GLQ3402 **Traitement statistique des données géologiques**
9h30 à 12h00 **EXAMEN FINAL** **6 juillet 2005**

Calculatrice et documentation permises.
L'examen comprend 5 questions totalisant 50 points répartis de la façon suivante :
1-6 pts, 2-16 pts, 3-5 pts, 4-8 pts, 5-15 pts

---

6 1- La matrice suivante présente les distances euclidiennes entre les six observations mesurées dans un espace de 4 variables.

	Obs 1	Obs 2	Obs 3	Obs 4	Obs 5	Obs 6
Obs 1	0.00	9.01	8.26	12.32	10.46	10.15
Obs 2	9.01	0.00	6.10	11.86	9.84	11.21
Obs 3	8.26	6.10	0.00	8.69	5.57	9.89
Obs 4	12.32	11.86	8.69	0.00	3.78	5.84
Obs 5	10.46	9.84	5.57	3.78	0.00	6.01
Obs 6	10.15	11.21	9.89	5.84	6.01	0.00

Indiquez, en faisant les calculs appropriés, les trois premiers regroupements à se produire si l'on utilise la méthode hiérarchique du « complete linkage ».

---

16 2- On a effectué l'ACP de la matrice de corrélations d'un jeu de 10 données avec 3 variables. Les principaux résultats obtenus sont :

	Vecteurs propres		
	u_1	u_2	u_3
Var 1	-0.70	0.05	-0.71
Var 2	-0.15	-0.98	0.08
Var 3	**a**	0.17	0.70
Valeurs propres	1.68	**b**	0.33

(Note : les valeurs propres λ_i représentent les variances, les dispersions sur chaque vecteur propre s'obtiennent en prenant « n λ_i »)

Coordonnées des observations

	u_1	u_2	u_3
Obs 1	1.48	-1.92	-0.08
Obs 2	0.02	0.44	1.38
Obs 3	-0.04	-0.85	0.02
•	•	•	•
Obs 10	-0.14	0.73	-0.56

Qualité de la représentation des observations

	u_1	u_2	u_3
Obs 1	0.37	**c**	0.00
Obs 2	0.00	0.09	0.91
Obs 3	0.00	1.00	0.00
•	•	•	•
Obs 10	0.02	0.62	0.36

Contribution des observations

	u_1	u_2	u_3
Obs 1	**d**	0.37	0.00
Obs 2	0.00	0.02	0.58
Obs 3	0.00	0.07	0.00
•	•	•	•
Obs 10	0.00	0.05	0.09

Coordonnées des variables

	u_1	u_2	u_3
Var 1	-0.91	0.05	-0.41
Var 2	-0.20	-0.98	0.05
Var 3	-0.90	0.17	0.40

Qualité de représentation des variables

	u_1	u_2	u_3
Var 1	0.83	0.00	0.17
Var 2	0.04	0.96	0.00
Var 3	**e**	0.03	0.16

Contribution des variables

	u_1	u_2	u_3
Var 1	0.49	0.00	**f**
Var 2	0.02	0.97	0.01
Var 3	0.48	0.03	0.49

i. Indiquez les valeurs pour les lettres « a » à « f ».

*ii. Si l'on reconstruit la matrice X à l'aide de la relation X=co*u', qu'obtiendra-t-on exactement, la matrice initiale ou la matrice X centrée réduite? Expliquez brièvement comment enlever l'effet du 2e vecteur propre dans la reconstruction.*

iii. Une 4e variable est disponible. Ses corrélations avec les coordonnées des observations sur les vecteurs propres 1 à 3 sont respectivement de –0.9, 0.2 et 0.39. Avec laquelle des variables originales cette variable est-elle la plus corrélée? Justifiez.

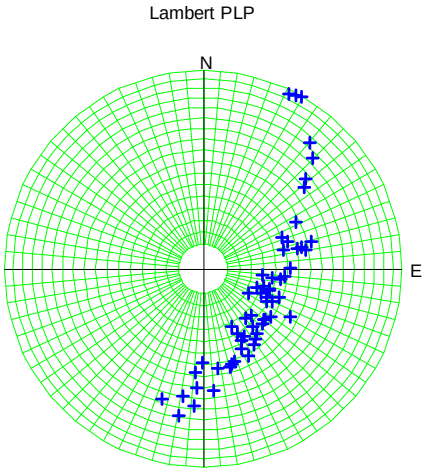
---

5 3- Sur un site contaminé de la région de Québec, on désire connaître la conductivité hydraulique (zone saturée) afin de mieux prévoir le déplacement des contaminants. On pense que la conductivité est étroitement liée à la nature des sédiments rencontrés sur le site. Quatre faciès principaux sont observés : sable grossier, sable moyen, sable fin, silt. Les faciès ont pu être observés grâce à des forages de type « rotosonic » qui permettent d'obtenir en continu et sur de grandes profondeurs des échantillons non remaniés du sol. Un pénétromètre a été aussi utilisé à proximité de plusieurs forages « rotosonic » pour déterminer s'il existait des corrélations intéressantes entre les paramètres mesurés par le pénétromètre (vitesse de pénétration, force exercée sur les tiges, friction enregistrée le long de la pointe du pénétromètre, etc.) et les faciès observés dans les échantillons « rotosonic ». Les forages « rotosonic » reviennent à environ 80$/m foré. Les essais au pénétromètre à environ 5$/m.

Quelle étude suggérez-vous d'effectuer afin de déterminer si le pénétromètre peut être utilisé avantageusement en complément des forages « rotosonic » pour déterminer les faciès sur le site étudié ?

---

8 4- La figure suivante montre les pôles de 60 fractures sur une projection polaire à aires égales.



Les résultats de l'ACP sur les pôles sont :

	Vecteurs propres		
	u_1	u_2	u_3
x	-0.39	-0.52	0.76
y	0.31	-0.85	-0.42
z	0.86	0.07	0.50
Valeurs propres	0.76	0.23	0.005

Note : les coordonnées des vecteurs propres sont donnés dans un système main droite (i.e. z croissant vers le haut). L'azimut et la plongée suivent la convention géologique (plongée positive pointant vers le bas).

a) Donnez l'azimut et la plongée du pôle moyen de ces fractures.

b) Quelle loi de distribution des pôles (spécifiez la loi et donnez ses paramètres) pourrait être compatible avec la distribution observée ?

c) Trouvez la projection du pôle ayant azimut=63^o et plongée=32^o sur les 2^e et 3^e vecteurs propres.

d) Le pôle décrit en c) se situe-t-il à l'intérieur de l'ellipse de confiance de niveau 90% autour du vecteur pôle moyen ?

15 5- Des relevés géophysiques ont identifié un grand nombre d'anomalies conductrices. Parallèlement, des variables géologiques ont été notées à proximité de ces zones conductrices :

- MET : niveau de métamorphisme (sur une échelle croissante de métamorphisme de 0 à 3)
- RHY :présence de rhyolites à proximité de la zone conductrice (0 : absence, 1 présence)
- FRA : présence de fractures importantes ((0 : absence, 1 présence)
- DEN : densité de la roche (obtenue par inversion gravimétrique)

Des forages ont été menés sur les anomalies conductrices. Un certain nombre ont révélé uniquement la présence de graphite (stérile), d'autres la présence de sulfures (minéralisé). On effectue une AD pour distinguer les conducteurs stériles des conducteurs minéralisés. Les principaux résultats obtenus de l'AD sont :

Ordre et statistique de sélection des variables selon le V de Rao :

Ordre	Variable	V de Rao
1	MET	225.79
2	FRA	291.01
3	DEN	303.76
4	RHY	309.17

	Vecteur propre u_1
MET	0.28
FRA	-0.03
DEN	-0.16
RHY	-0.56
Valeur propre	6.4

Les coordonnées des groupes stérile et minéralisé sur ce vecteur propre sont : 0.358 et –0.358

Le % de bien classés est de 0.98.

Les fonctions de classification sont :

	Stérile	Minéralisé
MET	-5	-15
FRA	24	25
DEN	20	25
RHY	888	907
Cte	-1286	-1294

a) Est-on justifié de retenir les 4 variables dans cette AD ? Pourquoi?

b) Que valent : u'Du, u'Eu et u'Tu ?

c) Interprétez en terme géologique le vecteur propre issu de cette AD.

d) Calculez la probabilité d'appartenir au groupe « minéralisé » pour une zone conductrice ayant un niveau métamorphique 1, présence de fractures importantes, densité de 2.7 et absence de rhyolite .

e) Indiquez trois conditions requises pour que la probabilité calculée en d) soit valable.

f) Commentez l'énoncé suivant : « Lorsque les matrices de covariance intra-groupes ne sont pas égales, l'on peut quand même, sous certaines hypothèses, effectuer le classement en fonction de la probabilité d'appartenance à chaque groupe. Ceci revient géométriquement à classer chaque observation dans le groupe dont le centre est le plus près selon la distance de Mahalanobis ».

Bon examen, et bonnes vacances!

Denis Marcotte

Corrigé

Question 1- 1er regroupement, 4 avec 5 au niveau d=3.78

On recalcule la matrice de distance

	Obs 1	Obs 2	Obs 3	Obs 4 et 5	Obs 6
Obs 1	0.00	9.01	8.26	12.32	10.15
Obs 2	9.01	0.00	6.10	11.86	11.21
Obs 3	8.26	6.10	0.00	8.69	9.89
Obs 4 et 5	12.32	11.86	8.69	0.00	6.01
Obs 6	10.15	11.21	9.89	6.01	0.00

On regroupe 6 avec 4 et 5 au niveau d= 6.01

On recalcule la matrice de distance

	Obs 1	Obs 2	Obs 3	Obs 4+5+6
Obs 1	0.00	9.01	8.26	12.32
Obs 2	9.01	0.00	6.10	11.86
Obs 3	8.26	6.10	0.00	9.89
Obs 4+5+6	12.32	11.86	9.89	0.00

On regroupe les observations 2 et 3 au niveau 6.1

Question 2
i.
a) il faut que u_1'u1=1 => a^2=0.4875 => a= + ou – 0.698. Le signe est déterminé en considérant que u_1'*u_2=0 => a= -0.698

b) la somme des valeurs propres est égale à la trace de la matrice des corrélation, soit le nombre de variables (puisque chaque élément de la diagonale vaut 1). Donc ici b=3-1.68-0.33=0.99

c) La somme des qualités de la représentation sur les vecteurs propres donne 1. Donc : c= 1-0.37=0.63.

d) d= coordonnée au carré /valeur propre du 1er vecteur : 1.48^2/(10*1.68)=0.13

e) e=1-0.03-0.16=0.81

f) 0.70^2=0.49 (aussi 1-0.01-0.49=0.49 aux arrondis près)

ii. La matrice centrée-réduite. On enlève la 2^{e} colonne de co et la 2^{e} colonne de u et on calcule co*u' donc :

iii. Les corrélations sont les coordonnées de cette variable. Par simple examen, on constate que celle-ci est très près de la variable 3. Le cosinus de l'angle est presque 1.

