par 11 variables explicitées dans le tableau 1.2. Les données proviennent des informations fournies sur

NOM	Nom complet de l'eau inscrit sur l'étiquette
PAYS	Pays d'origine : immatriculation automobile officielle ; parfois, une lettre supplémentaire est nécessaire, par exemple la Crête : GRC (Grèce Crête)
TYPE	M comme eau minérale ou S comme eau de source
PG	P comme eau plate, G comme eau gazeuse
CA	ions calcium en mg/litre
MG	ions magnésium en mg/litre
NA	ions sodium en mg/litre
K	ions potassium en mg/litre
SUL	ions sulfates en mg/litre
NO3	ions nitrates en mg/litre
HCO3	ions carbonates en mg/litre
CL	ions chlorures en mg/litre

Tableau 1.2. *Descriptif des variables*

les étiquettes des bouteilles. Les variables numériques forment un ensemble homogène et de ce fait elles seront considérées toutes comme des variables actives dans l'analyse (voir paragraphe 1.4.3). Si l'on disposait de variables de nature différente, par exemple du prix de la bouteille, on mettrait cette variable en supplémentaire. Par contre les variables qualitatives PAYS, TYPE, PG sont forcément mises en supplémentaire.

1.2.1.2. *Résumés numériques associés*

1.2.1.2.1. Point moyen ou centre de gravité

Le vecteur $\mathbf{g}$ des moyennes arithmétiques de chacune des p variables définit le point moyen. On a $\mathbf{g} = (\overline{x}^1, \ldots, \overline{x}^p)'$ avec $\overline{x}^j = \sum_{i=1}^n p_i x_i^j$.

Si les données ont été recueillies à la suite d'un tirage aléatoire à probabilités égales, les n individus ont tous la même importance dans les calculs des caractéristiques de l'échantillon. On leur affecte donc un poids $p_i = 1/n$.

Il peut cependant être utile pour certaines applications de travailler avec des p_i différents d'un individu à l'autre (échantillons redressés, données groupées,...). Ces poids qui sont des nombres positifs de somme 1 comparables à des fréquences sont regroupés dans une matrice diagonale de taille n :

$$D_p = \begin{pmatrix} p_1 & & \\ & \ddots & \\ & & p_n \end{pmatrix}$$

On a alors les expressions matricielles suivantes : $\mathbf{g} = X' D_p \mathbb{1}$ où $\mathbb{1}$ représente le vecteur de $\mathbb{R}^p$ dont toutes les composantes sont égales à 1. On obtient alors le tableau Y centré associé à X tel que $y_i^j = x_i^j - \overline{x}^j$ et on a : $Y = X - \mathbb{1}\mathbf{g}' = (I - \mathbb{1}\mathbb{1}' D_p) X$.

1.2.1.2.2. Matrice de variance et matrice de corrélation

On note $s_j^2 = \sum_{i=1}^{n} p_i (x_i^j - \overline{x}^j)^2$ et $v_{k\ell} = \sum_{i=1}^{n} p_i (x_i^k - \overline{x}^k)(x_i^\ell - \overline{x}^\ell)$ respectivement la variance de la variable j et la covariance des variables k et ℓ. On les regroupe dans la matrice de variance notée V. On a $V = X' D_p X - \mathbf{g}\mathbf{g}' = Y' D_p Y$.

De même, on définit le coefficient de corrélation linéaire entre les variables k et ℓ par $r_{k\ell} = \frac{v_{k\ell}}{s_k s_\ell}$. En notant Z le tableau centré réduit tel que $z_i^j = \frac{x_i^j - \overline{x}^j}{s^j}$, on a avec $Z = Y D_{1/s}$ avec $D_{1/s}$ la matrice diagonale des inverses des écarts-types :

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

On note R la matrice regroupant les coefficients de corrélation linéaire entre les p variables prises deux à deux ; on a $R = D_{1/s} V D_{1/s}$. R est la matrice de variance des données centrées réduites. Elle résume la structure des dépendances linéaires entre les p variables.

Les tableaux 1.3 et 1.4 contiennent les résumés numériques associés à l'exemple du tableau de données.

VARIABLE	MOYENNE	ECART-TYPE	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

Tableau 1.3. *Statistiques sommaires des variables continues*

	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

Tableau 1.4. *Matrice des corrélations*

1.2.2. *Espace des individus*

Dans l'approche géométrique de Pearson, on associe aux données un nuage de points : chaque individu étant défini par p coordonnées est alors considéré comme un élément d'un espace vectoriel de dimension p, appelé espace des individus. Le point moyen $\mathbf{g}$ défini au paragraphe 1.2.1.2 est alors le centre de gravité du nuage de points.

Le principe de l'ACP est de visualiser le plus fidèlement possible, dans un espace de faible dimension, ce nuage de points. L'analyse repose sur des distances entre les points représentant les individus. La méthode de calcul des ces distances influence dans une large mesure les résultats de l'analyse. Il est donc essentiel de la déterminer avant toute étude.

1.2.2.1. *La métrique*

Dans la géométrie de l'espace physique usuel à 3 dimensions, le calcul est simple par application de la formule de Pythagore. Par contre, en statistique le problème est compliqué car comment calculer des distances entre individus décrits par des variables exprimées dans des unités aussi différentes que des francs, des kg ou des km ? La formule de Pythagore est aussi arbitraire qu'une autre. On utilisera donc la formulation générale suivante : M étant une matrice symétrique définie positive de taille p, la distance entre deux individus $\mathbf{x}_i$ et $\mathbf{x}_j$ est définie par la forme quadratique :

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

En théorie, le choix de M dépend de l'utilisateur qui seul peut préciser la métrique adéquate. Mais en pratique, les métriques usuelles en ACP sont la métrique $M = I$ à utiliser si les variances ne sont pas trop différentes ou si les unités de mesures sont identiques, et la métrique $M = D_{1/s^2}$, métrique diagonale des inverses des variances, dans le cas contraire. Cette dernière est la plus utilisée (c'est l'option par défaut de beaucoup de logiciels d'ACP) car en plus de permettre de s'affranchir des unités de mesure, elle donne à chaque caractère la même importance quelle que soit sa dispersion dans le calcul des distances. En effet, cela revient à réduire les variables ce qui les rend sans dimension et toutes de même variance 1. Ainsi dans l'exemple, les écarts-types des huit variables étant très différents (figure 1.3), les variables seront réduites.

