3. ANALYSE EN COMPOSANTES PRINCIPALES

L'analyse des données consiste essentiellement à établir quelles sont les relations existant entre les observations, entre les variables, et entre les observations et les variables. Il s'agit donc de mettre un peu d'ordre dans le fichier de données, souvent de taille considérable (quelques centaines ou milliers d'analyses et quelques dizaines de variables. Ceci peut être fait à l'aide de diagrammes binaires, à l'aide des corrélations et corrélations partielles, à l'aide de diagrammes spécifiques propres au domaine étudié (ex. diagramme AFM et diagrammes de Harker en pétrologie). L'analyse en composantes principales est un autre outil permettant une meilleure visualisation de nos données.

3.1 Méthodologie de l'ACP

Le tableau de données n x p forme un nuage de n points dans un espace à p dimensions, ou un nuage de p points dans un espace à n dimensions. Un diagramme binaire consiste à projeter ces points sur deux des dimensions choisies plus ou moins arbitrairement. L'ACP consiste à projeter les points sur une droite, un plan... un sous-espace à s dimensions (avec $s \leq p$) choisi de façon à optimiser un certain critère. Intuitivement, on cherchera le sous-espace donnant la meilleure visualisation possible de notre nuage de points. Un bon choix consiste à rechercher la plus grande dispersion (le plus grand étalement) possible des projections dans le sous-espace choisi. On est amené ainsi à chercher une rotation de notre système d'axes initial (les variables) permettant de mieux voir notre nuage. Définissons u_1 le vecteur unitaire (i.e. de norme 1; $u_1'u_1=1$) recherché; c'est le vecteur présentant la plus grande dispersion des projections.

Soit la matrice X_{nxp}; chaque ligne représente une observation. Chaque colonne représente une variable. On supposera chaque variable centrée, i.e. on a soustrait la moyenne de chaque variable au préalable. Ceci est fait de façon à faire coïncider le centre de gravité du nuage de points avec l'origine.

Les projections des n observations sur le vecteur u_1 sont données par:

$$C=Xu_1 \qquad \text{(produit scalaire)}$$

La somme des carrés de ces projections (inertie) est:

$$C'C=u_1'X'Xu_1$$

On choisira u_1 de façon à maximiser cette dernière quantité. Le problème est donc:

$$\text{Maximiser} \quad u_1'X'Xu_1 \quad \text{sujet à } u_1'u_1=1$$

Il s'agit d'un problème classique d'optimisation sous contrainte que l'on peut solutionner par la méthode de Lagrange.

On forme le Lagrangien:

$$L = u_1'X'Xu_1 - \lambda(u_1'u_1-1)$$

On dérive par rapport à chacune des p composantes du vecteur u_1 ainsi que par rapport au multiplicateur de Lagrange (λ) et on pose les dérivées partielles égales à zéro.

$$2\left[X'Xu_1 - \lambda u_1\right] = 0$$
$$u_1' u_1 = 1$$

Simplifiant, on trouve:

$$X'X\, u_1 = \lambda u_1$$
$$u_1' u_1 = 1$$

On reconnaît là l'équation de vecteurs propres et de valeurs propres de la matrice X'X. On note que cette matrice est, à un facteur 1/n ou 1/(n-1) près, la matrice de variances (sur la diagonale) et de covariances (hors-diagonale) des p variables. Le vecteur donnant les projections ayant la plus grande dispersion est donc le 1er vecteur propre de la matrice de variances-covariances de X (c.f. remarque ii. ci-bas).

Remarques:

i. La matrice des produits scalaires (X'X) est, par construction, symétrique et au moins positive semi-définie. Ceci implique que les valeurs propres et les vecteurs propres seront réels. De plus les valeurs propres seront toutes positives ou nulles. Finalement, rappelons que les vecteurs propres d'une matrice symétrique sont toujours orthogonaux entre eux, i.e. $u_1'u_2=0$. Ce n'est généralement pas le cas lorsque la matrice n'est pas symétrique.

ii. Les projections sur le vecteur u_1 définissent une nouvelle variable qui est combinaison linéaire des variables originales. La variance de cette nouvelle variable est donnée par:

$$1/n\ (u_1'X'Xu_1) = 1/n\ (u_1'\lambda_1 u_1) = 1/n\ \lambda_1$$

La variance des projections est donc égale (au facteur 1/n) à la valeur propre. Le maximum est donc atteint avec la 1ère valeur propre. C'est pourquoi on retient le 1er vecteur propre.

iii. On sait que la somme des valeurs propres d'une matrice est égale à la trace de la matrice originale. La quantité totale de variation n'est donc pas modifiée. De plus, le rapport $\lambda_1/(\Sigma\lambda_i)$ indique la proportion de la variance totale prise en charge par le 1er vecteur propre.

Question 1: Quelle relation y a-t-il entre les vecteurs propres et les valeurs propres de X'X et ceux de X'X/n?

3.1.1 Recherche de sous-espaces optimaux

A la section précédente, on a établi que le vecteur pour lequel la variance des projections est maximale est le 1er vecteur propre de la matrice des variances-covariances. Qu'en est-il maintenant si on recherche le plan pour lequel la variabilité (l'étalement) des projections est maximale?

Théorème: Le plan expliquant le mieux le nuage de points (au sens des moindres carrés) contient nécessairement le vecteur u_1

Démonstration:

On suppose que u_1 n'est pas inclus dans le meilleur plan et on montre que l'on arrive alors à une contradiction.

i. Il existe au moins une droite (disons u_2) contenue dans le meilleur plan et qui est orthogonale à u_1.
Soit u_1^* la projection de u_1 dans le meilleur plan. On peut écrire $u_1 = u^*_1 + (u_1 - u^*_1)$
Soit u_2 la droite du meilleur plan orthogonale à u^*_1. Par construction, u_2 est aussi orthogonale à $(u_1 - u^*_1)$ donc u_2 est orthogonale à u_1.

ii. L'inertie (i.e. la somme des carrés des projections) sur tout plan est égale à la somme des inerties expliquées par deux droites orthogonales du plan (trivial par le théorème de Pythagore et la définition d'inertie).

iii. Le "meilleur" plan est défini par u_2 et u_1^*. Ce plan explique moins d'inertie que le plan défini par u_2 et u_1 car u_1 est la droite expliquant le plus d'inertie. On arrive donc à une contradiction et on conclut que le meilleur plan contient nécessairement u_1.

Le résultat est important puisqu'il nous permet de chercher le vecteur expliquant le maximum d'inertie puis le second vecteur, orthogonal au premier, expliquant le maximum d'inertie, etc.....

Recherche du 2e vecteur expliquant le maximum d'inertie.

On cherche u_2 tel que:

$$\begin{aligned} &u_2'X'Xu_2 \text{ soit maximum} \\ &\text{sujet à } \quad u_2'u_1 = 0 \\ &\text{et} \qquad\quad u_2'u_2 = 1 \end{aligned}$$

Le Lagrangien est:

$$L = u_2'X'Xu_2 - w(u_2'u_1) - \lambda_2(u_2'u_2 - 1)$$

Dérivant le Lagrangien par rapport à chacune des composantes du vecteur u_2 et par rapport à λ_2 et w_2 puis simplifiant, on trouve:

$$\begin{aligned} X'X\, u_2 &= \lambda_2 u_2 \\ u_2'u_2 &= 1 \\ u_1'u_2 &= 0 \end{aligned}$$

Le vecteur recherché est le vecteur propre associé à la 2e plus grande valeur propre de la matrice de variances-covariances. Ces résultats se généralisent aisément à plusieurs dimensions et on trouve que:

Théorème: L'espace à s dimensions $s \leq p$ donnant la meilleure explication (en termes d'inertie) du nuage original est défini par les s vecteurs propres de X'X associés aux s plus grandes valeurs propres.