Question 3- Il serait intéressant d'effectuer une AD. Les groupes seraient les faciès vus au rotosonic, les variables mesurées seraient les variables du pénétromètre. Si l'AD est un succès, elle permettra d'identifier les faciès à partir des variables au pénétromètre. On pourra alors obtenir à faibles coût un grand nombre d'identifications de faciès, ce qui permettra de mieux estimer les conductivités hydrauliques sur le site.

Question 4- a) Le pôle moyen est le 1er v.p. de l'ACP. L'azimut est donné par $\tan^{-1}(-0.39/(-0.31))=128.5^{o}$. La plongée est $\sin^{-1}(0.86)=59.3^{o}$. Visuellement, ceci semble correspondre au point moyen du graphe des pôles.

b) Une loi binormale asymétrique avec paramètres $\lambda_2 = 0.23, \lambda_3 = 0.005$ et pôle moyen correspondant au 1er vecteur propre.

c) Ses coordonnées cartésiennes (vecteur unitaire) sont :
[sin(63)cos(32), cos(63)cos(32), -sin(32)]=[0.756, 0.385, -0.53]
Les projections sur les v.p. 2 et 3 sont : [-0.76, 0.15]

d) On calcule $0.76^2/0.23+0.15^2/0.005=2.02$
Cette valeur est inférieure à une $\chi^2_{2,0.9} = 4.6$ qui délimite l'ellipse. Elle est donc à l'intérieur de l'ellipse.

Question 5

a) Oui, la valeur limite pour le V de Rao (ajout) est $\chi^2_{1,0.95} = 3.84$. On mesure des écarts de 225, 66, 13 et 5.4.

b) u'Du vaut 1 par construction; u'Eu=6.4 et donc u'Tu=7.4

c) Le groupe minéralisé se trouve sur le côté négatif de ce vecteur, donc la présence de rhyolites, la densité plus grande de la roche, l'absence ou le faible niveau de métamorphisme et dans une moindre mesure la présence de fractures sont tous des facteurs favorisant la présence de minéralisation.

d) g1=-5+24+2.7*20-1286=-1213
g_2=-15+25+2.7*25-1331=-1216.5

p2=$[e^{3.5}+1]^{-1}$ =2.9 %

e) 1- mêmes matrices de covariances, 2-distribution multinormale dans chaque groupe, 3-l'observation appartient nécessairement à un des groupes et 4- les probabilités à priori pour chaque groupe sont égales.

f) Faux. Oui on peut calculer les probabilités d'appartenance à chaque groupe sous hypothèse multinormale en estimant une matrice de covariance différente pour chaque groupe. Cependant ce critère n'est alors plus équivalent au critère de la distance de Mahalanobis la plus petite. L'équivalence existe seulement dans le cas linéaire.

GLQ3402 **Traitement statistique des données géologiques**
9h30 à 12h00 **EXAMEN FINAL** **30 juin 2004**

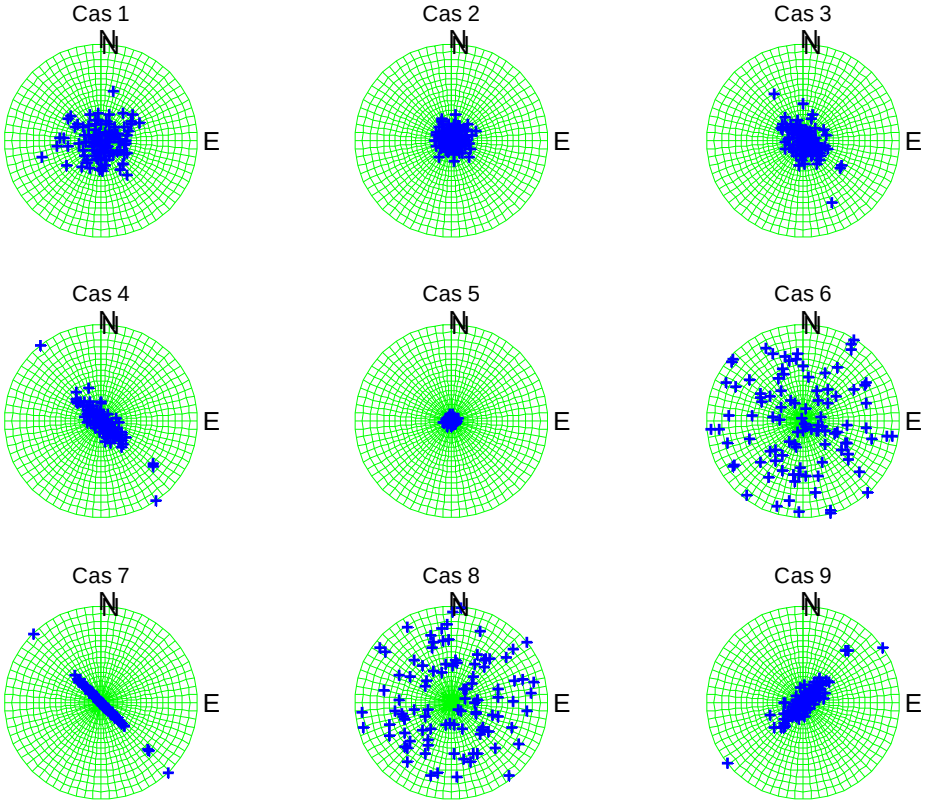
Calculatrice et documentation permises.
L'examen comprend 6 questions totalisant 50 points répartis de la façon suivante :
1-5 pts, 2-10 pts, 3-5 pts, 4-12 pts, 5-6 pts, 6-12 pts

---

5 1- *Effectuez la segmentation de l'image suivante. Le seuil de différence entre germe et pixels est 3.5. Un pixel est voisin d'un autre pixel s'ils partagent un côté commun (i.e. voisin par le côté et non par le sommet). Débutez la segmentation par le pixel #8 puis enchaînez avec le pixel #18. Indiquez votre réponse en donnant les # des pixels se rattachant à chacun des germes.*

#5 1	*#10* 7	*#15* 10	*#20* 2	*#25* 5
#4 2	*#9* 2	*#14* 3	*#19* 7	*#24* 10
#3 5	*#8* **2**	*#13* 10	*#18* **3**	*#23* 4
#2 5	*#7* 8	*#12* 7	*#17* 5	*#22* 3
#1 9	*#6* 1	*#11* 7	*#16* 9	*#21* 1

2- La figure suivante montre la distribution des pôles sur stéréonet. (Note : le pôle moyen a été tourné au centre du stéréonet pour une meilleure visualisation).



Chacune des lois suivantes correspond à au moins un des cas illustrés sur la figure.

Loi	Description
A	Distribution complètement aléatoire
B	Loi Fisher avec k=500
C	Loi binormale avec $\lambda_2 = 0.1$ et $\lambda_3 = 0.03$
D	Loi binormale avec $\lambda_2 = 0.15$ et $\lambda_3 = 0.0001$
E	Loi Fisher avec k=10
F	Loi binormale avec $\lambda_2 = 0.15$ et $\lambda_3 = 0.01$
G	Loi Fisher avec k=30
H	Loi Fisher avec k=>0

3 a) *Associez chacune des lois A à H à au moins un cas de la figure précédente.*

Pour un cas donné où l'on disposait de 50 pôles, on a effectué l'ACP de la matrice d'orientation. On a obtenu comme 1er vecteur propre : [0.3536 -0.3536 -0.8660]. La valeur propre associée est $\lambda_1 = 0.8$. Les autres valeurs propres sont $\lambda_2 = 0.1$ et $\lambda_3 = 0.1$.

2 b) *Donnez l'azimut et la plongée (convention géologique) du pôle moyen de cette famille de joints ainsi qu'un estimé approximatif du paramètre de concentration.*

2 c) Dans le même contexte que la question b), un pôle montre : azimut= 120 et plongée=50.

- *Quel est l'angle θ que forme ce pôle avec le vecteur moyen ?*
- *Quelle est la probabilité (loi de Fisher) de trouver un pôle formant un angle inférieur ou égal à θ avec le pôle moyen de cette famille ? (Prenez θ=10 si vous n'avez pas pu calculer la valeur de l'angle à la question précédente)*

3 d) Toujours dans le même contexte qu'en b), un pôle utilisé dans l'ACP de la matrice d'orientation avait pour coordonnées cartésiennes : [0.5567 -0.3214 -0.7660].

- *Quelle est la coordonnée de ce pôle sur le 1er vecteur propre ?*
- *Quelle est la contribution de ce pôle à la définition du 1er vecteur propre (aide : la dispersion totale sur le premier vecteur propre, par rapport à l'origine, est $n\lambda_1$) ?*
- *Quelle est la qualité de la représentation de ce pôle sur le 1er vecteur propre ?*

3- Soixante analyses géochimiques de ciment sont effectuées par un cimentier. Huit éléments majeurs sont analysés. Les analyses ferment toutes exactement à 100% (i.e. la somme des 8 éléments donne 100% pour chaque observation). De plus, deux variables supplémentaires sont calculées à l'aide des formules de Bogue :

$$C_3S = (4.071 \times CaO) - (7.6 \times SiO_2) - (6.718 \times Al_2O_3) - (1.43 \times Fe_2O_3) - (2.852 \times SO_3)$$

$$C_2S = (2.867 \times SiO_2) - (0.7544 \times C_3S)$$

Les 10 variables sont utilisées dans une ACP de la matrice des corrélations.