REMARQUE.– Toute matrice symétrique positive M peut s'écrire $M = TT'$. On a alors : $\mathbf{x}_i' M \mathbf{x}_j = \mathbf{x}_i' T' T \mathbf{x}_j = (T\mathbf{x}_i)'(T\mathbf{x}_j)$. Il est donc équivalent de travailler avec X et la métrique M ou avec la métrique I et le tableau transformé XT'. Ainsi, l'ACP usuelle consistera à réduire les variables et à utiliser la métrique I. C'est ce qu'on appelle une ACP normée.

1.2.2.2. *L'inertie*

C'est une notion fondamentale de l'ACP. On appelle inertie totale du nuage de points la moyenne pondérée des carrés des distances des points au centre de gravité,

Elle mesure la dispersion du nuage autour de son centre de gravité. On note

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

On montre que l'inertie en un point $\mathbf{a}$ quelconque définie par

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

se décompose suivant la relation de Huygens

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

Par ailleurs, on peut montrer que l'inertie totale est égale à la moitié de la moyenne des carrés de toutes les distances entre les n individus. Mais l'égalité la plus utilisée est la suivante :

$$I = trace(MV) = trace(VM).$$

Ainsi lorsque $M = I$, l'inertie est égale à la somme des variances des p variables et dans le cas de la métrique $M = D_{1/s^2}$, elle est égale à la trace de la matrice de corrélation, c'est-à-dire à p le nombre de variables ; l'inertie ne dépend alors pas des valeurs des variables mais uniquement de leur nombre.

Dans la suite de cet article, on se placera dans ce cas. Pour l'ACP générale avec une métrique M quelconque, on pourra se reporter aux ouvrages de Saporta [SAP 90] ou de Lebart, Morineau et Piron [LEB 95].

1.2.3. *Espace des variables*

Chaque variable est définie par n coordonnées ; on la considère alors comme un vecteur d'un espace à n dimensions appelé espace des variables. Pour le calcul des « distances » entre variables, on utilise la métrique D_p diagonale des poids qui possède, lorsque les variables sont centrées, les propriétés suivantes :

– le produit scalaire de deux variables $\mathbf{x}^k$ et $\mathbf{x}^\ell$ est $(\mathbf{x}^k)' D_p \mathbf{x}^\ell = \sum_{i=1}^{n} p_i x_i^k x_i^\ell$ qui est égal à la covariance $v_{k\ell}$,

– le carré de la norme d'une variable est égal alors à sa variance $||\mathbf{x}^j||_D^2 = s_j^2$ et l'écart-type représente la « longueur » de la variable,

– en notant $\theta_{k\ell}$ l'angle entre deux variables, on a $\cos\theta_{k\ell} = \frac{<\mathbf{x}^k, \mathbf{x}^\ell>}{||\mathbf{x}^k|| ||\mathbf{x}^\ell||} = \frac{v_{k\ell}}{s_k s_\ell} = r_{k\ell}$, le coefficient de corrélation linéaire.

Dans l'espace des variables, on s'intéressera donc aux angles plutôt qu'aux distances et on représentera les variables comme des vecteurs plutôt que comme des points.

1.3. L'analyse en composantes principales

1.3.1. *Principe général de la méthode*

Rappelons que le but de l'ACP est de fournir des représentations synthétiques de vastes ensembles de données numériques essentiellement sous forme de visualisations graphiques planes. Les espaces initiaux de représentations des individus et des variables étant de trop grandes dimensions, il est impossible d'y visualiser le nuage de points. On cherche donc des espaces de dimension réduite qui ajustent au mieux le nuage de points, c'est-à-dire qui respectent le plus possible la configuration initiale.

La méthode consiste à projeter le nuage de points en minimisant les déformations des distances inhérentes à la projection. Cela revient à choisir l'espace de projection F qui maximise le critère

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

Le sous-espace recherché est tel que la moyenne des carrés des distances entre points projetés est maximale (la projection raccourcit les distances), c'est-à-dire qu'il faut que l'inertie du nuage projetée soit maximale.

On montre [SAP 90] que la recherche du sous espace F peut s'effectuer de façon séquentielle : on cherche d'abord le sous-espace de dimension 1 d'inertie maximale, puis le sous-espace de dimension 1 orthogonal au précédent d'inertie maximale et ainsi de suite.

1.3.2. *Facteurs principaux et composantes principales*

Nous commençons par la recherche d'un sous-espace de dimension 1 représenté par une droite définie par un vecteur unitaire $\mathbf{u} = (u_1, \ldots, u_p)'$. Comme on l'a expliqué au paragraphe précédent, le vecteur doit être tel que les projections des points sur cette direction aient une inertie maximale. La projection ou coordonnée c_i d'un individu i sur Δ est définie par : $c_i = \sum_{j=1}^{p} \mathbf{x}_i^j u_j$ (figure 1.1). La liste des coordonnées c_i des individus sur Δ forme une nouvelle variable artificielle $\mathbf{c} = (c_1, \ldots, c_n)' = \sum_{j=1}^{p} \mathbf{x}^j u_j = X\mathbf{u}$; c'est une combinaison linéaire des variables initiales. L'inertie (ou variance) des points projetés sur Δ s'écrit donc :

$$\text{var}(\mathbf{c}) = \sum_{i}^{n} p_i c_i^2 = \mathbf{c}' D \mathbf{c} = \mathbf{u}' X' D X \mathbf{u} = \mathbf{u}' V \mathbf{u}.$$

Rappelons qu'on se limite au cas usuel de l'ACP normée ; la matrice de variance des données centrées réduites correspond donc à la matrice de corrélation R. Le critère de maximisation de l'inertie des points projetés sur Δ s'écrit alors :

$$\max_{\mathbf{u}} \mathbf{u}' V \mathbf{u} = \max_{\mathbf{u}} \mathbf{u}' R \mathbf{u}$$

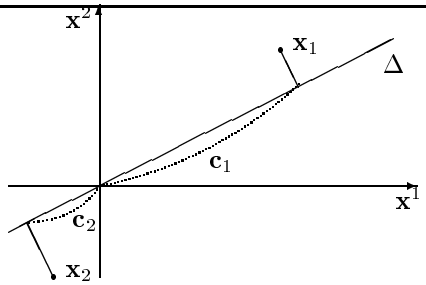

Figure 1.1. *Projection sur une direction* Δ

sous la contrainte $\mathbf{u}'\mathbf{u} = 1$.

