

Exercice 1

Assez P !

20 points

La figure 2 (ci-jointe) présente les données d'un fichier `petrology_data.xls` contenant les mesures d'oxydation de 35 échantillons de pétrole recueilli dans le pacifique nord. Sur chaque échantillon la concentration de 11 oxydes a été mesurées. Ces oxydes sont (dans l'ordre)

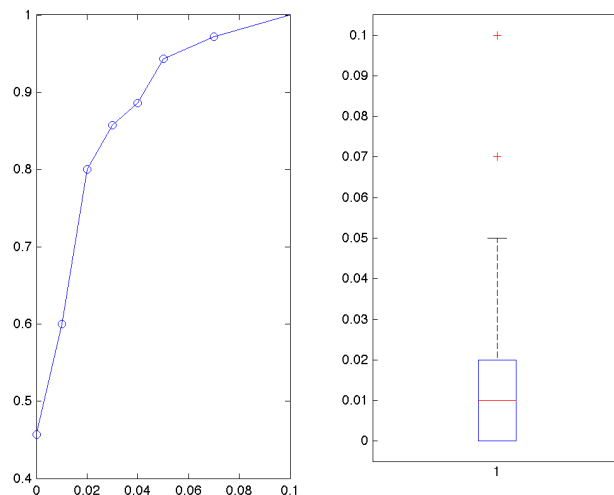
SiO_2 , TiO_2 , Al_2O_3 , Cr_2O_3 , FeO , NiO , MnO , MgO , CaO , Na_2O , et K_2O .

Chaque échantillon a été associé à trois variables supplémentaires : la latitude, la longitude et l'élévation du point de mesure.

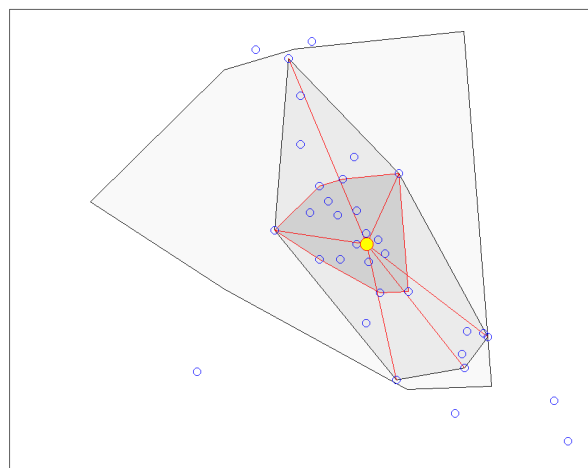
À la suite de l'énoncé, vous trouverez la copie d'une session Matlab. On y trouve des lignes de code qui commencent par le chargement

Rappel : la fonction matlab `[V,lam] = eig(M)` calcule les vecteurs propres (V) et les valeurs propres associées (sur la diagonale de `lam`) de la matrice M .

1. Tracer la fonction de répartition empirique de la concentration en NiO
2. Dessinez la boîte à moustache de la variable concentration en NiO



3. Sur la figure 3, tracez le sac à moustaches du nuage de points associé au couple de variable TiO_2 et Al_2O_3 . Quelle est la corrélation et la covariance entre ces deux variables ?



la corrélation est donné par le tableau : -0.6744 Pour la covariance il faut le multiplier par les écart types des deux variables TiO_2 et Al_2O_3 . $cov = -0.6744 \times 0.34 \times 0.89 = -0.2041$

4. Quelle(s) est/sont, selon vous, l'(es) étape(s) indispensable(s) à appliquer avant d'effectuer une ACP ?
il faut centrer et réduire les données et éliminer les valeurs aberrantes
5. Quelle relation permet de représenter les données par rapport à un axe factoriel ?