2 a) Quelle sera la somme des valeurs propres issues de cette analyse ?

3 b) Combien de valeurs propres égales à 0 devraient être obtenues de cette ACP ? Justifiez.

4- Dans une étude portant sur les dommages aux dalles de béton des résidences causés par le gonflement du remblai, on a identifié l'âge de la résidence (en années), le calibre du remblai (0 : étalé, 1 : net), l'IPPG du remblai (0 à 100) et la présence de polyéthylène (0 : présent; 1 : absent) sous la dalle comme des facteurs potentiellement importants pouvant expliquer l'occurrence de dommages. Deux groupes de résidences furent formés, le groupe des résidences sans dommages observables (groupe 1) et le groupe des résidences avec dommages observables (groupe 2).

Voici les principaux résultats de l'AD :

Valeur propre : 0.38

Vecteur propre : [-0.008 0.064 -0.002 -.093]

Sur le vecteur propre, le centre du groupe sans dommages (gr 1) est projeté en 0.05, celui du groupe avec dommages (gr 2) en -0.08.

Fonctions de classification :

Variable	g1	g2
Âge	0.16	0.24
Calibre	1.6	0.8
IPPG	0.04	0.06
Polyéthylène	1.6	2.6
Cte	-2.1	-5.0

2 *a) Interprétez le premier vecteur propre.*

2 *b) Que vaut u_1'Eu_1 ?*

Une résidence a 10 ans, le remblai est net, l'IPPG du remblai est 20 et il y a un polyéthylène sous la dalle. Calculez les probabilités d'appartenance à chaque groupe sous hypothèse normale et égalité des matrices de covariance.

2 c) *Présentement, dans quelle groupe cette résidence devrait-elle être classée ? Calculez la probabilité d'appartenance à chaque groupe. Au fil des années qu'advient-il de cette probabilité* ?

2 *d) Y a-t-il quelque chose dans les informations fournies qui pourrait inciter à douter de la validité de l'hypothèse de normalité des distributions ?*

2 *e) Si les probabilités calculées ne sont pas valables, quelle interprétation peut-on tout de même donner au classement obtenu ?*

2 *f) Outre l'hypothèse de multinormalité et l'égalité des matrices de variance-covariance des deux groupes, quelle est la condition essentielle pour que les probabilités calculées en c) soient fiables ?*

5- On désire regrouper des calcaires selon leur degré de ressemblance géochimique. On a 7 observations pour lesquelles 8 éléments majeurs sont analysés. La distance euclidienne calculée dans l'espace des 8 éléments majeurs est donnée par la matrice suivante :

0	50	12	32	31	30	39
50	0	59	39	33	26	29
12	59	0	33	33	21	31
32	39	33	0	23	27	8
31	33	33	23	0	21	28
30	26	21	27	21	0	1
39	29	31	8	28	1	0

2 a) *Réalisez la première fusion et recalculez la distance entre le groupe fusionné et les autres observations par la méthode « single linkage »; quelle serait la seconde fusion à intervenir ?*

2 *b) Faites la même chose qu'en a) mais cette fois par « complete linkage »;*

2 *c) Tracez les dendrogrammes représentant ces deux premières fusions pour chacune des deux méthodes.*

6- Commentez les énoncés suivants :

2 *a) On peut utiliser les résultats de l'ACP pour définir une équation de prédiction de n'importe qu'elle variable par l'ensemble des (p-1) autres variables. Cette équation de prédiction est définie par le premier vecteur propre issu de l'ACP.*

2 *b) Si l'on effectue une classification automatique non-hiérarchique en minimisant la dispersion dans les groupes, on est assuré d'une excellente discrimination entre les groupes obtenus.*

2 *c) On peut discriminer parfaitement des basaltes de différentes origines par une AD linéaire. Si l'on soumet les mêmes données à une méthode de classification non-hiérarchique en demandant le même nombre de groupes, on devrait retrouver à peu de choses près les mêmes groupes que les groupes initiaux (i.e. les basaltes de différentes origines).*

2 *d) On a trois variables bruitées. Le bruit est d'importance comparable sur chaque variable. On effectue une ACP et l'on identifie un des vecteurs propres comme représentant essentiellement le bruit sur les trois variables. En reconstruisant la matrice X sans ce vecteur propre, on vérifie que l'on obtient des données exemptes de bruit. On peut conclure que le bruit sur chaque variable était indépendant.*

2 *e) En ACP, une observation très bien représentée sur un vecteur propre contribue toujours beaucoup à la définition de ce vecteur propre.*

2 *f) En ACP, une observation contribuant beaucoup à la définition d'un vecteur propre est toujours bien représentée.*

Bonne chance et bonnes vacances !
Denis Marcotte

Corrigé

1- À partir du germe #8, on peut joindre : 2 3 4 5 8 9 14
À partir du germe 18, on joint 17 18 21 22 23
Les autres sont orphelins

2- a)

Loi	Description	Cas
A	Distribution complètement aléatoire	6 et 8
B	Loi Fisher avec k=500	5
C	Loi binormale avec $\lambda_2 = 0.1$ et $\lambda_3 = 0.03$	3
D	Loi binormale avec $\lambda_2 = 0.15$ et $\lambda_3 = 0.0001$	7
E	Loi Fisher avec k=10	1
F	Loi binormale avec $\lambda_2 = 0.15$ et $\lambda_3 = 0.01$	4 et 9
G	Loi Fisher avec k=30	2
H	Loi Fisher avec k=>0	6 et 8

b) atan(.3536/(-.3536))=-45 azimut=90+45=135
plongée : asin(0.8660)=60
paramètre de concentration : $1/(1-0.8^{0.5})=9.47$

c) On convertit le pôle en coordonnées cartésiennes
x=sin(120)*cos(50)=0.5567
y=cos(120)*cos(50)=-0.3214
z=-sin(50)=-0.7660

Le cosinus de l'angle entre les deux vecteurs est simplement le produit scalaire des deux vecteurs :
[0.3536 -0.3536 -0.8660]*[0.5567 -0.3214 -0.7660]'=0.9739
acos(0.9739)=13.1 degrés

La probabilité est donnée par : (exp(k)-exp(k*cos(theta))/(exp(k)-1)=13.4% avec theta=10 et 21.8% avec theta=13.1

d) la coordonnée est 0.9739 (même calcul qu'en c)
contribution : 0.9739^2/(50*0.8)=2.37%
0.9739^2/1=0.95

3- a) somme=nombre de variales soumis à ACP de la matrice des corrélations = 10
b) Il y aura 3 valeurs propres nulles. La fermeture indique que les 8 variables occupent un espace à 7 dimensions. Les 2 variables supplémentaires sont dans ce même espace. Donc 7 v.p. couvrent tout l'espace.

4-
a) les maisons plus âgées ont une plus grande probabilité de montrer des dommages, un remblai au calibre étalé est plutôt défavorable, un IPPG élevé est défavorable et la présence de polyéthylène est très favorable.

b) c'est la valeur propre par définition 0.38
c) Il suffit de calculer les gi avec ces valeurs. On trouve :
g1=10*0.16+1*1.6+20*0.04 +0*1.6-2.1 = 1.96
g2=10*0.24+1*0.8+20*0.06+0*2.6-5.0= -0.6

$P(1|x)=(1+\exp(-2.56))^{-1}=0.93$
$P(2|x)=(\exp(2.56)+1)^{-1}=0.07$

On doit classer l'observation dans le groupe 1 (sans dommages). Au fur et à mesure que la maison vieillit, on a une contribution de supplémentaire de .08 par année à g2. Comme l'écart est de 2.56, il faudra 2.56/.08=32 ans de plus avant que la maison passe dans le groupe 2.

d) oui, deux des variables « x » sont discrètes donc sûrement non normales.

e) l'observation est classée dans le groupe dont elle est le plus près au sens de la distance de Mahalanobis.

f) il faut que l'observation à classer appartienne nécessairement à l'un des groupes étudiés.

5- 1ère fusion single et complete : 6 avec 7

Nouvelles distances 6+7 vs

#	Single	Complete
1	30	39
2	26	29
3	21	31
4	8	27
5	21	28

La prochaine fusion : single : 4 avec 6+7 (distance=8)
Complete : 1 avec 3 (distance =12)

c) dessin

6- a) Faux, c'est le dernier vecteur propre qui définit cette équation de prédiction.
b) oui car c,est précisément le critère utilisé en AD
c) Pas nécessairement, on pourrait identifier des groupes encore plus différents
d) Faux. Au contraire, si le bruit sur les 3 variables est représenté par un seul vecteur c,est donc qu'il est parfaitement corrélé d'une variable à l'autre.

e) pas toujours, ex, une observation près de la moyenne peut être très bien représentée et ne rien contribuer.

f) pas nécessairement, une observation peut contribuer beaucoup à plus d'un facteur et donc la qualité de sa représentation est partagée entre ces facteurs.

GLQ3402 -- Examen de mi-session

Jeudi le 12 juin 2003
8h30 à 11h00

Toute documentation permise.
Calculatrice permise

L'examen comporte 9 questions totalisant 30 points. Chaque réponse doit être accompagnée des calculs et justifications appropriés.