On montre que la solution de ce problème de maximisation d'une forme quadratique est $\mathbf{u}_1$ le vecteur propre de R associé à la plus grande valeur propre λ_1. On cherche ensuite le vecteur $\mathbf{u}_2$ orthogonal à $\mathbf{u}_1$ tel que l'inertie des points projetés sur cette direction soit maximale. De façon analogue, on montre que $\mathbf{u}_2$ est le vecteur propre de R associé à la deuxième plus grande valeur propre λ_2. Plus généralement le sous-espace à q dimensions recherché est engendré par les q premiers vecteurs propres de la matrice R associés aux plus grandes valeurs propres.

Les vecteurs $\mathbf{u}_j$ sont appelés les *facteurs principaux*. Ils contiennent les coefficients des variables initiales dans la combinaison linéaire $\mathbf{c} = X\,\mathbf{u}$.

Les composantes principales sont les variables artificielles définies par les facteurs principaux : $\mathbf{c}^j = X\mathbf{u}_j$; elles contiennent les coordonnées des projections orthogonales des individus sur les axes définis par les $\mathbf{u}_j$.

En pratique, l'ACP va donc consister à diagonaliser la matrice R pour obtenir les $\mathbf{u}_j$ et à calculer les composantes principales $\mathbf{c}^j = X\,\mathbf{u}_j$.

Les résultats pour l'exemple des eaux figurent dans les tableaux 1.5 et 1.6.

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE
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

Tableau 1.5. *Valeurs propres* λ_1, λ_2,...

VARIABLES	VECTEURS PROPRES				
IDEN - LIBELLE COURT	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

Tableau 1.6. *Vecteurs propres*

1.3.3. *Propriétés des facteurs principaux et des composantes principales*

1.3.3.1. *Variance d'une composante principale*

La variance d'une composante principale est égale à la valeur propre λ : var$(\mathbf{c}_i) = \lambda_i$. En effet, par définition on a $V\mathbf{u} = R\mathbf{u}$ et $\mathbf{u}'\mathbf{u} = 1$ et donc

$$\text{var}(\mathbf{c}) = \mathbf{c}'D\mathbf{c} = \mathbf{u}'X'DX\mathbf{u} = \mathbf{u}'V\mathbf{u} = \mathbf{u}'R\mathbf{u} = \mathbf{u}'(\lambda\mathbf{u}) = \lambda\mathbf{u}'\mathbf{u} = \lambda.$$

Ainsi les composantes principales sont les combinaisons linéaires des variables initiales de variance maximale.

1.3.3.2. *Propriété supplémentaire*

$\mathbf{c}^1$ est la variable la plus liée aux $\mathbf{x}^j$ au sens de la somme des carrés des corrélations : $\sum_{j=1}^{p} r^2(\mathbf{c}, \mathbf{x}^j)$ est maximal. En effet, on montre [SAP 90] que

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

où Z est le tableau centré réduit. Le maximum de ce quotient est donc atteint pour $\mathbf{c}$ vecteur propre de $ZZ'D$ associé à sa plus grande valeur propre : $ZZ'D\mathbf{c} = \lambda\mathbf{c}$.

La composante principale $\mathbf{c}$ est donc combinaison linéaire des colonnes de Z : $\mathbf{c} = Z\mathbf{u}$ et alors $ZZ'D\mathbf{c} = \lambda\mathbf{c}$ devient $ZZ'DZ\mathbf{u} = \lambda Z\mathbf{u}$. Mais on a $Z'DZ = R$ d'où : $ZR\mathbf{u} = \lambda Z\mathbf{u}$ et si Z est de rang p : $R\mathbf{u} = \lambda\mathbf{u}$.

1.3.3.3. *Formules de reconstitution*

On a $X\mathbf{u}_j = \mathbf{c}^j$; en post-multipliant les deux membres de cette relation par $\mathbf{u}'_j$ et en sommant sur j, il vient $X\sum_{j=1}^{p}\mathbf{u}_j\mathbf{u}'_j = \sum_{j=1}^{p}\mathbf{c}^j\mathbf{u}'_j$. On montre facilement que $\sum_{j=1}^{p}\mathbf{u}_j\mathbf{u}'_j = I$ car les $\mathbf{u}_j$ sont orthonormés. On trouve alors $X = \sum_{j=1}^{p}\mathbf{c}^j\mathbf{u}'_j$. On peut donc ainsi reconstituer le tableau de données (centré) à partir des facteurs et composantes principales. Si l'on se contente de sommer sur les q premiers termes

correspondant aux q plus grandes valeurs propres, on obtient alors la meilleure approximation de X par une matrice de rang q au sens des moindres carrés (théorème d'Eckart-Young).

En résumé, on peut dire que l'ACP consiste à transformer les variables initiales $\mathbf{x}^j$ corrélées en de nouvelles variables, les composantes principales $\mathbf{c}^j$ combinaisons linéaires des $\mathbf{x}^j$ non corrélées entre elles, de variance maximale et les plus liées aux $\mathbf{x}^j$: l'ACP est une méthode factorielle linéaire.

Il existe des extensions de l'ACP au cas non linéaire : on recherche des transformations des variables par exemple par des fonctions splines [DEL 88] disponibles dans certains logiciels (procédure prinqual de SAS).

1.4. Interprétation des résultats d'une ACP

L'ACP fournit des représentations graphiques permettant de visualiser les relations entre variables ainsi que l'existence éventuelle de groupes d'individus et de groupes de variables. Les résultats d'une ACP se présentent sous la forme de graphiques plans et de tableaux dont l'interprétation constitue une des phases les plus délicates de l'analyse et doit se faire selon une démarche précise que nous expliquons dans la suite.

Avant d'aborder la phase d'interprétation proprement dite, il est utile de commencer par une brève lecture préliminaire des résultats dont le but est en gros de s'assurer du contenu du tableau de données. En effet, il est possible qu'en examinant le premier plan principal on observe quelques individus complètement extérieurs au reste de la population traduisant la présence soit de données erronées telles que des fautes de frappes ou une erreur de mesure qu'il faut corriger soit d'individus totalement différents des autres qu'il faut retirer de l'analyse pour mieux observer les individus restants ; on pourra les réintroduire *a posteriori* comme éléments supplémentaires.