$$X_n V$$

6. Caractérisez (en pourcentage) la qualité de la représentation du nuage de points par rapport à la première composante de l'ACP ? Qu'en pensez vous ?
le premier axe représente $100 \frac{208}{480} = 42,5\%$ de l'information, ce qui peut paraître beaucoup en soit, surtout ramené au nombre de variables.
7. Quelle sont les moyennes et variances des variables de X et de X_n ?
il suffit de lire l'énoncé. $mean(X)$ donne la moyenne de X et la variance est donné par $std(X)^2$ (attention il ne faut pas oublier de mettre au carré. la moyenne de X_n est 0 et sa variance est 1
8. Quelle interprétation (relations entre les variables et les composantes principales) donnez vous à la représentation des variables dans le plan des deux premières composantes principales (donnée figure 1 ci dessous) ?
les variables proches du cercle unité sont bien représentées. les variables proche du centre (NiO , MnO et Cr_2O_3) sont mal représentées L'axe 1 oppose le groupe de variables Al_2O_3 , CaO , MgO avec K_2O , SiO_2 et TiO_2 . les variables latitude et élévation sont fortement corrélées et opposées à la longitude.
9. Ecrire en une ligne de code une instruction permettant de visualiser les données par rapport aux deux premiers axes de l'ACP
`plot(V(:,14),V(:,13),'o')`
10. Une donnée d'élévation est manquante (celle de la troisième observation). Suggérez un moyen de traiter cette valeur manquante.
On peut utiliser la moyenne des autres valeurs observées pour cette variables. Ou utiliser des variables très corrélées. Dans ce cas il faudra recentrer les variables. On peut aussi effectuer une régression linéaire avec des variables explicatives, et prédire par ce moyen la variable manquante.

Barème provisoire :

- question 1 : 2 points
- question 2 : 2 points
- question 3 : 5 points
- question 4 : 1 points
- question 5 : 1 points
- question 6 : 2 points
- question 7 : 1 points
- question 8 : 3 points
- question 9 : 1 points
- question 10 : 2 points