Les questions valent dans l'ordre : 3, 7, 3, 3, 3, 4, 2, 3, 2

Points

3 1- Dans Excel, la fonction « Growth » (ou « Croissance » en français) permet d'estimer le modèle :

$$\hat{Y} = b_0 b_1^x$$

a) Indiquez comment vous pourriez « linéariser» ce modèle et ainsi obtenir des estimés pour b_0 et b_1 avec un programme de régression linéaire. Indiquez clairement le vecteur « y » et la matrice « x » qui seront soumis au programme de régression, les coefficients obtenus par la régression et le lien avec les coefficients recherchés.

b) Les prédictions $\hat{Y}$ *obtenues avec ce modèle minimiseront-elles la somme des carrés des erreurs (si l'on définit « e » comme* $e = Y - \hat{Y}$ *) ? Justifiez.*

---

7 2- On effectue une régression pas à pas (avant) avec 100 observations. On veut expliquer une variable Y avec au plus 4 variables X (modèle avec constante). Les 2 premières variables à être incluses dans la régression sont X_2 et X_4. À cette étape, le R^2 vaut 0.8 et SCE vaut 1000. L'on vous fournit la matrice de **variances-covariances partielles** (étant donné l'effet de X_2 et X_4 fixé) suivante :

	Y	X_1	X_3
Y	10	14.23	28.3
X_1	14.23	25	33.54
X_3	28.3	33.54	125

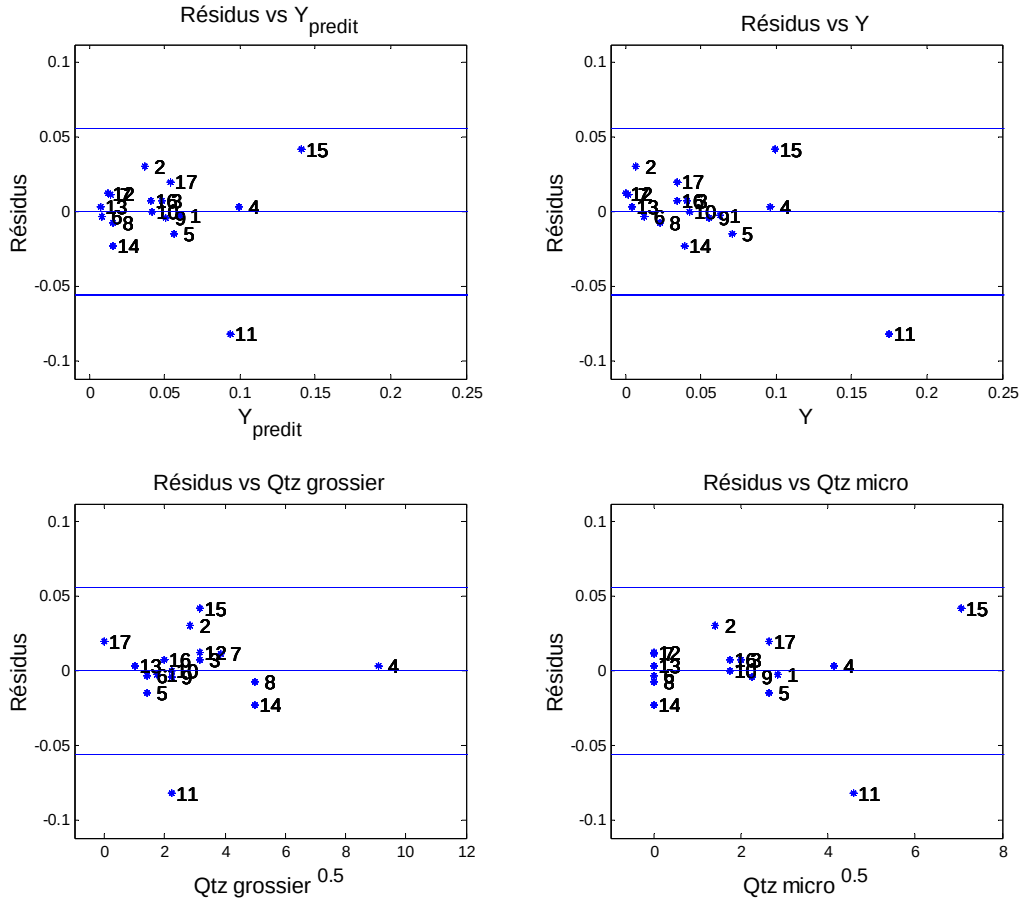
a) Quelle sera la prochaine variable à être sélectionnée par la procédure pas à pas avant dans le modèle de régression? (Aide : quel est le lien entre covariance partielle et corrélation partielle?)
b) Quelle est le coefficient de la régression associé à cette variable?
c) Les coefficients de régression associé à X_2 et X_4 seront-ils modifiés suite à l'inclusion de cette nouvelle variable?
d) Que devient le R^2 suite à l'inclusion de cette variable?
e) L'ajout dû à cette variable est-il significatif? Faites le test requis en indiquant clairement les degrés de liberté associés. (Aide, calculez d'abord SCT_m).

3 3- On a déterminé la porosité d'échantillons de résidus miniers provenant de 2 sites différents.

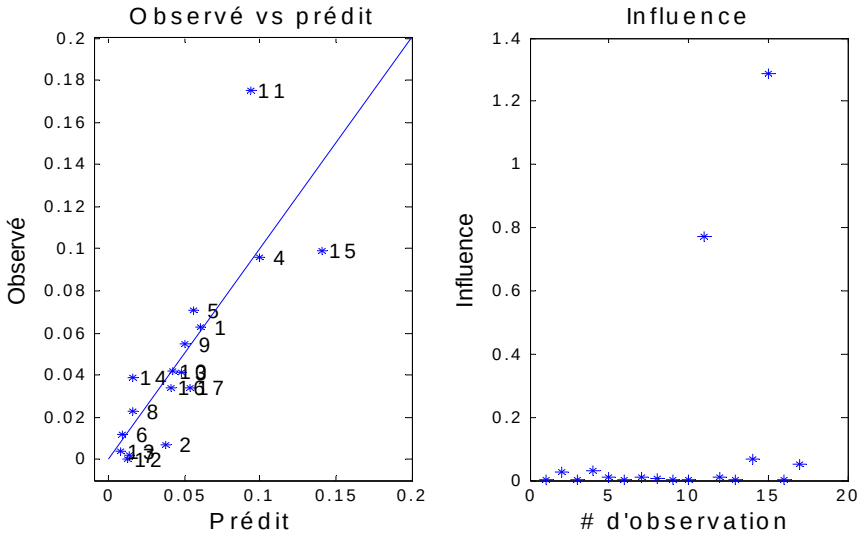
Expliquez comment on peut utiliser un programme de régression pour tester l'égalité des moyennes des porosités des 2 sites. (Aide : on doit utiliser le test d'ajout)

---

3 4- On a mesuré la déformation de 17 éprouvettes de béton après une cure humide d'un an. On a aussi déterminé le % de quartz micro-cristallin (chert et calcédoine) et le % de grains de quartz plus grossiers pour chaque éprouvette. Ces déterminations ont été réalisées par comptage microscopique. L'objectif de la régression est de prédire la déformation après un an en utilisant les % de quartz comme variables prédictives. Les graphiques suivants montrent les résidus obtenus en fonction de certaines variables. Les limites correspondant à + ou – 2 *CME$^{0.5}$ sont aussi illustrées :



Le graphique suivant montre le graphe des valeurs observées vs prédites et celui de l'influence des observations :



Les résultats présentés vous semblent-ils conformes à ce qui est attendu ? Justifiez votre réponse.

---

3 5- Toujours avec le jeu d'éprouvettes de béton de la question précédente, on a appliqué la norme ACNOR qui indique qu'un béton doit être rejeté s'il subit une déformation supérieure à 0.04. Ceci permet de définir deux groupes de béton, ceux ayant réussi et ceux ayant échoué le test. On applique une régression logistique à ces données et l'on obtient le modèle suivant :

$$\hat{W} = -3.6 + 0.07 * Qtz\ \ grossier(\%) + 0.8 * Qtz\ \ micro(\%)$$

a) Supposons qu'un béton contienne 5% de quartz grossier. Quelle quantité de quartz micro-cristallin peut-il posséder avant que l'on ne doive rejeter le béton (si l'on se base sur l'équation de prédiction logistique)?

b) Le béton contient 20% de quartz grossier et 8% de quartz micro-cristallin. Quelle est la probabilité qu'il échoue le test de la cure humide?

---

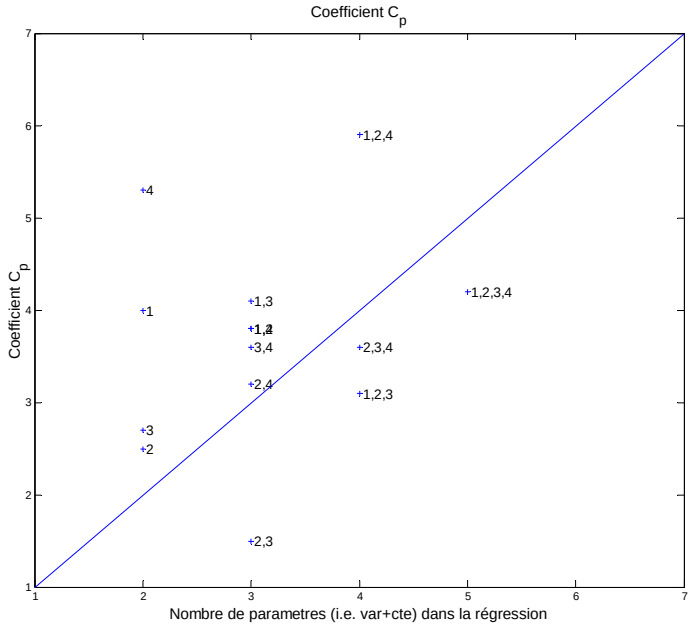
4 6- Soit la matrice de corrélation simple suivante obtenue avec les données géochimiques de l'Abitibi utilisées lors du TP-4.

	SiO_2	Al_2O_3	FeO	MgO	CaO	Na_2O	K_2O	TiO_2
SiO_2	1.00	-0.84	-0.65	-0.54	-0.80	-0.48	0.31	-0.84
Al_2O_3	-0.84	1.00	0.31	0.27	0.61	0.51	-0.08	0.65
FeO	-0.65	0.31	1.00	0.59	0.31	-0.02	-0.37	0.63
MgO	-0.54	0.27	0.59	1.00	0.32	-0.21	-0.27	0.39
CaO	-0.80	0.61	0.31	0.32	1.00	0.44	-0.25	0.64
Na_2O	-0.48	0.51	-0.02	-0.21	0.44	1.00	-0.44	0.43
K_2O	0.31	-0.08	-0.37	-0.27	-0.25	-0.44	1.00	-0.35
TiO_2	-0.84	0.65	0.63	0.39	0.64	0.43	-0.35	1.00

a) Calculez la corrélation partielle entre FeO et MgO quand l'on fixe l'effet du SiO_2.
b) On effectue la régression de « FeO expliqué par SiO_2 » et la régression de « MgO expliqué par SiO_2 ». Quelle est la corrélation simple entre les résidus de ces deux régressions?
c) Quelle est la corrélation simple entre le résidu de la régression « FeO expliqué par SiO_2 » et le SiO_2?