À la suite de cette étude préalable, on peut alors examiner de plus près les résultats de l'ACP ; on passe à la phase d'interprétation qui comporte plusieurs étapes.

REMARQUE.– Bien qu'il existe une manière de représenter simultanément individus et variables appelée « biplot » [GOW 96], nous préconisons de représenter séparément les deux ensembles pour éviter des confusions.

1.4.1. *Qualité des représentations sur les plans principaux*

L'ACP permet d'obtenir une représentation graphique des individus dans un espace de dimension plus faible que p mais celle-ci n'est qu'une vision déformée de la réalité. L'un des points les plus délicats de l'interprétation des résultats d'une ACP

consiste à apprécier cette déformation ou autrement dit la perte d'information engendrée par la réduction de la dimension et à déterminer le nombre d'axes à retenir.

Le critère habituellement utilisé pour mesurer la qualité d'une ACP est le pourcentage d'inertie totale expliquée. Il est défini par :

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

C'est une mesure globale qui doit être complétée par d'autres considérations. D'abord au niveau des variables initiales, leur nombre doit être pris en compte : un pourcentage de 10% n'a pas le même intérêt sur un tableau de 20 variables et un tableau de 100 variables.

Ensuite, pour les individus, il faut envisager pour chacun d'eux la qualité de sa représentation indépendamment du pourcentage d'inertie global. En effet, il est possible d'avoir un premier plan principal F_2 avec une inertie totale importante et qu'en projection deux individus soient très proches, cette proximité peut être illusoire si les deux individus sont éloignés dans $F_2^{\perp}$ (figure 1.2).

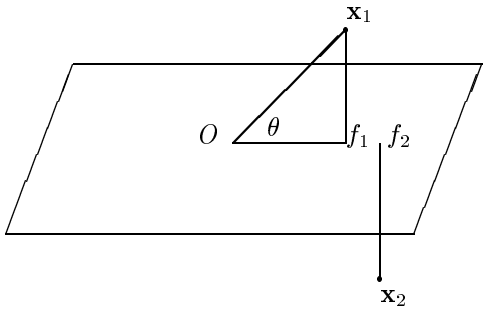

Figure 1.2. *Projections proches de points éloignés*

La mesure la plus fréquente de la qualité de représentation d'un individu est le cosinus de l'angle entre le plan principal et le vecteur $\mathbf{x}_i$. Si ce cosinus est grand, $\mathbf{x}_i$ sera voisin du plan, on pourra alors examiner la position de sa projection sur le plan par rapport à d'autres points ; si ce cosinus est faible, on se gardera de toute conclusion.

1.4.2. *Nombre d'axes à retenir*

C'est un point essentiel de l'ACP mais qui n'a pas de solution rigoureuse. Il existe des critères théoriques qui reposent sur des tests statistiques ou des intervalles de confiance sur les valeurs propres mais ces derniers ne sont utilisables qu'en ACP sur données non-réduites dans le cas gaussien p-dimensionnel. Ainsi, dans le cas pratique

le plus fréquent des matrices de corrélation, les seuls critères applicables sont des critères empiriques dont le plus connu est celui de Kaiser : en données centrées réduites, on retient les composantes principales correspondant à des valeurs propres supérieures à 1 ce qui revient à ne s'intéresser qu'aux composantes qui « apportent » plus que les variables initiales.

On utilise aussi la règle du coude qui consiste à détecter sur le diagramme de valeurs propres l'existence d'un coude. Mais ceci n'est pas toujours aisé en pratique.

Pour l'exemple des eaux, l'utilisation conjointe de la règle de Kaiser et de l'histogramme des valeurs propres (tableau 1.7, qui présente un coude après la deuxième valeur propre, nous conduit à retenir deux axes correspondant à un pourcentage d'inertie expliquée de 73.56%. Le troisième axe s'interprète aisément mais comme il s'identifie à la variable NO3 (corrélation de -0.96), et n'est pas corrélé aux autres , son intérêt est limité.

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
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	*

Tableau 1.7. *Histogramme des valeurs propres*

1.4.3. *Interprétation interne*

Les résultats d'une ACP sont obtenus à partir des variables et individus appelés éléments actifs par opposition aux éléments supplémentaires qui ne participent pas directement à l'analyse. Les variables et individus actifs servent à calculer les axes principaux ; les variables et individus supplémentaires sont projetés ensuite sur ces axes. Les variables actives (numériques) sont celles dont les intercorrélations sont intéressantes : ce sont les variables principales de l'étude. Les variables supplémentaires apportent une information utile pour caractériser les individus mais ne servent pas directement à l'analyse : on ne s'intéresse pas aux corrélations des variables supplémentaires entre elles mais seulement à leurs corrélations avec les variables actives à travers les composantes principales. L'interprétation interne consiste à étudier les résultats en se basant sur les variables et les individus actifs. L'étude des éléments supplémentaires se fait à travers la phase d'interprétation externe.

1.4.3.1. *Les variables*

L'ACP construit les composantes principales, nouvelles variables artificielles combinaisons linéaires des variables initiales. Interpréter une ACP, c'est donner une signification à ces composantes principales (en fonction des variables initiales). Cela se

fait de façon naturelle à travers le calcul de coefficients de corrélation linéaire $r(\mathbf{c}, \mathbf{x}^j)$ entre composantes principales et variables initiales ; et on s'intéresse aux coefficients les plus forts en valeur absolue et proches de 1 (tableau 1.8.)

Dans le cas usuel de l'ACP normée, on travaille avec des données centrées réduites et le calcul de $r(\mathbf{c}, \mathbf{x}^j)$ est particulièrement simple. On montre en effet que $r(\mathbf{c}, \mathbf{x}^j) = \sqrt{\lambda}\mathbf{u}_j$.

VARIABLES	COORDONNEES				
IDEN - LIBELLE COURT	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

Tableau 1.8. *Corrélations variable-facteur ou coordonnées des variables*

On synthétise usuellement les corrélations des variables pour un couple de composantes sur un graphique appelé cercle de corrélation sur lequel, chaque variable $\mathbf{x}^j$ est repérée par une abscisse $r(\mathbf{c}^1, \mathbf{x}^j)$ et une ordonnée $r(\mathbf{c}^2, \mathbf{x}^j)$. L'examen du cercle de corrélation permet de détecter les éventuels groupes de variables qui se ressemblent ou au contraire qui s'opposent donnant ainsi un sens aux axes principaux.