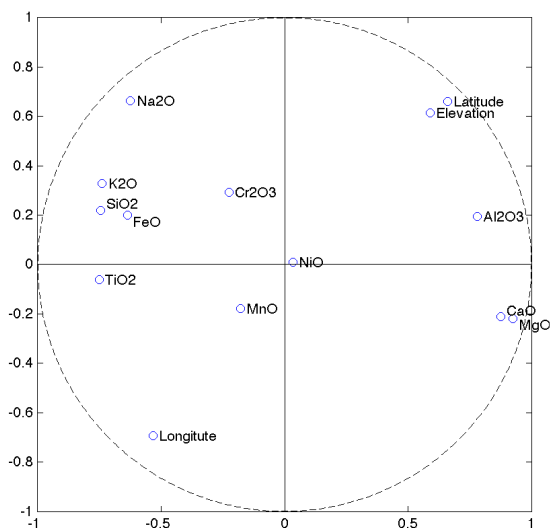


Figure 1 - Représentation des variables sur les deux premiers axes factoriels.

```

load petrology_data.xls;
X = petrology_data;
[n,p] = size(X)
n = 35
p = 14

Xn = (X - ones(n,1)*mean(X))./(ones(n,1)*std(X,1));
mean(X) =
50.8063    1.5526    14.8766    0.0306    9.6586    0.0160    0.4026    7.5380    11.3851    2.8691    0.1297    14.1439 -105.5052

std(X) =
1.2670    0.3402    0.8928    0.0331    1.9898    0.0224    1.4848    1.0305    1.0363    0.3776    0.0827    7.8353    4.3200    252.9691

cov(Xn) =
 1.0000    0.2725   -0.4306    0.1094    0.2042   -0.1294    0.0056   -0.8336   -0.8179    0.6123    0.7310   -0.4193    0.3555   -0.2786
 0.2725    1.0000   -0.6744    0.0940    0.8221    0.0878    0.2690   -0.6501   -0.6412    0.3886    0.3334   -0.4196    0.2308   -0.5535
-0.4306   -0.6744    1.0000   -0.0568   -0.4201   -0.0258   -0.1538   -0.7008   -0.4618   -0.2173   -0.4354   -0.6020   -0.5654    0.4964
 0.1094    0.0940   -0.0568    1.0000    0.1479    0.1914   -0.1047   -0.1996   -0.2085    0.4131    0.1587   -0.1017    0.0178   -0.0332
 0.2042    0.8221   -0.4201    0.1479    1.0000    0.0716    0.1625   -0.5757   -0.5688    0.4656    0.3772   -0.1763    0.0422   -0.2941
-0.1294    0.0878   -0.0258    0.1914    0.0716    1.0000    0.0486    0.0713    0.0423   -0.0378   -0.1053   -0.0142   -0.0688    0.0280
 0.0056    0.2690   -0.1538   -0.1047    0.1625    0.0486    1.0000   -0.1255   -0.0919   -0.0049    0.0892   -0.1731    0.1282   -0.0777
-0.8336   -0.6501    0.7008   -0.1996   -0.5757    0.0713   -0.1255    1.0000    0.8549   -0.7075   -0.8153    0.4720   -0.3631    0.3685
-0.8179   -0.6412    0.4618   -0.2085   -0.5688    0.0423   -0.0919    0.8549    1.0000   -0.6985   -0.6759    0.4597   -0.3239    0.4535
 0.6123    0.3886   -0.2173    0.4131    0.4656   -0.0378   -0.0049   -0.7075   -0.6985    1.0000    0.6954   -0.0401   -0.1028   -0.0237
 0.7310    0.3334   -0.4354    0.1587    0.3772   -0.1053    0.0892   -0.8153   -0.6759    0.6954    1.0000   -0.3224    0.3251   -0.1411
-0.4193   -0.4196    0.6020   -0.1017   -0.1763    0.0142   -0.1731    0.4720    0.4597   -0.0401   -0.3224    1.0000   -0.9289    0.8196
 0.3555    0.2308   -0.5654    0.0178    0.0422   -0.0688    0.1282   -0.3631   -0.3239   -0.1028    0.3251   -0.9289    1.0000   -0.6448
-0.2786   -0.5535    0.4964   -0.0332   -0.2941    0.0280   -0.0777    0.3685    0.4535   -0.0237   -0.1411    0.8196   -0.6448    1.0000

[U,d]=eig(Xn'*Xn)
diag(d)' =