---

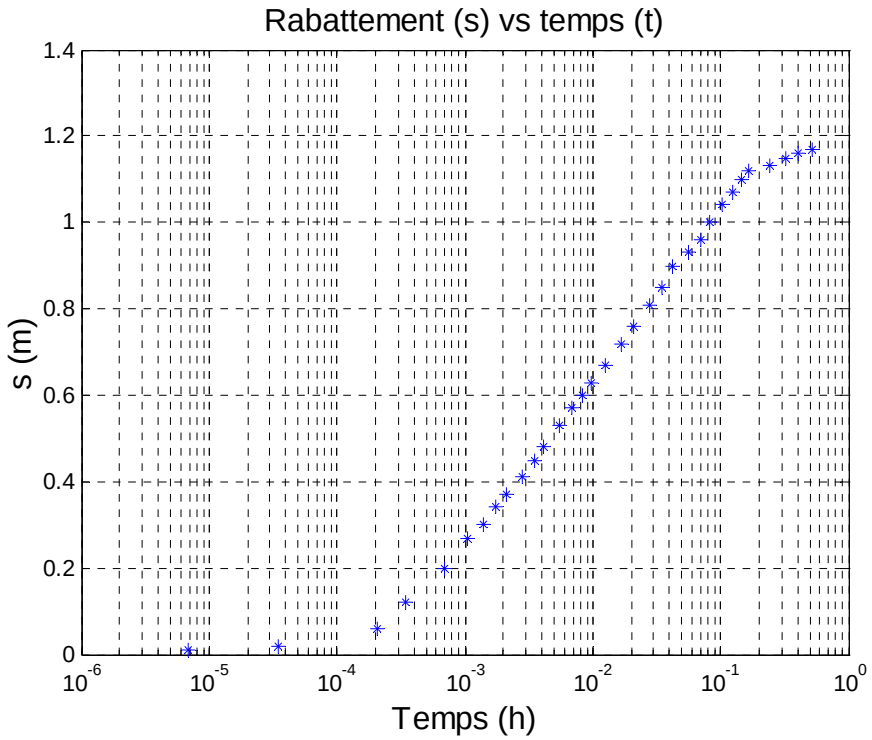
2 7- Soit le diagramme suivant illustrant le coefficient C_p obtenu pour différents sous-ensembles de variables.



Quel ensemble de variables ce graphique vous suggère-il retenir? Justifiez.

---

3 8- Dans un aquifère à nappe confinée, homogène, de grande extension latérale et d'épaisseur constante, l'on dispose d'un piézomètre ayant une crépine sur toute l'épaisseur de l'aquifère. On mesure le rabattement (s) en fonction du temps en vu d'estimer la transmissivité (T) et le coefficient d'emmagasinement (S) à l'aide de la méthode de Cooper-Jacob qui consiste à ajuster une droite sur la partie linéaire du graphe s vs ln(t). Le graphique suivant montre le rabattement en fonction du temps.



Suggérez un outil vu au cours qui pourrait permettre d'identifier et d'éliminer de façon itérative les points qui s'éloignent trop de la partie droite de la courbe.

---

2 9- On veut utiliser une sonde mesurant la conductivité électrique en continu sur un convoyeur pour identifier les fragments minéralisés et les fragments non-minéralisés en vu de les séparer. La teneur économique pour le gisement est 0.5% Ni équivalent (le seuil définissant « minéralisé »). On définit le Ni équivalent comme Ni+0.8*Cu. Vous avez mesuré la teneur chimique en Ni et en Cu sur 100 fragments et avez mesuré sur ces mêmes fragments la conductivité électrique. Vous notez une relation linéaire assez forte entre le log(conductivité) et le Ni équivalent.

Quel modèle de régression utiliserez-vous pour rencontrer les objectifs de l'étude?

Bonne chance,

Denis Marcotte

Corrigé :

1- a) En prenant les logs, on obtient : $ln(\hat{Y}) = ln(b_0) + ln(b_1) * x$. On a donc une régression linéaire avec constante de ln(Y) prédit par x. On obtient c_0 et c_1 de cette régression et $b_0=\exp(c_0)$ $b_1=\exp(c_1)$.

b) Comme la régression s'effectue dans un espace transformé, elle minimise la somme du carré des erreurs de la variable transformée ln(Y) et non du Y lui-même. Si l'on veut minimiser la somme du carré des erreurs, il faut effectuer une régression non-linéaire.

2- a) On cherche la variable avec la plus forte corrélation partielle avec Y. On calcule pour
X_1 : $14.23/(10*25)^{0.5}=0.9$
X_3 : $28.3/(10*125)^{0.5}=0.8$
On doit donc inclure X_1.

b) Le coefficient de la régression est : 14.23/25=0.569
c) Oui bien sûr puisque b est obtenu par $(X'X)^{-1}X'Y$ et que la matrice X change de dimension.
d) Le nouveau R^2 peut être calculé par la relation 2.30 donnée en page 37.
$R^2=0.8+0.9^2*(1-0.8)=0.962$
e) Le SCE avec X_2 et X_4 vaut 1000, on déduit que $1-R^2=0.2=1000/SCT_m$ et $SCT_m=5000$.
Après avoir inclus X_1, on a $1-R^2=0.038=SCE(c)/5000$ donc SCE(c)=190.
Le test d'ajout est donc $F_{calculé}=\dfrac{(1000-190)/1}{190/(100-4)}=409.3$, ce qui est très supérieur à la valeur

$F_{table,.1,96,05}=3.94$

3- On forme un premier modèle réduit $Y=b_o+e$. On forme un second modèle complet : $Y=b_0+b_1*I+e$ où I est une variable indicatrice prenant la valeur 0 pour un site et 1 pour l'autre. Ce modèle permet d'avoir deux moyennes différentes comme estimation pour les deux sites alors que le modèle réduit ne permets qu'une seule moyenne. Il ne reste qu'à tester le caractère significatif de l'ajout. Si c'est significatif alors les moyennes diffèrent, significativement, sinon, les moyennes peuvent être considérées égales.

4- On note que l'observation 15 montre une trop grande influence. De plus l'observation 11 montre aussi une grande influence et son résidu est à l'extérieur de l'intervalle de confiance calculé avec $CME^{0.5}$. Cette observation est donc aussi très suspecte. Il faudrait possiblement refaire l'analyse sans ces 2 observations.

5- a) Le seuil de classification correspond à $\hat{W}=0$. On calcule donc
% qtz micro<(3.6-0.07*5)/0.8=4.06%.
b) On calcule $\hat{W}=-3.6+0.07*20+0.8*8=4.2$
$\hat{Y}=exp(4.2)/(1+exp(4.2))=0.985$

6- a) On calcule $r_{Feo,MgO|SiO2} = \frac{0.59 - (-0.65) * (-0.54)}{\left((1 - 0.65^2)(1 - 0.54^2\right)^{0.5}} = 0.37$

b) Par définition, elle est égale à la corrélation partielle calculée en a)

c) Les résidus d'une régression sont non-corrélées aux variables « x » entrant dans la régression. Donc cette corrélation est 0.

7- Tous les ensembles ayant les variables x_2 x_3 sont sous la diagonale. Parmi ceux-ci le couple (2,3) est celui le plus sous la diagonale. Il s,agit probablement du meilleur sous-ensemble à retenir.

8- La notion d'influence d'une observation peut permettre de détecter les observations s'écartant de la droite. En enlevant les observations les plus influentes et en faisant le suivi du R^2 obtenu, on pourra déterminer un sous-ensemble de points situés sur une droite.

9- Le modèle aura la forme : Ni équivalent=b_0+b_1 log(conductivité électrique)+e

GLQ3402 **Traitement statistique des données géologiques**
13h à 15h30 **EXAMEN FINAL** **19 juin 2002**

Calculatrice et documentation permises.
L'examen comporte 6 questions totalisant 100 points répartis de la façon suivante :
1-14, 2-21, 3-15, 4-20, 5-15, 6-15

---

14 1- On mesure sur le terrain la direction de 109 stries glaciaires (azimut).Vous inspirant du traitement appliqué pour les pôles des fractures, expliquez comment vous pourriez utiliser une ACP pour :

a) déterminer la direction moyenne des stries;

b) déterminer si les directions des tries sont très dispersées ou non autour de la direction moyenne.

Note : Indiquez clairement quelles données seraient soumises à l'ACP et la matrice dont seraient extraits les valeurs propres et vecteurs propres.

---

21 2- On a effectué l'ACP des coordonnées des intersections centrales de forages avec une zone minéralisée d'épaisseur non négligeable (ACP de la matrice des covariances). On a obtenu les résultats suivants :

Valeurs propres		Vecteurs propres		
		u_1	u_2	u_3
$\lambda_1 = 5.6$	x	-0.96	0.20	-0.21
$\lambda_2 = 3.2$	y	-0.06	0.58	0.81
$\lambda_3 = 3.0$	z	0.28	0.79	-0.54

a) Quelle est l'attitude (direction et plongée) de l'axe parallèle à la principale dimension de la zone minéralisée? Note, les coordonnées sont dans un système de référence main droite (i.e. z pointe vers le haut, y vers le nord et x vers l'est).

b) La variance des coordonnées « x » vaut 5.4 , celle des coordonnées « y » 3.1. Que vaut la variance des coordonnées « z » ?

c) Que vaut la contribution de la variable x au 1er vecteur propre.

d) Quelle est la coordonnée de la variable y sur le 3^{e} vecteur propre?

e) Les coordonnées des variables sont-elles à l'intérieur du cercle unité? Pourquoi?