3.1.2 Interprétation géométrique de l'ACP.

L'ACP revient à effectuer une rotation du sytème d'axes initial puisque les vecteurs propres sont orthogonaux entre eux et constituent donc un nouveau repère de coordonnées. Les cosinus entre les nouveaux axes et les anciens sont les composantes des vecteurs propres.

Une autre interprétation procède par analogie avec la régression. Le 1er vecteur propre est le vecteur qui explique le mieux, simultanément toutes les variables de la matrice X. Ce vecteur minimise la somme des carrés entre les projections sur le vecteur et la position des points dans l'espace original. Le deuxième vecteur propre est celui qui explique le mieux, simultanément, l'ensemble des résidus obtenus, et ainsi de suite...

3.1.3 Analyse dans l'espace des échantillons

On s'est intéressé, jusqu'à maintenant, au nuage des n observations dans l'espace des p variables. On pourrait également considérer le nuage de p variables dans l'espace des n observations. On cherche le sous-espace de dimension s ($s \leq p$) pour lequel la somme des carrés des projections est maximale.

On applique la même technique que précédemment et on trouve que la solution, pour le premier vecteur est donnée par le système suivant:

$$XX'v_1 = \beta_1 v_1$$
$$v_1'v_1 = 1$$

Le premier vecteur propre de XX' est celui qui maximise la variance des projections. Comme tantôt, le plan maximisant la variance des projections sera formé des deux premiers vecteurs propres, et ainsi de suite.

Théorème: Les valeurs propres λ_i et β_i sont identiques.

Prémultiplions par X':

$$X'XX'v_i = \beta_i X'v_i$$

Cette équation nous indique que $X'v_i$ est vecteur propre de X'X associé à la valeur propre β_i. Or, les valeurs propres de X'X sont données par λ_i. Donc, on conclut que $\beta_i=\lambda_i$

Remarque: Le théorème précédent nous indique qu'il n'est pas nécessaire de rechercher explicitement les valeurs propres et les vecteurs propres de XX'. Les valeurs propres sont les mêmes, et les vecteurs propres u_i sont donnés par $X'v_i$ à une constante de normalisation près. En effet, la norme de $X'v_i$ est:

$$v_i'XX'v_i = \lambda_i$$

donc

$$\frac{1}{\sqrt{\lambda_i}} X' v_i$$

est de norme 1. Par conséquent, on a nécessairement:

$$\frac{1}{\sqrt{\lambda_i}} X' v_i = u_i$$

et, de façon similaire

$$\frac{1}{\sqrt{\lambda_i}} X u_i = v_i$$

Ces formules sont appelées formules de transition.

Sous forme matricielle, on peut les écrire comme:

$$V = XU\Lambda^{-1/2} \qquad U = X'V\Lambda^{-1/2}$$

où U et V contiennent les vecteurs propres u_i et v_i placés en colonne.

3.1.4 Reconstruction complète et partielle de la matrice X

Soit la matrice V ayant les p vecteurs propres de XX' placés en colonne, et U ayant les p vecteurs propres de la matrice X'X placés en colonne. Soit la matrice Λ, une matrice diagonale pxp ayant les p valeurs propres sur la diagonale. Des formules de transition on a:

$$XU=V\Lambda^{1/2}$$

Postmultipliant cette expression par U' et notant que UU'=I, on obtient:

$$XUU'= X = V\Lambda^{1/2}U'$$

Cette dernière expression nous indique que l'on peut reconstruire la matrice X si on connaît les valeurs propres et les vecteurs propres de XX' et X'X. Cette décomposition de la matrice X est connue sous le nom de décomposition en valeurs singulières (Singular Value Decomposition (SVD) en anglais).

Remarque: Dans l'équation précédente, on peut sélectionner le nombre désiré de vecteurs propres de façon à limiter la reconstruction de X aux seuls vecteurs qui nous semblent vraiment signifiants. On peut ainsi filtrer de la matrice X toute information qui nous semble non-pertinente.

3.1.5 Coordonnées des observations et des variables sur les vecteurs propres (composantes principales)

Les coordonnées (en anglais "scores") des observations et des variables sont simplement les projections sur les vecteurs propres:

observations: $C_o= XU = V\Lambda^{1/2}$
variables: $C_v= X'V = U\Lambda^{1/2}$

3.1.6 Qualité de la représentation des observations et des variables

En comparant la projection d'une observation (au carré) avec la distance (au carré) par rapport à l'origine de cette observation, on obtient une mesure permettant de juger jusqu'à quel point une observation est près du vecteur, du plan,..., considéré.

observation i sur vecteur j: $Qlt_o(i,j) = (C_o(i,j)^2)/l_i^2$
variable i sur vecteur j: $Qlt_v(i,j) = (C_v(i,j)^2)/l_i^2$

Remarque: On peut, avec Matlab exécuter ce calcul sous forme matricielle (voir acp.m). l_i^2 se lit sur la diagonale de XX' pour les observations et sur la diagonale de X'X pour les variables.

Question 2: Si on effectue la somme des qualités de réprésentation d'une observation ou d'une variable sur les différents vecteurs propres, qu'obtiendra-t-on?

Remarque: La qualité de la représentation des variables est aussi égale au carré de la corrélation entre la variable et la composante principale. Dans le cas d'une acp de matrice des corrélations, la qualité de la représentation est le carré de la coordonnée (puisque $l_i=1$ pour toutes les variables); la coordonnée de la variable i sur le vecteur j est donc égale à la corrélation entre la variable i et le vecteur propre j (j^e composante principale).

3.1.7 Contributions des observations et des variables

Cette mesure vise à quantifier l'importance de chaque variable et de chaque observation dans la définition d'un vecteur propre. C'est une mesure de première importance quand vient le temps d'interpréter chaque vecteur propre.

observations: $Ctr_o = (C_o(.)^2)\ \Lambda^{-1}$
variables: $Ctr_v = (C_v(.)^2)\ \Lambda^{-1} = U(.)^2$

Remarque: La contribution est donnée par la coordonnée (au carré) d'une observation ou d'une variable sur un vecteur propre divisée par la valeur propre associée à ce vecteur propre. Pour les variables, il suffit de prendre les éléments, au carré, des vecteurs propres.

Question 3: Si on effectue la somme des contributions des observations ou des variables sur un vecteur propre, quel résultat obtiendra-t-on?

3.1.8 Règles d'interprétation des vecteurs propres

Etapes d'une interprétation.

i. Considérer l'analyse dans son ensemble. Répondre aux questions suivantes:
A-t-on pu réduire la dimension du problème?
Combien a-t-on de vecteurs importants?

Ces vecteurs expliquent quel pourcentage de l'inertie totale?

Remarque: -Sous hypothèse multinormale pour X, on peut effectuer des tests de signification sur les valeurs propres. Ceci n'est pas présenté ici car l'hypothèse multinormale est rarement vérifiée en pratique.

-Il arrive parfois que des vecteurs propres associés à des valeurs propres faibles soient plus intéressants en termes géologiques que des vecteurs associés à des valeurs propres plus fortes. Souvent, avec des analyses géochimiques de roches volcaniques par exemple, le premier facteur représentera les variations reliées à la différenciation magmatique. Les facteurs de variation vraiment intéressants apparaîtront alors sur d'autres vecteurs propres.

-La proportion d'inertie exigée avant de considérer un vecteur propre comme important décroît lorsque le nombre de variables et d'observations augmente. Pour plus de détails, voir Lebart et al. (1984, 162-190); Legendre et Legendre (1984, 123-124); Cooley et Lohnes (1971, 103-105).

-Interpréter, si possible, chaque vecteur retenu en se servant des contributions et des qualités de représentation. Répondre aux questions suivantes:

- Quelles sont les variables qui contribuent le plus à définir un vecteur?

- Quelles sont les observations qui contribuent le plus à définir un vecteur?

- Parmi les variables (ou observations) qui contribuent beaucoup, attacher plus d'importance à celles qui sont bien représentées.