0.3982    0.8273    1.6474    3.2782    4.5793    7.7280    12.1677    19.2695    25.6475    31.9167    41.4316    52.8413    80.1650    208.1023

sum(diag(d)) = 490

V = Xn*U
V =
-0.0926    0.0445   -0.0952    0.2118    0.2737   -0.3333   -0.5921    0.0289   -0.5152    0.1204   -0.4737    0.0047    0.9335    1.4804
 0.1359    0.1582   -0.2449   -0.4511   -0.3151   -0.3167    0.1989    0.1771   -0.1524   -0.2629   -1.0220    0.4654    0.7435    1.7968
 0.1206   -0.0272    0.1607   -0.0161    0.1724   -0.4102    0.1914   -0.3430   -0.7402   -0.2154   -0.7341   -0.3039    0.6754    0.5380
-0.0165    0.1717    0.1674    0.0748    0.2447    0.2844    0.4328    0.2653   -0.4645    0.6327    1.3659   -0.8506    1.3602    0.5611
-0.1007   -0.3005    0.1649    0.3696   -0.3670    0.2157   -0.1420   -1.0320    0.8819   -0.2987   -0.1667    0.4653    0.7633    2.7656
 0.0957    0.1009    0.2163    0.4877   -0.2230    0.3560    1.1415    0.6280   -0.9026   -1.3202   -1.1222   -2.5062    0.7065   -1.7059
-0.0502   -0.0323   -0.1348   -0.1879   -0.0125    0.2139   -0.2770   -0.8564    0.2501   -0.2236   -0.7221    0.7717    0.0712    3.6984
 0.0811    0.1076   -0.2529   -0.3380    0.1254   -0.2282    0.0184    0.3372   -0.1555   -0.2407   -0.9221    0.5356    1.0265    1.1525
-0.1707   -0.1107    0.2489   -0.7055   -0.0606   -0.1095    0.0711    0.4778   -0.1298   -0.3162   -0.9501    0.3626    0.1965    2.4691
-0.0985    0.2687    0.3275    0.0982    0.0085   -0.4491   -0.6185    0.1906   -0.4894   -0.1170   -1.1915    0.8672    1.0068   -1.3723
 0.0263   -0.1373    0.3356    0.0144    0.3286   -0.5242   -0.4899    1.2835    0.9226    0.1922    1.4697   -0.1389    1.2719   -1.0105
 0.0317    0.3457    0.0041    0.6127    0.2258    0.1470   -0.8695   -1.1291    1.1003    0.2117    0.8046    0.8844    0.4601    1.5427
 0.0319    0.1354   -0.0915   -0.1727   -0.5816    0.3927   -0.1795   -0.1881    0.2056    0.3025    0.6112    0.4705   -0.1361    1.0684
-0.0896    0.1377   -0.0471   -0.2525    0.0252   -0.2441   -0.7544    1.2032    0.9190    0.2803    0.8707    0.3860    0.0741    0.7001
-0.1532   -0.0187   -0.1201    0.3021   -0.3668    0.4081   -0.0920   -0.1996   -0.8304    0.1119    0.9874   -1.2152   -0.5820   -1.9163
-0.0002   -0.0044    0.0203   -0.0750    0.0490   -0.0465    0.2512   -0.3432    1.7838    3.9363   -2.6936   -2.4959   -1.4678   -2.3764
 0.0842   -0.0286   -0.3885    0.0662   -0.1333    0.1473    0.2953    1.3518    2.1809   -0.5501    2.2748   -0.9456   -0.1228   -1.7252