Ainsi pour notre exemple illustratif (figure 1.3), l'axe 1 est corrélé négativement à toutes les variables (à l'exception de NO3 qui n'est pas significatif) : les eaux les plus à gauche dans le plan factoriel sont celles les plus chargées en minéraux. L'axe 2 oppose les eaux à forte teneur en calcium et sulfates à celles riches en potassium et carbonates.

REMARQUE.– Effet « Taille » : Lorsque toutes les variables initiales sont positivement corrélées entre elles, la première composante principale définit « un facteur de taille ». En effet, on sait qu'une matrice symétrique ayant tous ses termes positifs admet un premier vecteur propre dont toutes les composantes sont de même signe. On peut les choisir positifs et alors la première composante sera positivement corrélée avec toutes les variables et les individus sont rangés sur l'axe 1 par valeurs croissantes de l'ensemble des variables (en moyenne).

La deuxième composante principale différencie alors des individus de « taille » semblable ; on l'appelle facteur de forme.

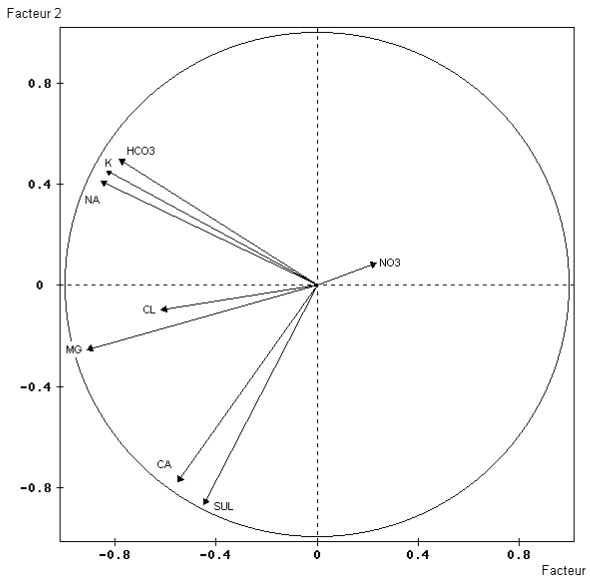

Figure 1.3. *Représentation des variables dans le plan 1,2*

1.4.3.2. *Les individus*

L'étude des individus consiste à examiner leurs coordonnées et surtout les représentations graphiques, appelés plans principaux (figure 1.4) qui en résultent, le but étant de voir comment se répartissent les individus, quels sont ceux qui se ressemblent ou qui se distinguent. Dans le cas où les individus ne sont pas anonymes, ils aideront ainsi à donner un sens aux axes principaux ; par exemple on recherchera des individus opposés le long d'un axe. Inversement, l'utilisation des résultats de l'étude des variables permet d'interpréter les individus. En effet, lorsque par exemple la première composante est très corrélée avec une variable initiale, cela signifie que les individus ayant une forte coordonnée positive sur l'axe 1 sont caractérisés par une valeur de la variable nettement supérieure à la moyenne (l'origine des axes représente le centre de gravité du nuage de points).

Dans cette étude des individus, il est aussi très utile de s'intéresser pour chaque axe aux contributions apportées par les différents individus car elles peuvent aider à l'interprétation des axes. Elle est définie par $\frac{p_i c_{ki}^2}{\lambda_k}$ où c_{ik} où représente la valeur pour l'individu i de la k^{e} composante $\mathbf{c}_k$ et $\lambda_k = \sum_{i=1}^{n} p_i c_{ki}^2$.

On s'intéressera aux contributions importantes c'est-à-dire celles qui excèdent le poids de l'individu concerné.

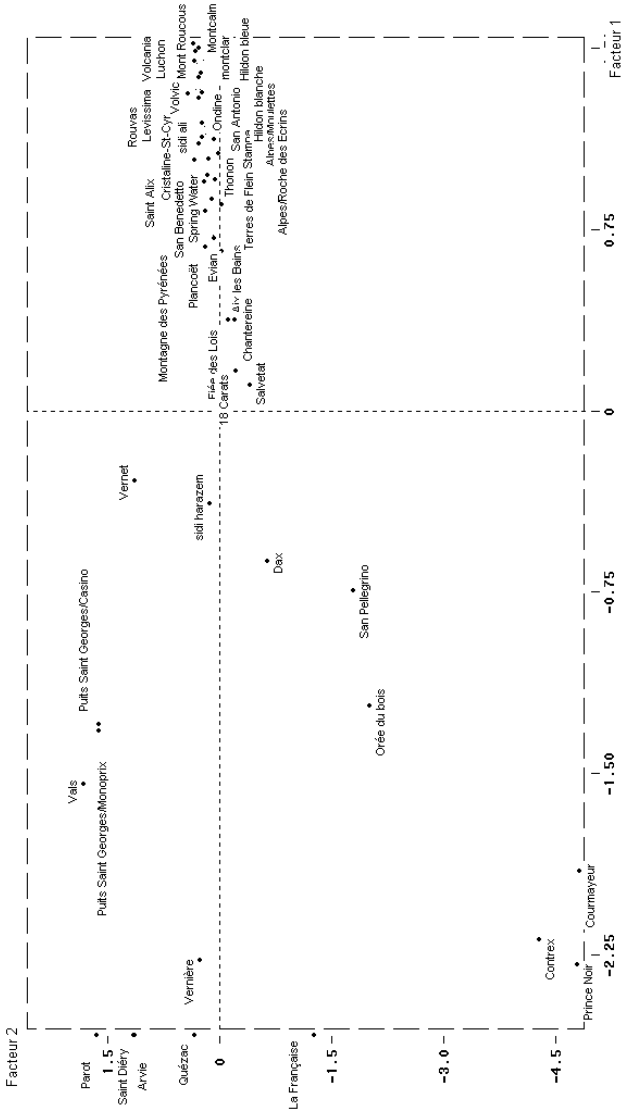

Figure 1.4. *Représentation des individus dans le plan 1,2*

Cependant, il faut prendre garde à un individu ayant une contribution excessive qui constitue un facteur d'instabilité : le fait de le retirer peut modifier profondément le résultat de l'analyse. On a alors intérêt à effectuer l'analyse en l'éliminant puis à le rajouter ensuite en élément supplémentaire si, bien sûr, il ne s'agit pas d'une donnée

erronée qui a été ainsi mise en évidence. Des points comme ARVIE et PAROT, sont des exemples de tels points.