- Parmi les variables (ou observations) qui ne contribuent pas (ou peu), considérer tout de même celles qui sont très bien représentées sur ce vecteur.

- Quelles sont les signes affectant les coordonnées des variables (observations); interpréter les vecteurs en tenant compte de ces associations et de ces oppositions.

- A-t-on de l'information extérieure (non incluse dans l'analyse) que l'on peut utiliser (ex. type de roche, altération connue, provenance, proximité d'un gisement,....) pour aider à l'interprétation? Si oui examiner la disposition de ces facteurs externes sur les différentes composantes principales.

- Y a-t-il des regroupements, tendances, évolutions au niveau des observations? Si oui, comment les expliquer?

Il n'y a pas de recette magique pour interpréter les résultats de l'analyse. Il s'agit d'un travail de synthèse qui demande habituellement de considérer simultanément plusieurs des pistes données précédemment.

Disons simplement qu'habituellement, l'interprétation des facteurs se fait à partir des variables, en considérant les variables qui contribuent fortement ou qui sont très bien représentées et en tenant compte évidemment de leur position (coordonnées) sur les différents vecteurs propres (ou sur les différents plans factoriaux).

Les observations servent surtout soit à corroborer l'interprétation faite sur les variables, soit à donner des pistes d'interprétation en termes des variables, soit à relier les variables à des informations extérieures non incluses dans l'analyse.

On pourra déceler des valeurs "bizarres" par leur contribution (pour les observations) à la définition des différents vecteurs. Une trop forte contribution peut indiquer par exemple une erreur de mesure ou un mauvais échantillonnage, etc... En effet une très forte contribution d'une observation n'est possible que si l'observation s'écarte considérablement du nuage de points. Si on décèle de tels problèmes, on doit vérifier et, s'il y a lieu, reprendre ou éliminer l'analyse.

Une technique qui peut être utile à l'interprétation consiste à choisir des observations qui sont très bien représentées à différents emplacements sur le vecteur propre, puis de cheminer d'une extrémité à l'autre du vecteur propre en notant ce qui change au niveau des observations retenues.

Mis à part les détails techniques présentés en justification, une interprétation ne dépasse que rarement une page de texte. C'est une formidable synthèse si l'on considère que le tableau initial de données comprend parfois plusieurs milliers de valeurs.

3.2 Exemple numérique

Voici un petit exemle numérique tiré de Davis (1973, 479-483). Les données sont :

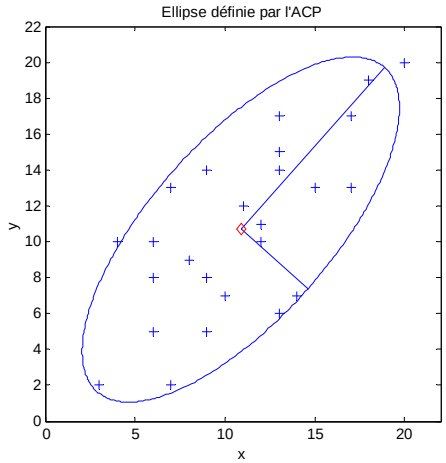
x=	3	4	6	6	6	7	7	8	9	9	9	10	11	12	12	13	13	13	13	14	15	17	17	18	20
y=	2	10	5	8	10	2	13	9	5	8	14	7	12	10	11	6	14	15	17	7	13	13	17	19	20

On effectue l'ACP de la matrice des covariances : $\Sigma = \begin{bmatrix} 20.28 & 15.59 \\ 15.59 & 24.06 \end{bmatrix}$

Les deux valeurs propres sont égales à 37.9 et 6.5. Les vecteurs propres sont :

	u_1	u_2
x	0.66314	0.7485
y	0.7485	-0.66314

Les vecteurs $\lambda_1 u_1$ et $\lambda_2 u_2$, définissent les axes majeurs d'une ellipse passant par les points (Var(x),Cov(x,y)) et (Cov(x,y),Var(y)). Les vecteurs $2\sqrt{\lambda_1}\ u_1$ et $2\sqrt{\lambda_2}\ u_2$ sont les axes majeurs d'une ellipse qui tracée dans le plan des coordonnées des observations devraient renfermer approximativement 86% des observations lorsque la distribution est (bi) normale.



Question 4: Quelle est la proportion de la variance expliquée par chaque vecteur propre? La somme des valeurs propres est égale à quelle quantité? Les vecteurs propres ont-ils été calculés avec X centré ou non?

Exercice 1: Vérifiez l'orthogonalité des vecteurs propres.

Question 5: Vous voulez utiliser le premier vecteur propre de l'ACP comme une équation de prédiction. Vous observez (nouvelle observation) $X_2=19$. Quel estimé donnerez-vous pour X_1?

3.3 Observations et variables supplémentaires

Il peut arriver que l'on veuille représenter sur les différents plans factoriels (i.e. plans définis par deux vecteurs propres) des observations ou des variables qui n'étaient pas incluses dans l'analyse initiale. Par exemple, on pourrait disposer d'observations de contrôle ou de nouvelles observations pourraient s'ajouter dans des contextes légèrement différents ou provenant de localisations différentes,... Les variables supplémentaires pourraient être des variables de nature très différentes des variables originales. On pourrait effectuer l'ACP à partir d'analyses géochimiques d'éléments majeurs et vouloir projeter sur les plans factoriels d'autres variables telles les éléments mineurs, des variables indicatrices du type ou de la texture de la roche, une variable donnant la distance du gisement le plus près,...

Observations:

On transforme chaque observation supplémentaire de la même façon que les données originales (s'il y a lieu). Pour cela on utilise les moyennes et écarts-types calculés avec les données originales seulement. On calcule X^+U où X^+ désigne la matice mxp contenant les m observations supplémentaires.

Variables:

On effectue la même transformation sur la variable supplémentaire que celle effectuée sur les variables originales (s'il y a lieu). On calcule X_+V où X_+ désigne la matrice kxn des k nouvelles variables à projeter. En utilisant les formules de transition, on a que les coordonnées des variables supplémentaires sont:

$$X_+XU\Lambda^{-1/2}$$

Remarque: La notion de projections supplémentaires peut être utilisée pour étudier la stabilité des plans factoriels obtenus. Pour ce faire, on effectue l'ACP et on détermine U. On échantillonne, avec remise, n rangées de notre matrice initiale X. on obtient ainsi une nouvelle matrice de données X_2 (de même taille). On calcule la matrice de variances-covariances de cette nouvelle matrice et on projette chaque ligne de la matrice de variances-covariances sur les vecteurs $U\Lambda^{-1/2}$. En effet, la projection d'une variable est donnée par $XV=X'XU\Lambda^{-1/2}$. Or X'X est, au facteur 1/n près la matrice de variances-covariances des variables. On répète l'échantillonnage un grand nombre de fois et on observe la dispersion des coordonnées des variables sur les différents vecteurs propres. Cette procédure est connue sous le nom de "**bootstrapping**". Pour plus de détails, voir Greenacre (1984) et Lebart et al. (1984).

3.4 ACP de la matrice des covariances

Dans cette forme d'ACP, on cherche les vecteurs propres et valeurs propres de la matrice X'X/n au lieu de X'X (X est centrée). Habituellement, on projette ensuite X sur les vecteurs propres. Les valeurs propres de cette matrice représentent des variances plutôt que des sommes de carré. Toutefois rien d'important ne change par rapport à l'analyse de X'X. Les vecteurs propres demeurent les mêmes, les projections également.