---

15 *3- Esquissez, pour un cas avec 2 variables, des jeux de données pour lesquels l'ACP fournit les éléments suivants (un jeu différent pour chaque question) :*

a) Deux observations contribuent chacune à 50% à la définition du 2e vecteur propre.

b) La 2e valeur propre de l'ACP vaut 0.

c) Le 1er vecteur propre de l'ACP est orthogonal au 1er vecteur propre discriminant (cas de 2 groupes).

d) Le 1er vecteur propre de l'ACP est parallèle au 1er vecteur propre discriminant (cas de 2 groupes).

---

20 4- Dans le but de prédire le résultat au test de gonflement ACNOR des bétons, l'on a déterminé les % de phases siliceuses pour 17 agrégats soumis au test ACNOR. L'AD a été effectuée en prenant comme variables x_1 : % de quartz et x_2 % de quartz microcristallin. Le groupe 1 est constitué des agrégats ayant passé le test ACNOR, le groupe 2 de ceux ayant échoué le test. Voici les principaux résultats obtenus :

Pas	Variable inclue	V de Rao
1	Quartz microcristallin	9.6
2	Quartz grossier	19.3

Le vecteur propre discriminant est :

Quartz grossier	-0.1
Quartz microcristallin	-0.17

La valeur propre associée à ce vecteur vaut 1.29

Les coordonnées des groupes ayant passé ou échoué le test sur le v.p. discriminant sont :

Passé le test	0.26
Échoué le test	-0.29

Les moyennes globales sont : quartz grossier : 9.5%, quartz microcristallin : 4.9%.

Les fonctions de classification (g_i) sont :

	Groupe passé le test	Groupe échoué le test
Quartz grossier	0.43	1.23
Quartz microcristallin	0.33	1.75
Constante	-0.35	-4.13

a) Selon ces résultats, la discrimination est-elle significative? La variable « quartz grossier » améliore-t-elle la discrimination? Faites le test requis. Note : la table χ^2 est fournie en annexe.

b) Que vaut le Lambda de Wilks?

c) Un agrégat présentant 5% de quartz grossier et 3% de quartz microcristallin devrait-il être utilisé dans les bétons selon le résultat de l'AD? Justifiez

d) Calculez la probabilité qu'un béton construit avec cet agrégat échoue le test ACNOR.

---

15 5- En agriculture de précision, les agronomes essayent d'identifier des parcelles de terrain le plus homogène possible pour les propriétés physiques et chimiques des sols. Le but est de déterminer la quantité d'amendement optimale pour chaque parcelle homogène. Le critère d'optimalité ici est le rendement obtenu en biomasse en fonction de la quantité d'engrais utilisé. Les données disponibles sont habituellement obtenues sur une grille régulière ponctuelle (espacée) et comprennent des analyses géochimiques, des analyses granulométriques des sols, la mesure de la conductivité électrique du sol (EM-38), des descriptions de texture du sol, etc.

a) Quelle méthode permettrait d'identifier des parcelles homogènes ?

b) Comment identifieriez-vous le nombre de parcelles homogènes à former?

c) Supposons que les groupes obtenus soient spatialement très éclatés (i.e. une parcelle est constituée de plusieurs sous-zones non contiguës). Que pourriez-vous tenter pour aider former des parcelles spatialement plus continues ?

---

15 6- Supposons que l'on effectue une ACP (matrice de corrélation) de roches volcaniques où les variables analysées sont les 8 éléments majeurs habituels (Si, Al, Fe, Mg, Ca, Na K, Ti). Les roches couvrent les compositions allant de basaltes à rhyolites.

a) À quoi devrait correspondre normalement le 1er vecteur propre dans ce contexte (en terme géologique) ?

b) Si l'on effectue une classification non-hiérarchique en utilisant une distance euclidienne habituelle (i.e. sans normalisation), obtiendra-t-on le même résultat avec les 8 variables originales (sans transformation) qu'avec les 8 composantes principales issues de l'ACP de la matrice des corrélations ? Justifiez.

c) Comment pourriez-vous obtenir des groupes, par classification non-hiérarchique, libres de l'influence du facteur géologique décrit en a) ?

Bonne chance et bonnes vacances !

Corrigé :
1- On exprime les azimuts sous forme de vecteurs unitaires (repère cartésien). Chaque vecteur est doublé pour éviter les problèmes de mauvaise orientation du vecteur. On obtient ainsi une matrice X_{218x2}. On effectue l'ACP de X'X/218. La somme des valeurs propres sera 1.
a) Le 1er vecteur propre donnera la direction moyenne.
b) La 2e valeur propre sera une mesure de dispersion autour de la direction moyenne. Plus la 2e valeur propre sera faible, plus les directions des stries seront concentrées autour du vecteur moyen.

2-.
a) la direction et plongée sont données par le 1er vecteur propre
direction atan(-0.96/-0.06)=86.2°
pendage=asin(0.28)=16.3°
b) la somme des valeurs propres est 11.8. La somme des variances de x et y est 8.5, donc celle de z est 11.8-8.5=3.3.
c)0.96^2=92%
d) $0.81*3^{0.5}$=1.40
e) non parce qu'on a fait une ACP de la matrice des covariances et non de la matrice des corrélations.

3- a) les données forment une droite, sauf 2 observations situées sur le 2e v.p. de part et d'autre du 1er vecteur propre et à égale distance.
b) les données forment une droite.
c) 2 nuages parallèles très allongés.
d) 2 nuages allongés qui se suivent selon la direction d'allongement.

4- a) On compare l'accroissement du V de Rao à une chi^2 avec 1 degré de liberté (=3.84 au niveau 0.05). Le quartz microcristallin est significatif. L'ajout du quartz grossier est aussi significatif.
b) le lambda de Wilks est 1/(1+1.29)=0.437
c) Calculons les gi
g1=5*0.43+3*0.33-0.35=2.79
g2=5*1.23+3*1.75-4.13=7.27

clairement, l'agrégat est classé dans le groupe 2 (échoué le test) et donc il ne devrait pas être utilisé dans le béton.

d) $[\exp(2.79-7.27)+1]^{-1}$=0.989

5 a) Une méthode de classification automatique, préférablement de type non-hiérarchique. On peut aussi utiliser une méthode hiérarchique de type average-linkage.
b) non-hiérarchique : on fait des essais avec 2,3,... groupes et on trace le graphe du critère de dispersion utilisé dans l'algorithme vs le nombre de groupes. On recherche un niveau ou le critère diminue rapidement puis se stabilise en fonction d'un nombre de groupe croissant.
hiérarchique : l'examen du dendrogramme peut permettre de détecter le nombre de zones vraiment distinctes.
c) On pourrait entrer les coordonnées x et y comme variables dans la classification en donnant un poids aasez grand à ces variables.

6- a) différentiation magmatique. C'est l'élément principal de variation de composition chimique dans les roches volcaniques.

b) Non, l'ACP de la matrice des corrélations travaille avec les variables normalisées (i.e. centrées-réduites). La classification, dans un cas utilisera des coordonnées dans un espace normalisé, dans l'autre cas dans l'espace original.

c) On pourrait effectuer la classification en utilisant les composantes principales des vecteurs 2 à 8. On pourrait aussi reconstruire la matrice X en filtrant l'influence du 1er vecteur propre et en effectuant la classification avec la matrice X filtrée.

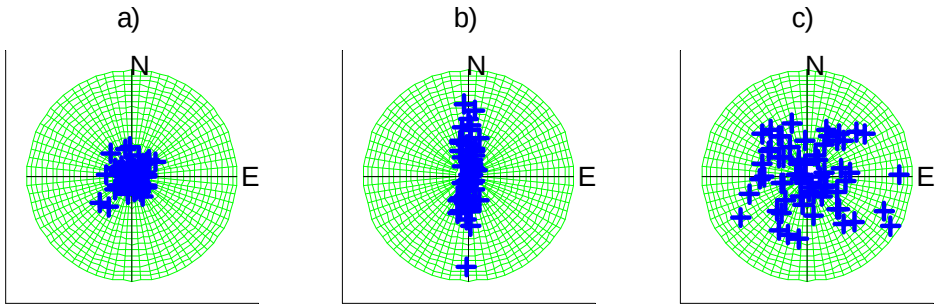
GLQ3402 **Traitement statistique des données géologiques**
13h à 15h30 **EXAMEN FINAL** **20 juin 2001**

Calculatrice et documentation permises.
L'examen comporte 7 questions totalisant 100 points répartis de la façon suivante :
1-18, 2-10, 3-10, 4-15, 5-10, 6-21, 7-16
Justifiez toutes vos réponses.

---

6 *1 a) Décrivez, en termes qualitatifs, le lien existant entre la 1ère valeur propre d'une ACP d'une matrice d'orientation et le paramètre de concentration "k=N/(N-R)".*

b) On vous présente les projections de 3 groupes de pôles de joints sur un stéréonet à aires égales (figure suivante). Le pôle moyen des joints a été placé au centre du stéréonet dans chaque cas pour faciliter la visualisation.



6 *Identifiez le groupe ayant, apparemment:*
i.- distribution de Fisher avec k=5
ii. distribution de Fisher avec k=50
iii. une distribution qui n'est pas Fisher

6 *c) Un ensemble de 30 fractures (plans) a été modélisé par une loi de Fisher. Le pôle moyen est orienté (0,0,1) (en x,y,z) et le vecteur somme des composantes montrait R=28.8 . Quelle est la probabilité qu'une fracture (pas son pôle) présente un pendage inférieur à 30°?*

---

10 2- Une ACP de la matrice des corrélations (5 variables) a donné une 1ère valeur propre de 1.
Que peut-on dire de l'ACP (autres valeurs propres et vecteurs propres) et de la disposition des points dans l'espace des variables (après avoir "centré-réduit" celles-ci)?