Il est à noter que lorsque les poids sont tous égaux, les contributions n'apportent pas plus d'information que les coordonnées. Le tableau 1.9 donne, en plus des coordonnées et des contributions, les cosinus carrés des angles avec les axes principaux qui permettent de juger de la qualité de la représentation (voir le paragraphe 1.4.1).

INDIVIDUS			COORDONNEES					CONTRIBUTIONS					COSINUS CARRES				
IDENTIFICATEUR	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
Vernet	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

Tableau 1.9. *Coordonnées, contribution et cosinus carrés des individus*

1.4.4. *Interprétation externe : variables et individus supplémentaires*

Rappelons que les éléments supplémentaires n'interviennent pas dans les calculs de détermination des axes factoriels ; par contre ils sont très utiles *a posteriori* pour conforter et enrichir l'interprétation de ces axes.

On distingue le cas de variables numériques supplémentaires et celui des variables qualitatives supplémentaires. Pour les variables numériques supplémentaires on les place sur les cercles de corrélation après avoir simplement calculé le coefficient de corrélation entre chaque variable supplémentaire et les composantes principales. L'interprétation s'effectue de la même manière que pour les variables actives à travers la détection des corrélations significatives.

Pour les variables qualitatives supplémentaires, en général, on se contente de représenter chaque modalité par son centre de gravité. Certains logiciels (notamment SPAD) fournissent des aides à l'interprétation : les valeurs-test qui sont des mesures de l'éloignement du point représentatif d'une modalité par rapport à l'origine. Plus précisément, la valeur-test mesure cet éloignement en nombre d'écart-type d'une loi normale. Elles permettent de mettre en évidence une position excentrée d'un sous groupe d'individus. Une modalité sera considérée comme significative d'un axe si la valeur-test qui lui est associée est supérieure en valeur absolue à 2 (au risque 5%).

Ainsi dans l'exemple, le centre de gravité des eaux gazeuses (et par voie de conséquence celui des eaux plates) est à plus de 3 écarts-types de l'origine (-3,5) : les eaux gazeuses sont donc très significativement éloignées de l'origine.

En ce qui concerne les individus supplémentaires, il est immédiat de les positionner sur les axes principaux ; en effet à partir du moment où l'on dispose de la formule permettant de calculer les composantes principales, il suffit de calculer des combinaisons linéaires des caractéristiques de ces points supplémentaires.

MODALITES			VALEURS-TEST					COORDONNEES					
IDEN - LIBELLE	EFF.	P.ABS	1	2	3	4	5	1	2	3	4	5	DISTO.
1 . PAYS													
France	40	40.00	-1.9	0.7	2.1	-0.5	0.7	-0.33	0.09	0.18	-0.04	0.03	0.15
Grande-Bretagne	4	4.00	1.2	0.2	-2.7	0.5	-0.2	1.17	0.16	-1.28	0.22	-0.05	3.13
Grèce	2	2.00	0.8	0.2	-3.5	0.4	-1.0	1.09	0.15	-2.44	0.26	-0.29	7.36
Grèce-Crête	2	2.00	0.8	0.2	0.8	0.2	0.0	1.12	0.19	0.52	0.10	0.00	1.58
Italie	7	7.00	0.7	-1.5	0.4	-0.5	0.1	0.49	-0.77	0.13	-0.17	0.01	0.92
Maroc	2	2.00	0.3	0.2	0.6	1.0	-0.6	0.36	0.20	0.45	0.61	-0.19	0.86
2 . TYPE													
Minérale	38	38.00	-2.5	-0.5	-1.2	-0.7	0.4	-0.46	-0.07	-0.11	-0.06	0.01	0.23
Source	19	19.00	2.5	0.5	1.2	0.7	-0.4	0.92	0.13	0.22	0.11	-0.03	0.94
3 . PG													
Gazeuse	16	16.00	-3.5	2.7	-0.5	-1.8	0.3	-1.44	0.82	-0.11	-0.34	0.03	2.89
Plate	41	41.00	3.5	-2.7	0.5	1.8	-0.3	0.56	-0.32	0.04	0.13	-0.01	0.44

Tableau 1.10. *Coordonnées et valeurs-test des modalités*

1.5. Application à la maîtrise statistique des procédés

1.5.1. *Introduction*

Le contrôle statistique en cours de fabrication est essentiellement basé sur l'utilisation de cartes de contrôle dites aux mesures traçant l'évolution des caractéristiques d'un produit ou d'un procédé. La carte de contrôle est un outil qui permet de détecter à travers des prélèvements successifs de petits échantillons $(\mathbf{x}_i, i = 1, 2, \ldots, n)$ le déréglage d'un paramètre de centrage (moyenne) ou de dispersion (écart-type, étendue) par rapport à des valeurs de référence fixées.

Il existe plusieurs sortes de cartes de contrôle [MON 85, NIA 94] toutes basées sur l'hypothèse que les $\mathbf{x}_i$ sont distribués selon une loi normale $\mathcal{N}(\mu_0, \sigma_0)$; les valeurs cibles ou de référence μ_0 et σ_0 sont supposées connues ou fixées, dans le cas contraire, elles sont remplacées par des estimations sans biais.

Nous nous intéressons exclusivement aux cartes de Shewhart pour la détection des déréglages de la moyenne d'un procédé. Les cartes de contrôle de Shewhart classiques utilisent à chaque instant i la seule valeur $\overline{x}_i$ moyenne des n mesures dont on dispose à l'instant i que l'on compare aux limites inférieure (LCL) et supérieure (UCL) de contrôle :

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

Cette carte de contrôle peut être vue comme une représentation graphique d'une suite de tests $H_0 : \mu = \mu_0$ contre $H_1 : \mu \neq \mu_0$ pour une série d'échantillons ; on supposera que l'écart-type $\sigma = \sigma_0$ est connu. La région critique correspond à la zone hors des limites de la carte de contrôle. La carte de contrôle est donc la visualisation en fonction du temps de ce test répété. Cette équivalence avec les tests facilitera l'extension à plusieurs variables.