3.5 ACP de la matrice des corrélations

Si, en plus de centrer la matrice X, on avait divisé chaque valeur par l'écart-type de la variable correspondante, alors le produit X'X/n serait la matrice des corrélations. Il s'agit de l'ACP la plus courante. L'avantage de cette approche est que les variables ne possédant plus d'unités, les résultats sont indépendants des unités originales choisies pour mesurer les variables. On peut ainsi considérer des variables dont l'ordre de grandeur des variations est très différent. Auparavant, une variable présentant une variation très grande aurait contribué beaucoup plus à la définition des vecteurs propres qu'une variable montrant peu de variations. Ainsi, le TiO_2 exprimer en pourcentage, aura une importance beaucoup moins grande dans l'ACP de la matrice des variances-covariances que s'il est exprimé en ppm. Ce n'est pas le cas avec l'ACP de la matrice des corrélations et c'est ce qui fait que cette méthode est habituellement préférée.

Particularités de l'ACP de la matrice des corrélations

i. La trace de la matrice des corrélations (donc la somme des valeurs propres) est égale au nombre de variables. Comme pour l'ACP de la matrice des covariances, les valeurs propres représentent la variance des projections sur chaque vecteur. Dans les calculs des contributions des observations, il faudra en tenir compte (cf. ACP matrice des covariances).

ii. Les coordonnées des variables sur les vecteurs propres sont égales aux corrélations entre ces variables et les composantes principales correspondantes. Par conséquent les coordonnées des variables sont toutes inférieures à 1 et supérieures à -1. Également, si on trace un cercle unité sur un plan factoriel quelconque, on est assuré que les projections des variables tomberont toutes à l'intérieur de ce cercle.

En effet, les coordonnées des variables sont données par:

$$X'V=U\Lambda^{1/2}$$

La variance d'une composante principale est donnée par λ_i (note: ici λ_i est valeur propre de la matrice des corrélations).
La variance d'une variable originale est 1.
Les covariances entre composantes principales et variables originales sont données par:

$$1/n\ X'C= 1/n\ X'XU = U\Lambda$$

La corrélation est donc : $U\Lambda\Lambda^{-1/2} = U\Lambda^{1/2}$, i.e. la corrélation est égale à la coordonnée des variables sur les vecteurs propres.

Finalement, puisque la qualité de la représentation est ici égale à la coordonnée, au carré, de la variable et que la somme des qualités de représentation donne nécessairement 1, il suit que les coordonnées des variables sont sur la surface d'une hypersphère de rayon unité. Plusieurs programmes impriment automatiquement le cercle unité sur les divers plans factoriels.

Question 6: Deux variables très corrélées apparaîtront comment sur les différents plans factoriels? Comment peut-on évaluer visuellement si une variable est bien représentée sur un plan factoriel donné?

Question 7: Lorsqu'on a que deux variables, quelle forme très particulière prennent les vecteurs propres?

3.5.1 Cas particulier de l'ACP avec 2 variables

Lorsqu'il n'y a que deux variables, des relations très simples existent pour les valeurs propres et vecteurs propres de l'ACP de la matrice des corrélations. Les composantes des vecteurs propres sont toutes égales à $\pm (1/2)^{0.5}$. Les valeurs propres sont égales à $1 \pm r_{xy}$. Dans une figure mettant en relation X/s_x et Y/s_y, la pente du 1er vecteur propre est donc 1 ou -1 selon que la corrélation entre x et y est positive ou négative. Si on construit la figure Y vs X, alors on aura une droite dont la pente est s_y/s_x que l'on peut utiliser pour effectuer des prédictions lorsque X et Y ont des rôles similaires (ex. longueur et largeur d'un fossile). Cette droite de prédiction est connue sous le nom de "reduced major axis". On peut démontrer que c'est la droite qui minimise la somme des produits croisés des erreurs sur X et des erreurs sur Y.

$$\text{Min}\left[\sum_{i=1}^{n}\left|(X_i - \hat{X}_i)(Y_i - \hat{Y}_i)\right|\right]$$

avec les valeurs prédites données par le modèle unique:

$$\hat{Y} = b_0 + b_1 X$$

$$\hat{X} = \frac{Y - b_0}{b_1}$$

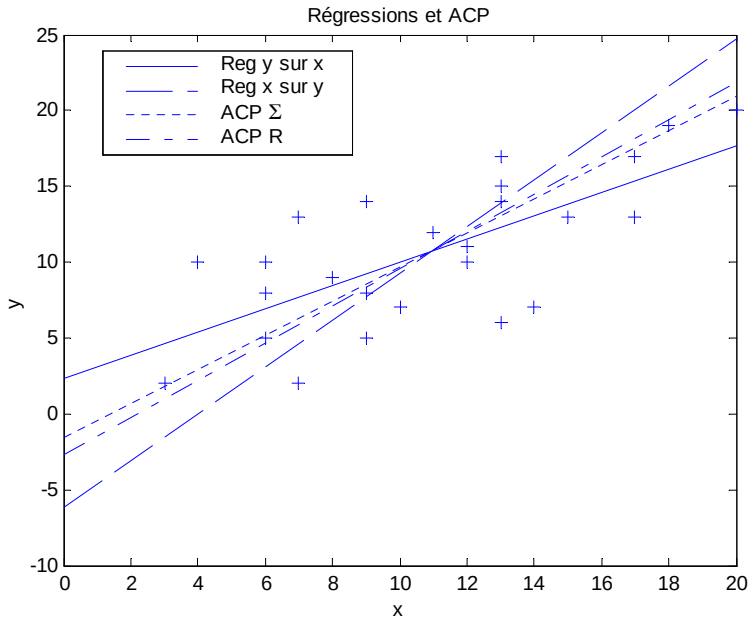
Géométriquement, ceci revient à minimiser la somme des surfaces des triangles rectangles formés par la droite et les points expérimentaux.

Pour l'ACP de la matrice des covariances, on parle d'axe majeur, mais il n'y a pas les mêmes relations simples pour déterminer les valeurs propres et vecteurs propres.

On peut aussi généraliser la notion d'axe majeur ou d'axe majeur réduit à plus de deux dimensions. On obtient ainsi, en général, un hyperplan de dimension p-1 orienté dans un espace à p dimensions et qui permet de prédire n'importe quelle variable étant donné que l'on connaît les p-1 autres variables. Cette équation s'obtient à partir du vecteur propre associé à la **plus petite** valeur propre. En effet, la projection sur le plus petit vecteur propre est Xu_i. Si cette projection est 0, alors l'observation se trouve dans l'espace complémentaire (i.e. hyperplan de dimension p-1). Donc, on doit avoir, dans le cas de l'ACP de la matrice des corrélations:

$$0 = \frac{(x_1 - \bar{x}_1)}{s_1} u_{p1} + \frac{(x_2 - \bar{x}_2)}{s_2} u_{p2} + ... + \frac{(x_p - \bar{x}_p)}{s_p} u_{pp}$$

qui permet d'exprimer une variable en fonction des p-1 autres.



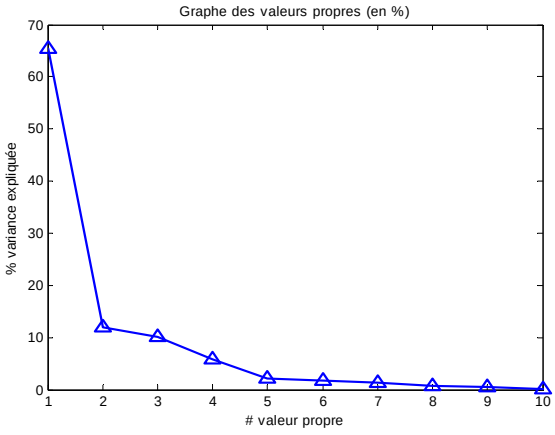
3.6 Exemples d'application

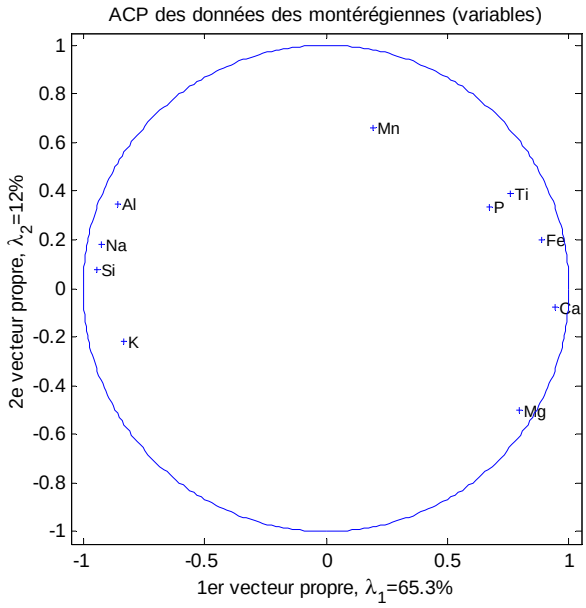
3.6.1 ACP des données des montérégiennes