---

10 3. *Effectuez la segmentation de l'image suivante à partir des germes des cellules #12 et #14 dans cet ordre. Utilisez comme distance la différence (en valeur absolue) entre pixel et germe avec un seuil pour cette différence de 2.5. Indiquez votre solution en donnant les # des cellules dans chaque région ainsi formée.*

#21 2	#22 9	#23 6	#24 8	#25 10
#16 2	#17 2	#18 2	#19 3	#20 2
#11 9	#12 ***3***	#13 7	#14 ***9***	#15 5
#6 10	#7 2	#8 5	#9 9	#10 8
#1 4	#2 4	#3 7	#4 10	#5 9

---

4. Dans une étude statistique des expertises CTQ-M100, l'on a pu établir par analyse discriminante une équation de prédiction de l'occurrence de dommages à une dalle de béton, au sous-sol, en fonction de l'IPPG du remblai, du calibre de celui-ci et de l'âge de la maison. L'AD a été faite sur deux groupes, celui des dalles avec dommages importants et celui des dalles avec très peu de dommages. On a obtenu près de 84% de bien classés. L'équation linéaire discriminante est la suivante:

Projection sur v.p. discriminant= -69+0.8*âge + 1.1*IPPG+36*calibre du remblai

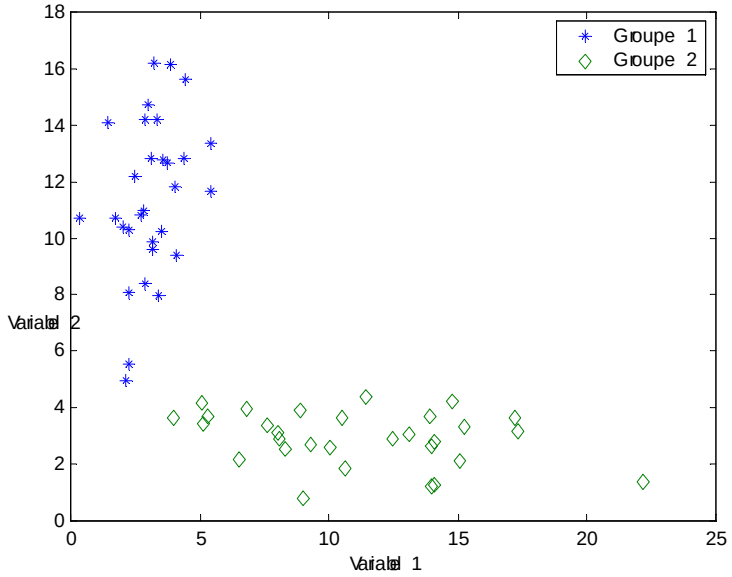
L'âge est donné en années,
l'IPPG en %,
le calibre du remblai vaut 1 si remblai net, 2 sinon.
La constante effectue le centrage des données.
Les groupes endommagés et non-endommagés présentaient des moyennes respectivement de 56 et –2.

8 *a) Une dalle sur remblai ayant IPPG de 20 et un calibre net, devrait-elle être endommagée après 15 ans selon ces résultats?*

7 *b) À très long terme, que se passe-t-il avec cette équation de prédiction indépendamment de la nature du remblai? Suggérez une transformation de la variable âge qui atténuerait ou éliminerait ce problème.*

---

10 5- La figure suivante montre les points de 2 groupes pour lesquels 2 variables sont mesurées.



Peut-on calculer la probabilité d'appartenance de chaque observation à chaque groupe en utilisant les résultats d'une analyse discriminante linéaire ? Si oui, expliquez la démarche à suivre. Sinon expliquez pourquoi.

---

21 6. Des analyses géochimiques de roches de l'Abitibi sont soumises à une ACP de la matrice des corrélations. Trois variables sont utilisées dans l'analyse: Si, Mg, et Na. On obtient les résultats suivants:

2 premiers vecteurs propres:

Si	u1=	a	u2= -0.63
Mg		0.42	0.78
Na		0.72	0.01

1ère valeur propre

$?_1 =$ 1.47

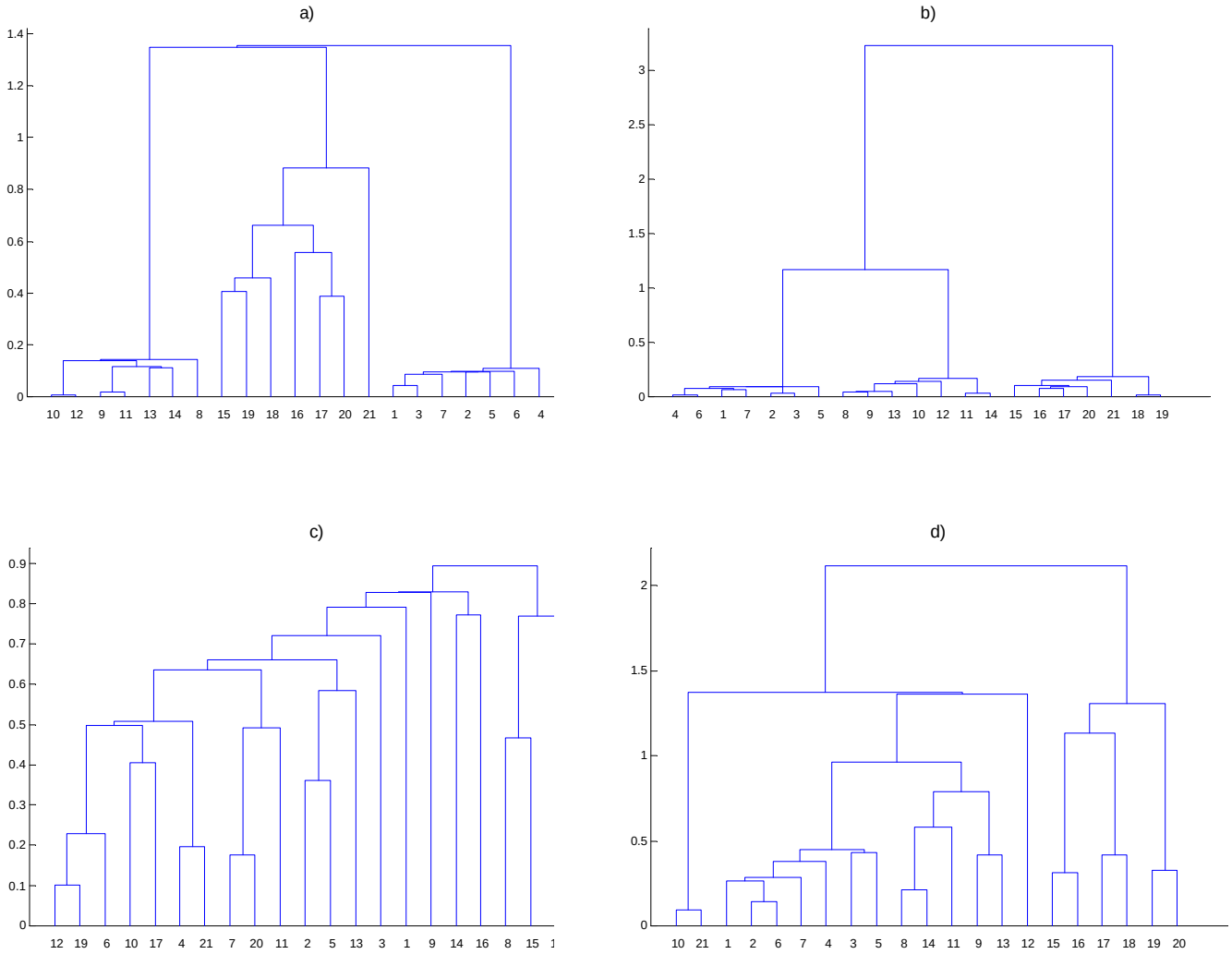
Observation 1 (centrée-réduite)

Si=-0.67 Mg= -0.25 Na=0.30

a) Que vaut a) dans u1 ?
b) Quelle est la coordonnée de l'observation 1 sur le 1er vecteur propre?
c) Quelle est la qualité de la représentation de cette observation sur le 1er vecteur propre?
d) Quelle est la contribution de cette observation (la somme des carrés des projections vaut 147) au 1er v.p.?
e) Quelle est la coordonnée de la variable Na sur le 1er v.p.?
f) Quelle est la contribution de la variable Na au 1er v.p.?
g) Quelle est la qualité de la représentation de la variable Na sur ce vecteur?

8 7. a) Les méthodes de classification non-hiérarchiques peuvent être influencées par la présence de bruit dans les données. *Expliquez comment l'ACP pourrait être utile dans ce contexte pour permettre d'améliorer les classifications.*

8 b) *Parmi les dendrogrammes suivants obtenus par classification hiérarchique (average linkage), indiquez celui qui correspond le mieux à la description suivante: "On remarque deux groupes très compacts et très séparés et un 3e groupe plus diffus mais néanmoins bien distinct des 2 autres".*



Bonne chance et bonnes vacances

Denis Marcotte

Corrigé

1- a) Plus la valeur propre augmente plus la concentration des pôles augmente et donc plus k augmente.

b) i.-c; ii-a) iii. b)

c) k=30/(30-28.8)=25
Prob(pendage <30)=(exp(k)-exp(kcos(30))/(exp(k)-exp(-k))= 0.96

2- La somme des valeurs propres donne 5. Si la plus grande est 1 il faut que toutes les autres soient aussi 1. Ceci implique que les vecteurs propres sont complètement arbitrairement orientés (i.e. toutes les directions sont équivalentes en terme de dispersion). Les points, une fois normalisés, sont donc nécessairement soit sur ou dans une sphère et ils sont parfaitement uniformément répartis.

3- 1er groupe : 1,2,7,8,12,15,16,17,18,19,20,21

2e groupe: 3,4,5,9,10,13,14

4- a) On calcule sco=-69+0.8*15+1.1*20+36*1=1.0. L'observation est beaucoup plus proche du groupe non-endommagé que endommagé. On la classe dans ce groupe.

b) En vieillissant toute maison sera classée comme présentant des dommages. Ainsi, le point milieu est situé en 27. Pour un IPPG de 0, un calibre net, toute maison plus vieille que (27+69-36)/0.8=75 ans sera classée comme endommagée.
On pourrait par exemple utiliser le log de l'âge de la maison ce qui diminuerait le taux de progression avec l'âge ou encore une fonction tronquée genre agecodé=min(ageréel,40) qui limiterait l'influence de l'âge pour les grandes valeurs.