Dans la plupart des cas les entreprises ont, non pas une, mais plusieurs caractéristiques à contrôler simultanément. La pratique habituelle consiste à utiliser conjointement p cartes unidimensionnelles. Cette méthode à l'inconvénient majeur de ne pas tenir compte de la corrélation entre les variables représentant ces p caractéristiques. Cela conduit (figure 1.5) alors à des situations indésirables de fausses alertes : les cartes univariées signalent une sortie des limites alors que le procédé multivarié est sous contrôle (région B et C) ou des situations, plus graves, qui correspondent à une non détection d'un déréglage du procédé multivarié (région A).

L'approche globale à travers des cartes multivariées est donc la seule adéquate [NIA 02]. L'analyse en composantes principales, qui fournit des variables artificielles mais non corrélées, constitue dans une certaine mesure une première solution. Nous verrons dans la suite qu'une fois un déréglage détecté, des cartes univariées adéquates peuvent cependant aider à déterminer les variables responsables de ce déréglage, c'est ce que l'on appelle rechercher les causes assignables.

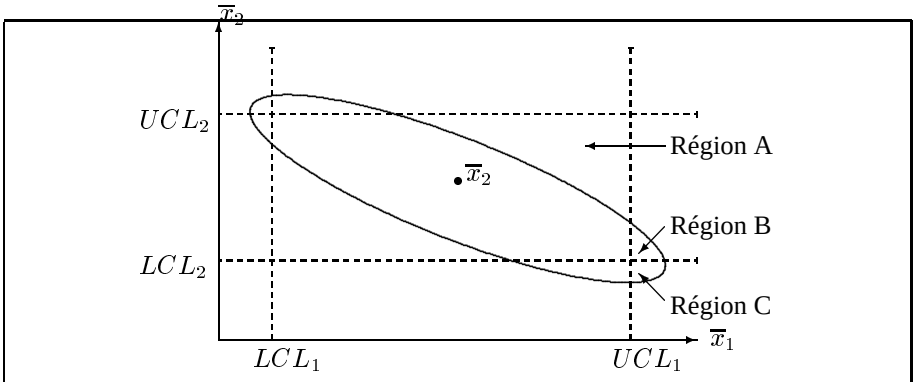

Figure 1.5. *Carte de contrôle multivarié*

1.5.2. *Cartes de contrôle et ACP*

1.5.2.1. *ACP et valeurs aberrantes*

Les cartes de contrôle multivariées considérées reposent sur une transformation d'un vecteur de $\mathbb{R}^p$ en un scalaire par l'intermédiaire d'une fonction quadratique. Elles peuvent être vues comme des méthodes de détection de valeurs aberrantes multidimensionnelles. En effet, ces méthodes consistent à trouver une relation de sous-ordre sur $\mathbb{R}^p$ basée généralement sur une mesure de distance multivariée (on trouvera une étude détaillée dans [BAR 84]). Celle-ci est alors utilisée dans un test sur la base duquel on décide qu'une observation est aberrante si la valeur de la statistique réduite est anormalement grande ou petite, sous une hypothèse de modèle (en contrôle de qualité, on fait souvent une hypothèse de normalité).

Dans le cas multidimensionnel, une valeur aberrante peut être le résultat d'une valeur extrême sur une seule des p composantes principales ou celui d'erreurs systématiques moyennement faibles sur différentes directions. Ce dernier type de valeurs aberrantes correspond à un problème au niveau de l'orientation (corrélation) et non à celui de la position (moyenne) ou de l'échelle (variance). L'utilisation de l'analyse en composantes principales non pas comme une méthode de réduction de la dimension mais de détection de valeurs aberrantes facilite la recherche de ces directions extrêmes.

On considère non seulement les premières composantes pour résumer au mieux la structure des données mais aussi les dernières considérées comme les résidus de l'analyse. En effet, Jolliffe [JOL 86] a montré que les premières composantes permettent de détecter des valeurs aberrantes qui détériorent, gonflent les variances et les covariances, mais de telles valeurs seront aussi extrêmes sur les variables initiales ; on pourra donc les détecter directement ; les premières composantes n'apportent pas d'information supplémentaire. Par contre, les dernières composantes principales permettent de détecter des valeurs aberrantes non visibles sur les variables initiales, celles qui violent la structure de corrélation entre les variables. De nombreuses méthodes de

détection de valeurs aberrantes utilisant l'ACP ont été proposées par plusieurs auteurs notamment Hawkins [HAW 74], Gnanadesikan [GNA 77], Jolliffe [JOL 86], Barnett et Lewis [BAR 84], ...

Les techniques proposées par ces auteurs consistent à faire des tests statistiques classiques sur les composantes principales prises séparément ou conjointement. Ces tests reposent sur des statistiques de résidus calculées sur les q dernières composantes principales. Les plus courantes sont

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

où $y^k = \frac{c^k}{\sqrt{\lambda_k}}$; R_{1i}^2 est une somme pondérée des composantes principales de l'ACP standardisée qui donnent plus d'importance aux composantes principales ayant de grandes variances et

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

Les distributions de ces statistiques sont aisément obtenues dans le cas où on suppose que les observations suivent une loi normale de moyenne μ et de variance Σ connus : y^k est distribué selon une loi normale $\mathcal{N}(0,1)$ et en l'absence de valeurs aberrantes, les statistiques des résidus R_{1i}^2 et R_{2i}^2 suivent alors des lois χ_q^2. Pour le cas μ et Σ inconnus on peut aussi, en utilisant leurs estimations, obtenir des distributions approximatives de ces statistiques. Il est donc possible de faire des tests.

1.5.2.2. *Cartes de contrôle associées aux composantes principales*

En contrôle de qualité, l'ACP est surtout utilisée comme une technique de détection des déréglages considérés comme des valeurs aberrantes. On utilisera donc les premières et les dernières composantes principales puisqu'on ne connaît pas *a priori* le sens ou la nature des déréglages.

Rappelons que les composantes principales sont par définition les combinaisons linéaires des variables initiales qui résument au mieux la structure des données. Elles tiennent compte de la corrélation entre les variables. Donc même prises individuellement elles aident à la détection des déréglages contrairement aux variables initiales.