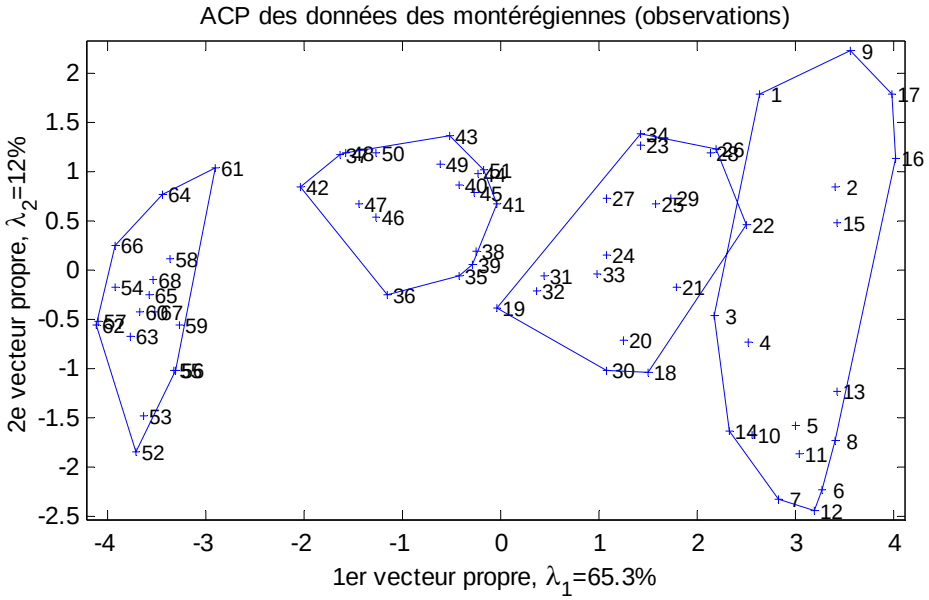
Les données sont des analyses géochimiques d'éléments majeurs (10 éléments) de roches prélevées dans les différentes montérégiennes de la région. On a 68 analyses de roches réparties en 4 groupes principaux de 17 observations et plusieurs sous-groupes.

On remarque que les deux premiers vecteurs propres expliquent à eux seuls 77% de la variance totale. Le premier vecteur propre oppose SiO_2, Al_2O_3, Na_2O et K_2O à FeO, MgO, TiO_2, CaO et P_2O_5. Les divers éléments contribuent à peu près également à la définition du vecteur (sauf MnO). Ce premier vecteur oppose donc les éléments felsiques aux éléments mafiques. A ce titre, on peut l'interpréter comme un axe de maturité magmatique. La disposition des observations le long de cet axe confirme d'ailleurs cette interprétation puisqu'on retrouve les roches mafiques à droite sur le graphe, les felsiques à gauche, et les intermédiaires au centre.

Groupe	Sous-groupe	Observations #
Ultramafiques	Jacupirangite	1, 2
	Alnoïte	3 à 13
	Mélilite à biotite	14
	Yamaskite	15 à 17
Mafiques	Essexite	18 à 21
	Montréalite	22
	Gabbro	23 à 32
	Diabase	33, 34
Roches intermédiaires	Dykes mésocratiques	35 à 51
Roches felsiques	Syénites	52 à 68







Le 2e vecteur propre oppose TiO_2 et MnO à MgO. Sa signification est moins claire, toutefois, on note que la dispersion au niveau des observations s'effectue surtout au niveau des roches mafiques, les autres roches présentant généralement des coordonnées plus homogènes sur ce facteur. Il semble opposer les Jacupirangites (1-2) et Yamaskites (15 à 17) aux Alnoïtes (3 à 13). On notera également que l'observation 9 des Alnoïtes se démarque nettement des autres Alnoïtes. Ceci suggère soit une erreur d'analyse, soit une erreur d'identification de cette roche, soit une particularité qui la distingue des autres Alnoïtes (ex. altération, minéralogie particulière, etc.). Si on compare les valeurs de l'analyse à celles des autres Alnoïtes, on note une teneur en MnO, P_2O_5 et CaO plus élevée et en MgO plus faible. L'augmentation des trois premiers éléments suggère la présence importante d'apatite dans cette roche.

3.6.2 Utilisation en télédétection

Une application très courante de l'ACP consiste à produire une carte synthèse à partir d'une information multivariable. Ce genre de situation se présente en télédétection et analyse d'images ou l'on dispose de l'intensité d'un signal mesurée pour différentes plages de longueurs d'ondes. Les systèmes sophistiqués de SIG (système d'information géographique) permettent habituellement d'effectuer l'ACP du signal mesuré et de produire des images composites correspondant aux diverses composantes principales. Ceci permet, entre autres, de filtrer du bruit et de rehausser l'information contenue dans l'image originale.

3.6.3 Détermination de l'orientation d'une veine à partir de données de forages

Souvent on a des intersections d'une veine obtenues dans quelques forages et l'on désire déterminer l'orientation de la veine. On cherche donc à déterminer 1- si les données sont réellement dans un plan, 2- si oui, quelle est l'orientation de ce plan. L'ACP fournit une méthode très simple pour solutionner ce

problème. On effectue l'ACP avec, comme variables, les coordonnées x,y, et z des intersections (habituellement le centre de l'intersection). Les deux premiers vecteur propres vont s'orienter dans le plan de la veine et fourniront le meilleur plan d'ajustement possible (les distances au carré mesurées perpendiculairement au plan sont minimisées). Le 3e vecteur propre (vecteur unitaire) sera la normale au plan et peut être utilisé pour calculer la direction et le pendage du plan.

ex. La compagnie Cambiex-Miramar a obtenu 112 intersections d'une veine d'or dans un dépôt des TNO. La variable x est orientée vers l'est, y vers le nord et z vers le haut. L'ACP de la matrice des covariances de ces données a indiqué que 99.7% de la variation était exprimée par les 2 premiers vecteurs-propres. On est donc effectivement sur un plan. Les composantes du 3e v.p. sont u_3=[0.9937 -0.1116 .0104]. On voit donc immédiatement que ce vecteur est essentiellement parallèle à l'axe des x et donc que la veine est subverticale et orientée suivant la direction y (nord).

Plus précisément, la direction du pôle est donnée par:

$\text{atan}(u_x/u_y)=\text{atan}(-0.9937/0.1116)=-83.59^0$

L'inclinaison du pôle est: $\text{asin}(0.0104)=0.59^o$

Passant du pôle au plan et respectant la convention de la main droite, on trouve que la direction du plan est: $90-83.59=6.4^o$ et le pendage: $90-0.59=89.41^o$.

3.6.4 Détermination des contraintes principales et de leur orientation en un point d'un massif rocheux

Cette application n'est pas à proprement parler de l'ACP, toutefois, elle utilise les mêmes outils. En effet, l'ACP consiste, en définitive, à extraire les valeurs propres et les vecteurs propres d'une matrice symétrique. Les vecteurs de cette matrice symétrique sont situés sur un hyper-ellipsoïde (ellipse si p=2, ellipsoïde si p=3); Les vecteurs propres donnent les directions des axes principaux et les valeurs propres les longueurs de ces axes.