-0.0390    0.2278   -0.1036   -0.2521   -0.3500   -0.2812   -0.6179   -0.5238   -2.0205    1.3855    2.3352   -1.5455   -0.5619   -1.0886
-0.3043   -0.1518    0.1091   -0.2407   -0.0253   -0.7931   -0.0574   -0.1548    0.2158   -1.2286    0.4984   -1.3402   -2.2339   -1.5489
-0.0424   -0.1336   -0.1211   -0.1989    0.0321    0.0461    0.6330   -0.7387   -0.2189    0.7429    0.3159    3.2194    3.7884   -9.3509
 0.0365   -0.0400   -0.0173   -0.1849   -0.2027    0.1682    0.2250    0.0161    0.9970   -0.3597    0.1970    0.2317    0.6929    2.5788
 0.1755   -0.3764    0.1424   -0.0164   -0.4407   -0.0375   -0.0244   -1.2773   -0.8363    0.9121    1.6315   -0.9999    0.7784    2.5385
 0.0597   -0.1128   -0.0017    0.0921   -0.3666   -0.7665   -0.1660    0.0392   -0.0923   -0.3270   -1.1066    0.3129    1.1631    0.9395
-0.0241   -0.2172   -0.4125    0.2995    1.2949    0.0928   -0.4134    0.5044   -1.5759    0.3743   -0.5374   -0.7667    0.2209    0.9918
-0.0645    0.1259    0.3510   -0.2055    0.1555    0.7101    0.9149    0.0217   -0.2762    0.0932    0.6135   -0.4335    0.7067    1.1616
-0.1011   -0.0267    0.1958    0.1458    0.4210    0.8078    0.7758    0.2827    0.6993   -0.4565    0.0509    0.0055    0.8775    0.8788
 0.0069    0.1878   -0.0460   -0.0103   -0.1597    0.1131    0.5312    0.0371   -0.2621   -0.4700   -1.1427    0.1458    0.5417    1.3293
-0.0034   -0.0622   -0.1596    0.7375   -0.2476   -0.4637   -0.4679   -0.0060    0.5204   -0.3298   -0.6289   -0.0287    1.6847    0.5473
 0.1070   -0.1220   -0.2610   -0.4211    0.3168    0.1749    0.3438    0.4865   -0.3055    0.2526    0.5347   -0.2920    0.3491    2.3608
 0.1240   -0.1007    0.1926   -0.0364   -0.1685    0.8588   -1.1834    0.7891   -0.4603   -0.2744   -0.7816    0.8251   -2.0386   -2.4736
 0.1128   -0.1017    0.1897   -0.0385   -0.1622    0.8542   -1.1763    0.7864   -0.4610   -0.2768   -0.7840    0.8204   -2.0294   -2.4690
-0.0067   -0.0214    0.0248    0.5264   -0.1518   -0.5910    1.3947    0.9928   -0.7674    1.4897    0.6454    3.6674   -4.3373    2.2167
-0.0707   -0.0124   -0.4792    0.0932   -0.4500   -0.0842    0.3493   -0.0781   -0.1807   -1.5422   -1.0047   -1.0372   -1.5519   -3.1036
-0.0366    0.0646   -0.1754   -0.1871    0.5249    0.6980    0.1427   -1.8434    0.5385   -0.7685    0.1684    1.0052   -2.6889   -0.2352
 0.2352    0.0620    0.3012   -0.1413    0.5863   -1.0100    0.2105   -1.1860    0.6216   -1.4598    0.6089   -0.5469   -2.3422   -2.9399

U = U*sqrt(d)/sqrt(length(Xn))