Cependant les cartes sur les composantes principales ne doivent pas être utilisées à la place des cartes multivariées mais en conjonction avec ces dernières. Le problème de fausses alertes et de non détection, que l'on a mis en évidence sur la figure 1.5, est atténué mais pas complètement supprimé pour les composantes principales non corrélées.

Les limites de contrôle à 3 écarts-types pour les composantes principales réduites sont alors $\pm 3/\sqrt{n}$. On peut tester la présence de valeurs aberrantes avec une carte de contrôle R_{2i}^2 dont la limite supérieure correspond au fractile $1-\alpha$ d'un χ_q^2.

Pour les statistiques des résidus, on peut tester la présence de valeurs aberrantes à l'aide d'une carte de contrôle définie par :

$$UCL = \chi^2_{p-k}, \quad LCL = 0 \quad \text{et} \quad Stat = R^2_{2i}.$$

EXEMPLE.– Nous avons simulé 30 échantillons de taille 5 d'une loi normale $\mathcal{N}_3(0, R)$; les trois variables sont supposées centrées avec des variances égales respectivement à 0.5 ; 0.3 ; et 0.1 ; la matrice de corrélation est

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

On a ensuite simulé un déréglage de la moyenne sur les cinq derniers échantillons qui a consisté à augmenter la moyenne de la première variable et à diminuer celle de la deuxième variable de la moitié de leurs écarts-types.

Cette situation est détectée par la carte de contrôle multidimensionnelle adéquate [NIA 94] ainsi que par la carte sur la troisième et dernière composante principale (figure 1.6) : les cinq derniers points hors contrôle sont clairement mis en évidence alors que le phénomène est invisible sur les deux premières composantes.

Lorsque le nombre de caractéristiques est faible et s'il est possible de trouver une interprétation simple des composantes principales en fonction d'un petit nombre de variables parmi les p alors les cartes sur les composantes principales permettent non seulement de détecter les déréglages mais elles aident aussi à la détermination des causes assignables.

Par contre, lorsque le nombre de variables est très grand, les méthodes proposées nécessitent de nombreuses cartes de contrôle pour les premières et les dernières composantes, ce qui est gênant. On peut se contenter des q premières composantes comme dans l'approche ACP pour la réduction de la dimension mais il faudra d'une part, tester la qualité de la représentation des p variables initiales par les q composantes, et d'autre part, utiliser des méthodes basées sur les résidus pour la détection des valeurs aberrantes ou déréglages.

De plus, même si on arrive à trouver q composantes principales résumant au mieux l'information présente dans les p variables initiales, ces q composantes dépendent d'un grand nombre sinon de toutes les p variables de départ. Pour simplifier l'interprétation

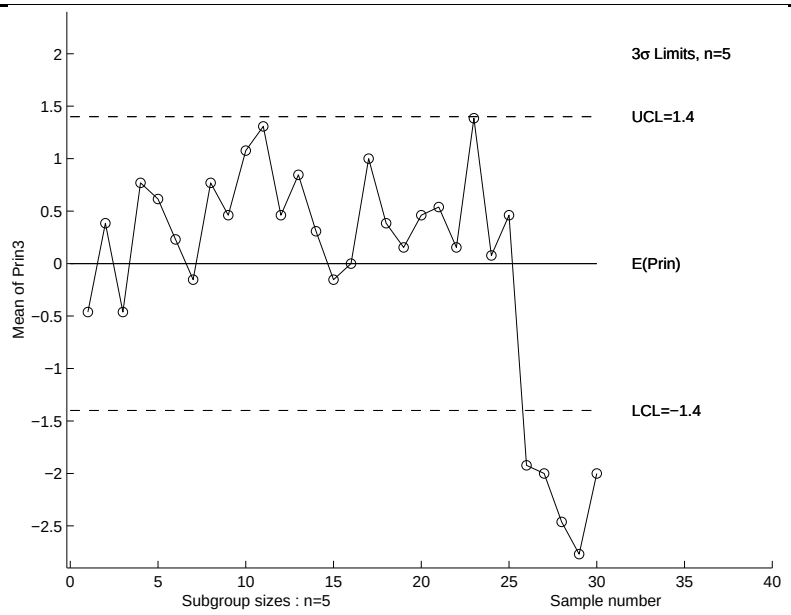

Figure 1.6. *Carte de contrôle pour la 3e composante principale (Prin 3)*

des composantes principales, d'autres méthodes de type Projection Pursuit ont été proposées [NIA 94]. Nous pensons que les travaux de Caussinus, Hakam et Ruiz-Gazen [CAU 02] peuvent être utiles à l'amélioration de ces méthodes.

1.6. Conclusion

L'ACP est une méthode très efficace pour représenter des données corrélées entre elles. Elle est largement utilisée dans des études de marché, d'opinion et de plus en plus dans le domaine industriel.

Nous avons présenté ici l'analyse en composantes principales essentiellement comme une méthode linéaire de réduction de la dimension dans laquelle on s'intéresse en général aux premières composantes principales. À travers son application à la maîtrise statistique des procédés, on a vu que l'ACP peut aussi être utilisée comme une technique de détection de valeurs aberrantes multidimensionnelles reposant plutôt sur les dernières composantes.

Des extensions non linéaires de l'ACP existent mais sont encore peu utilisées [DEL 88, SCH 99] et devraient se développer dans un avenir proche.

1.7. Bibliographie

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

[CAU 02] CAUSSINUS H., HAKAM S., RUIZ-GAZEN A., « Projections révélatrices contrôlées - recherche d'individus atypiques », *Revue de Statistique Appliquée*, vol. L, n° 4, p. 81-94, 2002.

[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, n° 346, June 1974.

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

[LEB 95] LEBART L., MORINEAU A., PIRON M., *Statistique exploratoire multidimensionnelle*, Dunod, Paris, 1995.

[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, Université Paris Dauphine, Paris, 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 90] SAPORTA G., *Probabilités, analyse de données et statistique*, Technip, Paris, 1990.

[SCH 99] SCHÖLKOPF B., SMOLA A., MULLER K., « Kernel Principal Component Analysis », B. SCHÖLKOPF C.J.C. BURGES A. S., Ed., *Advances in Kernel Methods – Support vector learning*, MIT Press, p. 327-352, 1999.