En mécanique des roches, lorsqu'on mesure un tenseur de déformations, on calcule par inversion un tenseur de contraintes, lequel est symétrique (toutefois, il n'est pas nécessairement positif défini ou positif semi-défini). On y retrouve les contraintes normales sur la diagonale et les contraintes tangentielles hors diagonale. Les vecteurs propres de cette matrice donnent les directions des forces principales et les valeurs propres en donnent l'amplitude. Il est possible d'observer des valeurs propres négatives, ce qui correspond alors à des forces de tension plutôt que la situation plus courante de forces de compression.

Exemple: On a calculé les contraintes selon 3 axes orthogonaux (x,y et z) dans un massif rocheux et on a obtenu la matrice des contraintes suivantes:

$$\begin{bmatrix} 1200 & 150 & 200 \\ 150 & 800 & 320 \\ 200 & 320 & 1800 \end{bmatrix}$$

Les valeurs sont en KPa. Calculant les valeurs propres et vecteurs propres de cette matrice, on trouve:

$$\begin{bmatrix} u_1 & u_2 & u_3 \\ .29 & -.94 & .19 \\ .29 & -.1 & .95 \\ .91 & .33 & .24 \end{bmatrix}$$

Les valeurs propres valent λ_1=1966, λ_2=1145 et λ_3=690. Ces valeurs propres sont les contraintes principales et la direction de ces contraintes est donnée par les vecteurs propres.
Ainsi, l'angle de la contrainte σ_1 avec l'axe z est $\cos^{-1}(0.91)=24.5^o$
l'angle de la contrainte σ_3 avec l'axe y est $\cos^{-1}(0.95)=18^o$.

3.6.5 Étude des joints et fractures

(Consultez aussi Fisher, Lewis et Embleton, 1987, Statistical analysis of spherical data; . G.S. Watson, 1983, Statistics on spheres)

Soit un système de coordonnée main droite (i.e. main droite tendue, paume vers le haut, le pouce pointe vers x, l'index vers y et le majeur vers z). On peut représenter une fracture par sa normale sous la forme d'un vecteur unitaire dans ce repère. Il existe plusieurs conventions possibles pour orienter une fracture (ex. direction-pendage; pendage et direction du pendage; etc.)

Ainsi, une fracture dont le vecteur pendage est orienté à 35^o et montre une inclinaison de 63^o (pendage positif vers le bas) présente un pôle dont le pendage est 63-90= -27^o orienté à 35^o. Les cosinus directeurs de ce vecteur sont:

x=cos(27)*sin(35)=0.5111
y=cos(27)*cos(35)=0.7299
z=sin(27)=0.4540

On a bien: $x^2+y^2+z^2=1$, i.e. le vecteur est unitaire. Note, on aurait pu tout aussi bien utiliser comme cosinus directeurs
-(x,y,z), i.e. les vecteurs ne sont pas orientés.

Si on considère "n" fractures, on peut former une matrice X_{nx3} de tels vecteurs. On définit la *matrice d'orientation* comme X'X (sans centrage). L'idée est d'effectuer l'ACP directement sur cette matrice. On cherche donc les valeurs propres et vecteurs propres de 1/n X'X. les résultats de l'ACP s'interprètent de la façon suivante:

#	*Apparence sur le stéréonet*	*Valeurs propres*	*Interprétation des vecteur propres*
1	Points regroupés (mode unique)	$\lambda_1 >> \lambda_2, \lambda_3$	Le 1er vecteur propre donne l'orientation moyenne des fractures.
2	Points sur un grand cercle	$\lambda_1 \cong \lambda_2 > \lambda_3$	Les fractures s'orientent autour d'un cylindre ex. pli cylindrique). Le 3e vecteur propre donne l'axe du cylindre.
3	Points sur une portion de grand cercle	$\lambda_1 > \lambda_2 > \lambda_3$	Même interprétation sauf que l'on a observé qu'une partie des positions possibles des fractures.
4	Points sur un petit cercle	$\lambda_1 > \lambda_2 \cong \lambda_3$	Les fractures sont sur un cône de grande ouverture (s'approche du cas 1). Le 1er v.p. donne l'orientation

			de l'axe du cône.
5	Points sur un petit cercle	$\lambda_1 \cong \lambda_2 > \lambda_3$	Les fractures sont sur un cône de petite ouverture (s'approche du cas 2). Le 3e v.p. donne l'orientation de l'axe du cône.
6	Points répartis uniformément	$\lambda_1 \cong \lambda_2 \cong \lambda_3$	Pas d'orientation préférentielle (distrib. aléatoire)
7	2 ou plusieurs groupes de points	$\lambda_1 > \lambda_2 > \lambda_3$	Plusieurs familles? Si oui, isoler les familles et reprendre l'analyse séparément pour chaque famille. Pas d'interprétation particulière des vecteurs propres et valeurs propres.

Note: Comme Trace(X'X)=Trace(XX') et que la diagonale de XX' vaut 1 partout (les vecteurs sont unitaires), il découle que les valeurs propres de 1/n X'X somment à 1.

Note: Lorsqu'il n'y a qu'une seule famille de joints, le 1er vecteur propre est un excellent estimateur du pôle moyen de la famille. Il possède l'immense avantage de traiter correctement, et de façon automatique, les pôles indépendamment de l'orientation qu'on leur donne. Ex. en 2D: Si on a un pôle à 0^o et un autre à 180^o, le pôle moyen est à 0^o (ou 180^o) et non à 90^o. L'ACP fournira ce résultat alors que les autres estimateurs nécessitent d'abord de corriger l'orientation du 2e joint de 180^o à 0^o avant de calculer le pôle moyen. Ces corrections ne sont pas toujours évidentes à faire lorsqu'on a une famille montrant une grande dispersion et c'est encore plus difficile si plusieurs familles se retrouvent sur le stéréonet.

Note: Lorsqu'il y a plus d'une famille de joints, on devrait au préalable séparer les familles avant de les soumettre à l'ACP. On peut aussi utiliser le principe de maximum de vraisemblance si on a une idée à priori du type de distribution des joints pour chaque famille.

Note: On peut tester l'uniformité de la distribution des joints (i.e. aucune orientation préférentielle) dans un espace à p dimensions (réf. G.S. Watson, 1983, Statistics on spheres, Wiley, p. 59).

$$n(p+2)\frac{p}{2}\sum_{i=1}^{p}\left(\lambda_i - \frac{1}{p}\right)^2 \approx \chi^2_{\left[\frac{p(p+1)}{2}-1\right]}$$

où n est le nombre d'observations et p la dimension (habituellement 2 ou 3).

Exemple: On a observé les 15 joints suivants:

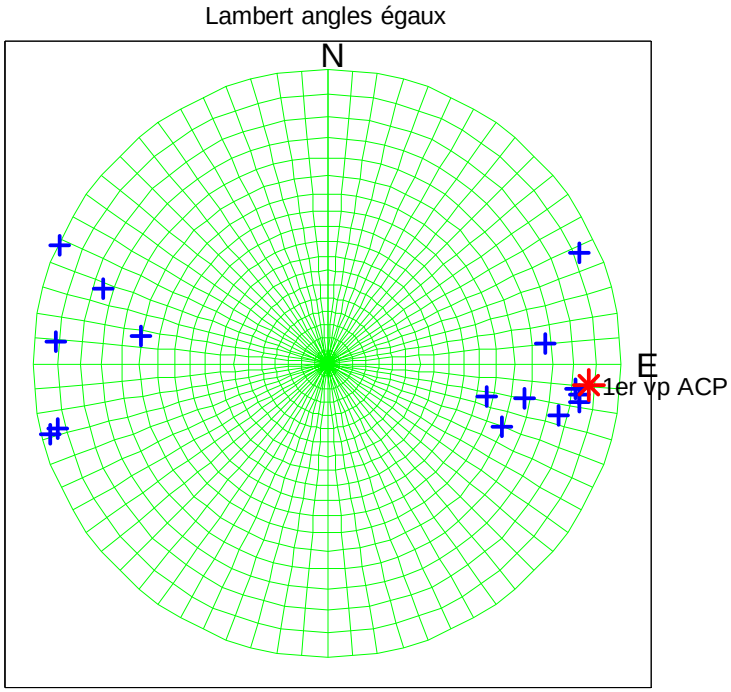
Pôle		Vecteur pôle (unitaire)		
Pendage	Direction	x	y	z
9.40	95.91	0.9813	-0.1016	-0.1633
8.12	98.53	0.9790	-0.1468	-0.1413
12.28	102.84	0.9527	-0.2171	-0.2127
32.10	101.66	0.8297	-0.1712	-0.5314
21.34	100.01	0.9172	-0.1619	-0.3640
25.49	109.66	0.8500	-0.3037	-0.4304
4.34	274.45	-0.9941	0.0774	-0.0757
3.33	256.40	-0.9703	-0.2347	-0.0581
24.66	278.28	-0.8994	0.1310	-0.4172
12.46	288.40	-0.9265	0.3082	-0.2158
0.49	293.80	-0.9149	0.4035	-0.0086

8.18	96.94	0.9826	-0.1195	-0.1423
16.68	84.58	0.9536	0.0905	-0.2870
3.88	66.42	0.9144	0.3991	-0.0676
1.77	255.66	-0.9684	-0.2475	-0.0308

Les 3 dernières colonnes contiennent les coordonnées des vecteurs pendages unitaires. On soumets les 3 dernières colonnes à une ACP (i.e. on extrait de la matrice d'orientation les vecteurs propres et valeurs propres). On obtient:

λ_1=0.89, λ_2=0.06 et λ_3=0.04. Clairement, on se retrouve dans le cas 1, i.e. des données regroupées autour d'un mode. Le 1er v.p. est: u_1=(0.99, -0.08, -0.11). Ce vecteur indique que le pôle moyen des fractures est essentiellement parallèle à x. Sa direction est: 94.9^o, son pendage est 6.4^o.

Le test précédent donne: 15*5*1.5*[(0.89-0.33)^2+(0.06-0.33)^2+(0.06-0.33)^2=51.7 alors qu'une chi^2 avec 3*4/2-1=5 degrés de liberté montre une valeur critique de 11.07. Clairement, on rejette fortement l'hypothèse d'uniformité et le regroupement de points peut donc être considéré comme significatif. Les fractures sont sub-verticales d'orientation nord-sud..



3.6.6 Filtrage de données géochimiques

Une application courante de l'ACP consiste à éliminer des données des facteurs de variation indésirables. Ceci peut être fait facilement en effectuant une reconstruction partielle de la matrice X en ne conservant que les vecteurs propres d'intérêt.

ex. 1. Dans le cas d'une campagne géochimique régionale (ex. sédiments de lac), des facteurs locaux et régionaux viennent influer sur la composition chimique des sédiments. Les facteurs régionaux peuvent être la géologie, la végétation, le relief, la température, etc. Les facteurs locaux peuvent être, la disponibilité de

matière organique, la position du lac dans le réseau hydrographique, la profondeur du lac, la nature des sédiments, la minéralisation, etc.

Une bonne façon d'identifier le caractère régional ou local d'un facteur de variation est de calculer le variogramme des coordonnées sur chaque vecteur propre. Les vecteurs montrant des variogrammes indicatifs d'une structure spatiale à grande échelle (grande portée) identifient les facteurs sûrement régionaux. Les variogrammes présentant une absence de structure de grande échelle identifient les facteurs locaux.

ex. 2. PFE de M. Guèvremont (1999)

Des diagrammes de classification sont utilisés pour reconnaître le type de roche à partir de l'analyse géochimique. Ces diagrammes nécessitent en général l'utilisation de roches fraîches et non altérées. M. Guèvremont a projeté des analyses prises à proximité d'un dépôt minéralisé. La plupart des observations tombaient carrément à l'extérieur des diagrammes de De la Roche(1980) servant à classer les roches plutoniques. De plus les observations tombant sur le diagramme fournissait des noms de roches différents de ceux attendus suite aux observations pétrographiques.

M. Guèvremont a effectué une ACP (matrice des corrélations) du tableau d'analyses géochimiques. Il en est ressorti un premier vecteur propre représentant 58% de la variation et dont l'interprétation était clairement la maturité magmatique de la roche. On a reconstruit les analyses en ne conservant que ce facteur et on a projeté les analyses reconstruites (ou filtrées) sur le diagramme de De la Roche. Cette fois, presque toutes les observations tombent dans le champ de la granodiorite, le type de roche suggéré par l'examen pétrographique et celui rencontré le plus souvent à proximité de la zone étudiée.

3.6.7 Analyse de données TBF (Marroquin, 1997)

Considérons une plaque mince conductrice au-dessus duquel un relevé TBF est réalisé.
La réponse TBF dépend des facteurs suivants :

- fréquence du signal émis par l'antenne
- conductivité du conducteur et du milieu encaissant,
- épaisseur du conducteur,
- profondeur du conducteur,
- pendage du conducteur.

Le long d'un profil donné traversant un conducteur, on note la variation du champ magnétique vertical. Cette variation est représentée par l'inclinaison de l'ellipse de polarisation par rapport à l'horizontal (phase). C'est habituellement la valeur étudiée bien que la quadrature (rapport entre les axes mineurs et majeurs de l'ellipse de polarisation) soit aussi parfois utilisée.

Marroquin (1997) a généré synthétiquement des réponses TBF de conducteurs dont les divers paramètres (conductivité, épaisseur, profondeur et pendage) variaient. Chaque conducteur était échantillonné aux mêmes 41 points le long du profil. Ces 41 points constituent les observations alors que chaque conducteur définit une colonne dans la matrice X. En soumettant cette matrice X à une ACP (ACP de la matrice des corrélations), il a été possible d'identifier des vecteurs propres pouvant être reliés à la profondeur, la conductivité, etc. Un premier vecteur propre représentait la forme « commune » des différents conducteurs et comptait pour 95% de la variation totale. Les autres vecteurs propres représentaient de légères variations par rapport à cette forme commune et dues aux différences existant entre les divers conducteurs étudiés.

3.7 Tableau récapitulatif

Le tableau suivant présente les diverses relations pour le cas générique d'une matrice X. On cherche alors les vecteurs propres de la matrice X'X.

Item	Observations	Variables	Remarques
Valeurs propres	Λ	Λ	$\lambda_i / \sum \lambda_j = \text{trace}(\Lambda)$:proportion expliquée par le v.p. "i"
Vecteurs propres	U	V	U'U=UU'=I U est pxp V'V=I V est nxp
Équation	$X'XU=U\Lambda$	$XX'V=V\Lambda$	Variances-covariances (au facteur 1/n près)
Relations de transition	$U=X'V\Lambda^{-1/2}$	$V=XU\Lambda^{-1/2}$	
Coordonnées	$C_o=XU$ $=V\Lambda^{1/2}$	$C_v=X'V$ $=U\Lambda^{1/2}$	Les coordonnées sont les projections des observations (ou des variables) sur les vecteurs propres
Qualités (observation ou variable i sur vecteur j)	$C_o(i,j)^2/l_i^2$	$C_v(i,j)^2/l_i^2$	Carré des coordonnées/carré de la longueur (l_i^2); pour la variable i, c'est aussi le carré de la corrélation entre la variable i et les coordonnées des observations sur le vecteur j.
Contributions	$C_o(.)^2\Lambda^{-1}$	$C_v(.)^2\Lambda^{-1}$ $U(.)^2$	Notation pointée: l'opération s'applique à chaque élément de la matrice pris séparément.
Reconstruction de X	$X=V\Lambda^{1/2}U'$ $=C_oU'$		Reconstruction partielle obtenue en utilisant les colonnes désirées de V et X ainsi que les éléments diagonaux correspondants de Λ

Note : Dans le cas de l'acp de la matrice des corrélations, les coordonnées des variables sont égales aux corrélations entre les variables et les différents vecteurs propres (i.e. corrélation entre les variables originales et les coordonnées des observations sur les différents vecteurs propres).