```

Sample_ID	SiO2	TiO2	Al2O3	Cr2O3	FeO	NiO	MnO	MgO	CaO	Na2O	K2O	Latitude	Longitude	Elevation
ALV0906-R007	50.18	1.42	15.53	0.01	9.3	0.02	0.06	7.99	11.93	2.83	0.17	20.87	-109.07	-2600
ALV0907-R001	50.67	1.36	15.34	0	8.79	0	0.08	7.9	12.42	2.95	0.09	20.87	-109.08	-2585
ALV0911-R004	50.78	1.57	15.01	0	9.7	0.02	0.23	7.58	11.41	2.93	0.13	20.85	-109.07	-2600
ALV0912-R004D	50.94	1.53	14.82	0.07	9.71	0.05	0.18	7.69	11.39	2.93	0.12	20.87	-109.07	-2600
ALV0916-R002	49.67	1.19	16.57	0.03	8.26	0	0.14	8.77	11.39	2.83	0.05	20.83	-109.1	-2615
ALV0920-R012	50.73	2.38	13.27	0.01	12.88	0.02	0.29	6.8	10.33	2.95	0.08	20.95	-109.3	-2800
ALV0923-R002	49.91	1.05	16.65	0	7.89	0	0.03	8.97	12.45	2.53	0.04	20.82	-109.12	-2602
ALV0923-R005	50.93	1.4	15.22	0.01	9.26	0	0	7.53	12.16	2.93	0.14	20.82	-109.12	-2602
ALV0923-R006	50.08	1.13	15.09	0	9.12	0	0.09	8.27	12.65	2.7	0.07	20.82	-109.12	-2602
CHRSR01-002-102	51.74	1.58	14.55	0	10.47	0	0.16	7.1	10.82	3.22	0.26	15.817	-105.4	-2750
CHRSR01-002-201	50.31	1.52	14.29	0.1	10.53	0.02	0.09	7.06	11.51	3.26	0.23	15.817	-105.4	-2750
CHRSR01-002RT-101	50.47	1.29	16.72	0.06	8.32	0.01	0.16	8.78	11.32	2.83	0.15	15.817	-105.4	-2750
CHRSR01-003-102	50.75	1.32	15.47	0.04	8.85	0.02	0.16	8.13	11.73	2.86	0.08	12.817	-103.97	-2650
CHRSR01-003-103	49.98	1.32	14.84	0.07	9.46	0.01	0.22	7.91	12.53	2.91	0.17	12.817	-103.97	-2650
CHRSR01-004-101	51.02	1.95	14.22	0.04	11.28	0.05	0.25	7.04	10.4	2.9	0.13	6.733	-102.6	-2900
CHRSR01-004-102	50.85	2.04	14.17	0.01	11.34	0.02	0.05	6.84	10.91	2.86	0.17	6.733	-102.6	-2900
CHRSR01-004-202	50.07	1.93	14.02	0.14	10.98	0.01	0.22	6.93	11.47	3.1	0.11	6.733	-102.6	-2900
CHRSR01-004-301	50.99	1.7	14.56	0.04	10.55	0.1	0.25	7.49	11.32	3	0.15	6.733	-102.6	-2900
CHRSR01-005-101	49.72	1.94	13.9	0.03	11.41	0.02	0.1	7.54	11.36	2.8	0.09	2.28	-102.78	-3350
CHRSR01-006-401	57.43	1.9	13.51	0.08	11.54	0	0.09	2.83	6.83	4.5	0.47	2.3	-101.75	-3000
CYA1978-002-003	49.88	1.24	15.83	0.05	8.62	0	0.02	8.48	12.25	2.82	0.04	20.88	-109.03	-2590
CYA1978-006-010	49.74	1.24	16.25	0.04	8.51	0.07	0.19	8.5	11.61	2.91	0.03	20.92	-109.07	-2655
CYA1978-007-012	50.26	1.48	15.26	0	9.15	0	0.13	7.67	11.69	3.17	0.15	20.92	-109.05	-2645
CYA1978-010-017	50.3	1.6	14.89	0	9.84	0.04	0.24	7.23	12.04	2.33	0.19	20.9	-109.03	-2660
CYA1978-010-019B	51.15	1.41	14.84	0.05	9.48	0.03	0.14	7.96	11.4	2.72	0.05	20.9	-109.05	-2645
CYA1978-012-036	50.92	1.48	14.97	0.06	9.62	0	0.15	7.72	11.19	2.69	0.08	20.9	-109.07	-2645
CYA1978-018-066B	50.85	1.52	15.06	0	9.11	0	0.14	7.92	11.72	2.8	0.08	20.88	-109.08	-2625
CYA1978-018-068	49.99	1.66	15.58	0.03	9.64	0	0.2	7.68	11.16	3.2	0.18	20.87	-109.08	-2585
CYA1978-019-072	50.15	1.28	15.24	0.04	8.64	0.03	0.02	7.94	12.69	2.59	0.06	20.85	-109.08	-2585
SON0012-155-001	51.44	1.65	13.8	0	11.58	0	0.22	6.72	10.8	2.59	0.21	2.7983	-95.049	-2914
SON0012-155-001	51.44	1.65	13.8	0	11.58	0	0.22	6.72	10.8	2.59	0.21	2.7983	-95.135	-2914
SON0012-165-001	51.09	0.8	14.87	0	0.54	0.01	0.12	8.58	12.91	2.04	0.07	2.3052	-96.017	-2900
SON0012-190-001	51.1	2.32	13.62	0	12.17	0	0.21	6.44	10.44	2.93	0.11	2.804	-102.23	-3180
SON0012-206-D001	51.65	1.47	15.72	0.02	8.92	0	0.13	7.97	10.88	2.3	0.05	2.182	-100.67	-3307
SON0012-207-001	51.04	2.02	14.2	0.04	11.01	0.01	0.11	7.15	10.57	2.92	0.13	1.83	-100.92	-3700
somme	1778.2	54.34	521.68	1.07	338.05	0.56	14.09	263.83	398.48	100.42	4.54	495.03	-3692.7	-95056
somme des carrés	90401	88.417	7802.6	0.0711	3403.7	0.0266	82.833	2025.9	4574.3	293.11	0.828	9150.4	390250	3E+08

FIGURE 2 – Données brutes utilisées pour l'examen. La première colonne donne le nom permettant de repérer la mesure, et les deux dernières lignes représentent, pour chaque variable, la somme $\sum_{i=1}^n x_i$ et la somme des carrés $\sum_{i=1}^n x_i^2$ des mesures.