5- Non, le postulat que chaque groupe doit présenter la même matrice de covariance intra-groupe n'est pas respecté ici.

6- a) –0.63*a+0.42*0.78+0.01*0.72=0 a=0.53
b) –0.67*0.53-0.25*0.42+0.30*0.72=-0.24
c) 0.24^2/(0.67^2+0.25^2+0.3^2)=.096
d) 0.24^2/147=0.0004
e) $0.72*1.47^{0.5}=0.87$
f) 0.72^2=0.52
g) 0.86^2/1=0.74

7 a) On pourrait effectuer une ACP, identifier les vecteurs pouvant représenter du bruit ou des facteurs de variation indésirables dans les données et reconstruire la matrice X sans l'effet de ces facteurs. Ensuite on procéderait à la classification basée sur la matrice reconstruite X.

b) Le dendrogramme a)

GLQ3402 Traitement statistique des données géologiques
13h à 15h30 EXAMEN FINAL 21 juin 2000

Calculatrice et documentation permises.
Petits Phils interdits.
L'examen comporte 7 questions totalisant 100 points répartis de la façon suivante :
1-20, 2-10, 3-15, 4-15, 5-20, 6-10, 7-10
Justifiez toutes vos réponses.

---

1. Vous effectuez le relevé de joints (structures planaires) dans des galeries de mine. Vous mesurez le pendage et la direction de ces joints. Chaque joint est ensuite représenté par son vecteur normal unitaire (système main droite). Vous avez relevé 50 joints et vous faites une ACP de la matrice d'orientation des pôles.

Vous obtenez: λ_1=44
λ_2=3.1
λ_3=2.9

De plus u_1 = [1/3 2/3 -2/3] u_2 = [-2/3 2/3 1/3] u_3 = [2/3 1/3 2/3]

4 a) Quelle est la direction (azimut) et l'inclinaison du pôle moyen ?

4 b) Est-il plausible que la distribution des pôles suive une loi de Fisher ? Quel serait le paramètre "k" de cette loi si R=45?

6 c) Un des pôles est: [.5 .5 -.707]. Calculez la coordonnée de ce joint sur le vecteur propre u_1. Calculez aussi la qualité de représentation de l'observation sur ce vecteur ainsi que la contribution de cette observation à la définition du vecteur.

6 d) La somme au carré de la composante x des vecteurs pôles donne 17. Quelles sont la coordonnée, la contribution et la qualité de représentation de cette composante sur le 1er vecteur propre ?

---

2. Esquissez des nuages de points, dans l'espace de deux variables, tel que:

4 a) Une observation A est très bien représentée sur le premier vecteur propre mais contribue peu à la définition de celui-ci. Indiquez la position approximative du 1er vecteur propre.

4 b) Une observation B contribue beaucoup à la définition du premier vecteur propre mais est moyennement bien représentée sur ce vecteur.

2 c) En régression, on pouvait détecter une donnée extrême par la mesure de son influence. Parmi les quantités définies pour l'ACP, en voyez-vous une qui pourrait jouer ce rôle ?

---

3. Identifiez pour chacun des problèmes ci-dessous la méthode de traitement la plus appropriée. Indiquez les détails nécessaires (groupes, variables, sorte d'analyse, etc.)

5 a) On a foré plusieurs anomalies électro-magnétiques. Certains de ces forages ont permis de détecter de la minéralisation, d'autres étaient tout à fait stériles. On a également analysé des roches d'affleurements voisins pour leurs éléments majeurs et mineurs. On aimerait pouvoir déterminer si une anomalie est intéressante sans avoir à effectuer le forage.

5 b) Vous avez plusieurs observations et plusieurs variables. Vous voulez construire un index qui synthétise la variation présente dans les données. Vous voulez que cet index soit indépendant du choix des unités de mesure de vos variables.

5 c) Dans une séquence stratigraphique de calcaires, d'apparence homogène, vous voulez identifier des calcaires semblables à partir des résultats d'analyse géochimiques.

---

10 4. a) Tracez le cercle unité de l'ACP de la matrice des corrélations et positionnez approximativement les variables X_1 à X_7 tel que:

- X_1 et X_5 sont très corrélées.
- X_1 est représentée à plus de 95% sur le 1^{er} v.p.
- X_4 est indépendante de X_5 et est mal représentée sur le plan des deux premiers v.p.
- X_2 et X_6 sont mal corrélées avec X_5 et contribuent beaucoup au 2^{e} v.p.
- La corrélation entre X_2 et X_6 est négative.
- X_3 est fortement corrélée négativement à X_5.
- X_7 est mal corrélée avec toutes les autres variables.

5 b) Donnez un estimé approximatif de la proportion de l'inertie expliquée par chacun des deux premiers vecteurs propres.

---

5. On a effectué une AD avec 100 observations réparties en trois groupes. On a mesuré deux variables.

Les valeurs propres obtenues sont :

$l_1 = 3.5, l_2 = 1.2$

On a trouvé les fonctions de classification suivantes:

	g_1	g_2	g_3
X_1	1	1	2
X_2	0	1	-1
cte	1	1	0

5 a) Classez l'observation ayant X_1=4, X_2=2.

5 b) Calculez la probabilité d'appartenance de cette observation au groupe 2.

5 c) Quelles sont les trois conditions pour que la probabilité calculée en b) soit valide?

5 d) Que vaut le Lambda de Wilks ?

---

10 6. Tracez un dendrogramme (complet) qui réponde à la description suivante: parmi 11 observations, on remarque deux groupes de quatre observations très compacts et très différents l'un de l'autre. Trois observations semblent intermédiaires entre ces deux groupes.

---

10 7. En AD, on suppose habituellement que les matrices de dispersion intra-groupes sont égales pour tous les groupes. Expliquez la conséquence sur l'AD de ne pas faire cette hypothèse (et donc de supposer que les groupes ont des matrices de covariance différentes).

Bonne chance et bonnes vacances !

Denis Marcotte

Corrigé

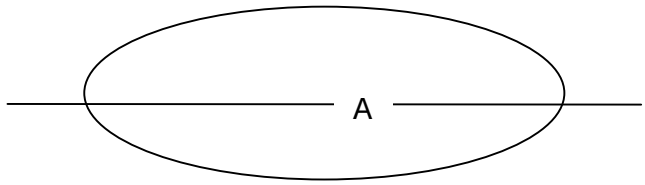
1 a) azimut : atan(1/2)=26.6
inclinaison : asin(2/3)=41.8

b) Oui c'est plausible car les valeurs propres 2 et 3 sont presque égales, ce qui semble indiquer une symétrie radiale. De plus, la 1ere valeur propre indique une forte concentration des pôles.
Le paramètre « k » est N/(N-R)=50/5=10;

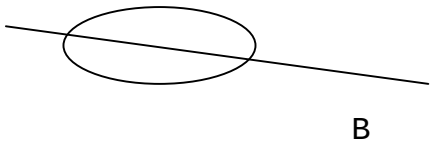
c) coordonnée : [.5 .5 -.707]*[1/3 2/3 –2/3]'=0.971
Qualité de la représentation : 0.971^2=94.3%
Contribution : 0.971^2/44=0.021, soit 2%

d) Coordonnée : 1/3*$(44)^{0.5}$=2.21
contribution : $1/3^2$=1/9
qualité : 2.21^2/17=28.7%

2- a)



b)



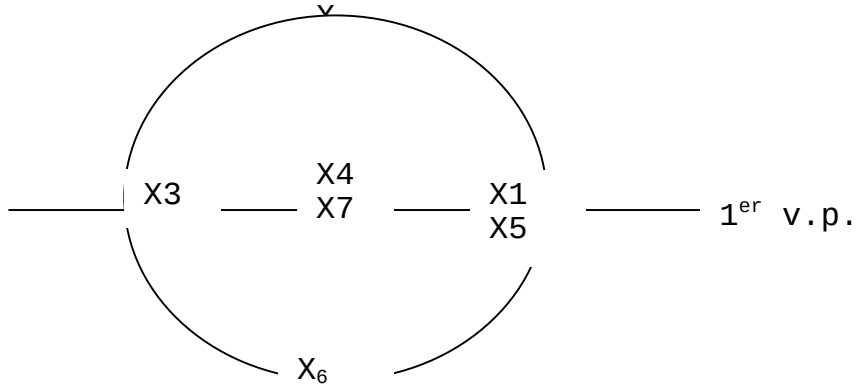
c) La contribution des observations est une mesure analogue à l'influence. Les observations contribuant beaucoup à définir un vecteur doivent être examinées avec soin.

3 a) On peut effectuer une AD o;u les groupes sont formés en fonction des résultats des forages. On essayerait de distinguer les groupes selon les compositions chimiques. Sur une nouvelle anomalie, on pourrait prendre des échantillons de roches et déterminer leur groupe probable (minéralisé ou non).

b) ACP de la matrice des corrélations.

c) Classification non-hiérarchqie. On pourrait aussi utiliser une ACP pour tenter de visualiser des regroupemeents d'observations.

4-



b) 1er vp.p 3/7, 2e v.p., 2/7; total 5/7

5- g1=5 g2=7 g3=6

a) L'observation se classe dans le groupe 2
b) $[\exp((5-7)+1+\exp(6-7)]^{-1}=0.66$
c) Probabilités à priori égales. Matrices de covariances intra-groupes égales. Distibution multinormale. (Aussi, il faut que l'observation appartienne nécessairement à un des 3 groupes étudiés).

d) On calcule le Lambda de Wilks. On trouve : gamma1=1/(1+3.5)=.222 et gamma 2=1/(1+1.2)=0.454
Le lambda de Wilks est donc : .222*.454=.108

6- dessin

7- Si les matrices de dispersion ne sont pas égales, alors la surface de séparation entre les groupes devient une surface quadratique. On a plus de degrés de liberté, mais en même temps, les résultats deviennent moins robustes.