Note : Si les vecteurs propres sont issus d'une matrice de covariance ou de corrélation, alors le tableau précédent s'applique intégralement en posant :

Matrice des covariances	Matrice des corrélations
$X = \frac{(Y - 1\,1'Y/n)}{\sqrt{n}}$	$X = \frac{(Y - 1\,1'Y/n)}{\sqrt{n}} D_\sigma^{-1}$

où Y est la matrice des données originales, 1 est un vecteur nx1 avec des 1 1'/n effectue le centrage de Y. D_{σ}^{-1} est une matrice diagonale avec $1/\sigma$ (cette multiplication normalise l'écart-type de X à la valeur 1).

Note importante : Plusieurs programmes effectuant l'ACP de la matrice des covariances ou des corrélations préfèrent utiliser comme coordonnées des observations $\sqrt{n}\,C$ plutôt que C pour rendre les coordonnées indépendantes du nombre d'observations. Il faut donc diviser les coordonnées fournies par ces programmes par le facteur $\sqrt{n}$ avant d'appliquer les formules du tableau.
Si l'on prend directement les coordonnées fournies par ces programmes, on aura alors :

Item	matrice des covariances	matrice des corrélations
Qualité (observation i, vecteur j)	$C_o(i,j)^2/l_i^2$	$C_o(i,j)^2/l_i^2$
Qualité (variable i, vecteur j)	$C_v(i,j)^2/\sigma_i^2$	$C_v(i,j)^2$
Contribution (observation i, vecteur j)	$(C_o(.)^2/n)\,\Lambda^{-1}$	$(C_o(.)^2/n)\,\Lambda^{-1}$
Contribution (variable i,, vecteur j)	$U(i,j)^2$	$U(i,j)^2$

3.8 Exemple numérique complet

ACP de la matrice des covariances.

Matrice des données Y

21	38	52
4	51	67
67	83	0
67	3	38
93	5	6

Matrice centrée X

-29.4	2	19.4
-46.4	15	34.4
16.6	47	-32.6
16.6	-33	5.4
42.6	-31	-26.6

Matrice des variances et des covariances S=(X'X/n)

1076.6	-368.6	-750.24
-368.6	897.6	-66.2
-750.24	-66.2	671.84

Valeurs propres

λ_1= 1732.20
λ_2= 906.65
λ_3= 7.28

Vecteurs propres U

	u_1	u_2	u_3
x_1	0.79	-0.06	0.62

x_2	-0.30	-0.90	0.31
x_3	-0.54	0.43	0.73

Coordonnées des observations (C_o=XU)

	u_1	u_2	u_3
obs 1	-34.15	8.16	-3.40
obs 2	-59.54	3.84	1.03
obs 3	16.25	-57.29	1.07
obs 4	20.21	31.14	3.94
obs 5	57.24	14.14	-2.65

ex. observation 4 sur le 2e vecteur:
16.6*(-.06) - 33*(-0.90) + 5.4*.43 = 31.14 (aux arrondis près)

Coordonnées des variables $U\Lambda^{1/2}$

	u_1	u_2	u_3
x_1	-32.73	-1.66	1.66
x_2	12.68	-27.13	0.83
x_3	22.37	12.95	1.96

ex. variable 2 sur le 1er vecteur:
$-.30*(1732.2)^{1/2} = -12.68$ (aux arrondis près)

Qualité de représentation des observations

	u_1	u_2	u_3
obs 1	0.94	0.05	0.01
obs 2	1.00	0.00	0.00
obs 3	0.07	0.93	0.00
obs 4	0.29	0.70	0.01
obs 5	0.94	0.06	0.00

ex. Observation 4 sur le vecteur 2
$31.14^2/(16.6^2+33^2+5.4^2) = 969.7/1393.72 = 0.70$

Qualité de représentation des variables

	u_1	u_2	u_3
x_1	0.99	0.00	0.00
x_2	0.18	0.82	0.00
x_3	0.74	0.25	0.01

ex. Variable 3 sur le vecteur 1:
$(-22.37)^2/671.24 = .74$

Contribution des observations $(C_o(.)^2/n)\ \Lambda^{-1}$
Note : le facteur n est introduit pour tenir compte que les coordonnées calculées plus haut sont XU et non $XU/n^{0.5}$.

	u_1	u_2	u_3
Obs 1	0.13	0.01	0.32
Obs 2	0.41	0.00	0.03
Obs 3	0.03	0.72	0.03

Obs 4	0.05	0.21	0.43
Obs 5	0.38	0.04	0.19

ex. Observation 4 sur le 3e vecteur
$3.94^2/(5*7.28) = 0.43$

Contribution des variables ($U(.)^2$)

	u_1	u_2	u_3
x_1	0.62	0.00	0.38
x_2	0.09	0.81	0.10
x_3	0.29	0.19	0.53

ex. Variable 1 sur le 3e vecteur
$.62^2 = .38$

Réponses aux questions et exercices

Question 1:

On a: $X'Xu_1=\lambda_1 u_1$

$1/n\ X'Xu_1=1/n\lambda_1 u_1$

Les vecteurs propres sont les mêmes, les valeurs propres sont divisées par "n".

Question 2:

$C_0C_0'=XUU'X'=XX'$ sur la diagonale, on a les longueurs carrés des observations. Donc la somme des qualités de représentations est 1. La démonstration est analogue pour les variables.

Question 3:

$C_0'C_0=U'X'XU=\Lambda^{1/2}V'V\Lambda^{1/2}=\Lambda$

La somme des contributions des observations sur un vecteur propre est donc 1.

Question 4:

Proportions: 37.9/44.4 6.6/44.4

La somme est égale à la trace de la matrice de covariances (acp de la matrice des covariances) donc elle est égale à la somme des variances des variables originales.

Les v.p. ont été calculés avec X centré (matrice des covariances). Toutefois, les projections des X sur le 1^{er} v.p. indiquent que X n'a pas été centré avant d'effectuer la projection (toutes les valeurs sont positives). Ceci ne change rien toutefois aux positions relatives des points (simple translation).

Question 5:

La valeur prédite est sur le 1^{er} v.p. et $x_2=19$. La projection sur le 2^e v.p. doit donner 0. donc:

$0 = (x_1, 19)* (.75 ; -.66)$

$.75\ x_1= 19*.66$

$x_1=16.7$

Question 6 :

Les points représentants les diverses variables seront très proches l'un de l'autre sur tous les plans factoriels. Une variable bien représentée sur un plan factoriel aura sa projection près du périmètre du cercle unité centré à l'origine.

Question 7 :

Les vecteurs ont toutes leurs composantes égales à $(1/2)^{1/2}$ à un signe près. Le vecteur u_1 sera orienté à 45^0 (antihoraire par rapport à « x ») si la corrélation entre les deux variables est positive, sinon il sera orienté à -45^0.

Exercice 1: (.66, .75) * (.75 ; -.66) =0

RÉFÉRENCES

Cooley W.W. et Lohnes P.R., 1971. Multivariate data analysis. Wiley, 364p.

Davis J. C., 1986. Statistics and data analysis in geology (2e ed.). Wiley, 646p.

Greenacre M.G., 1984. Theory and applications of correspondence analysis. Academic Press, 364p.

Lebart L., Morineau A. et Warwick K.M., 1984. Multivariate descriptive statistical analysis. Wiley, 231p.

Legendre L. et Legendre P., 1984. Écologie numérique. Masson, 2 tomes.