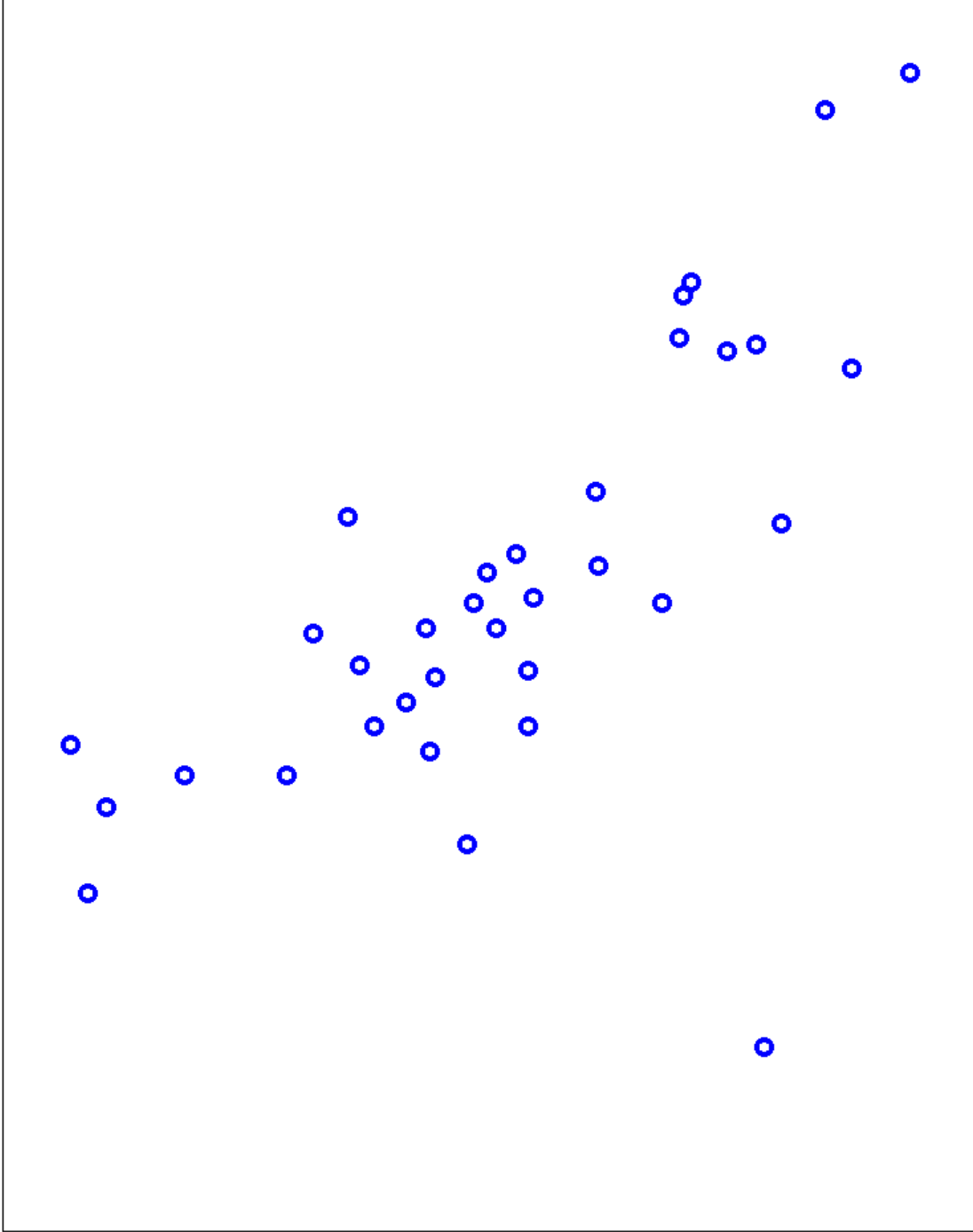


FIGURE 3 – Tracez le sac à moustache de ces deux variables (TiO_2 et Al_2O_3).