

Analyse de données
TP n°1
Analyse ACP



Introduction

Nous disposons de 14 variables représentant les notes obtenues dans différentes matières étudiées par 27 élèves de 6ème d'un collège.

Voici une partie de ces données :

	ORTH	GRAM	EXPR	RECI	MATH	ANGL	HIST	BIOL	EDMU	ARTS	TECH	EPS	GEO	EXPO
EL01	13.0	10.0	2.0	4.0	9.0	9.0	8.0	7.0	7.5	1.5	14	10.0	10.50	13.0
EL02	6.5	8.0	8.5	14.0	13.0	7.0	11.0	8.5	16.0	4.0	18	18.0	16.00	15.0
EL03	14.0	6.5	8.0	5.0	11.0	8.0	9.5	8.0	18.5	9.5	14	16.5	14.00	13.0
EL04	13.0	7.5	9.0	5.0	10.0	10.5	10.0	16.0	16.0	11.5	0	11.5	15.00	18.0
EL05	15.0	7.5	10.0	14.0	12.0	11.0	9.0	11.0	16.5	13.5	16	13.0	15.51	17.0
EL06	5.0	8.0	5.5	6.5	16.0	12.0	9.0	7.0	13.5	5.0	16	12.5	13.00	17.5
EL07	12.0	6.5	9.0	16.0	18.0	13.5	9.0	10.0	15.0	11.0	16	13.5	14.00	18.0
EL08	8.5	2.5	9.0	13.0	12.0	9.5	12.0	13.5	16.5	8.0	13	12.0	14.00	17.0
EL09	15.5	7.5	12.5	16.0	15.0	13.0	12.0	13.5	17.0	14.0	15	16.0	15.00	17.0
EL10	20.0	14.5	16.5	10.0	18.0	16.5	15.0	10.5	18.0	13.5	12	14.5	13.00	15.0
EL11	6.0	4.0	11.0	9.5	13.0	12.0	11.0	7.5	17.5	12.0	18	13.5	13.00	17.0
EL12	0.0	5.5	6.0	6.0	9.0	3.0	7.0	2.0	12.0	1.5	15	12.5	14.00	0.0
EL13	15.5	12.0	12.0	8.0	17.0	17.0	11.0	4.5	16.0	9.0	13	5.0	12.50	16.5
EL14	15.0	6.5	11.5	10.0	13.0	12.5	7.0	11.0	14.0	7.0	14	11.0	15.00	13.0
EL15	11.0	9.0	7.5	11.0	12.5	11.0	9.0	11.5	18.0	8.5	17	15.0	15.00	15.0
EL16	6.5	9.0	10.0	10.0	10.0	6.5	7.0	12.5	18.0	14.5	16	13.5	13.00	18.0
EL17	7.0	4.5	6.0	6.0	15.0	13.5	11.0	7.0	7.0	6.5	11	10.5	13.50	16.0

Showing 1 to 18 of 27 entries, 14 total columns

Nous allons analyser ces données en effectuant une ACP à l'aide de R.

1 - Statistiques sommaires

Au moment où l'enquête a été réalisée, l'éducation musicale était celle obtenant en moyenne les meilleurs résultats, nous pouvons voir cela avec la moyenne qui est ici de 15,200. C'est la plus élevée dans les données dont nous disposons. A l'inverse, la grammaire obtient les moins bons résultats avec une moyenne faible de 7,556. Les matières les mieux réussies sont donc l'éducation musicale suivie d'exposé, légèrement plus loin la géographie, le sport, la technologie ainsi que les mathématiques, suivie un peu plus loin par l'anglais, la récitation et l'orthographe, puis sous la barre des 10 de moyenne nous retrouvons la biologie, l'histoire et l'art plastique, enfin plus loin l'expression écrite et la grammaire.

Les critères les plus importants sont ceux variant le moins, ici l'histoire et les mathématiques présentent les plus petites variations (écart-type de 2,13 et 2,82 respectivement). La technologie et l'orthographe présentent des dispersions plus importantes (écart-type de 4,86 et 4,82 respectivement), suivies de près par l'exposé (écart-type de 4,50). En effet, on observe une dispersion maximale (en imaginant que les notes soient comprises entre 0 et 20) en orthographe de 20.

2 - Corrélations et valeurs-tests

Tableau des corrélations

	ORTH	GRAM	EXPR	RECI	MATH	ANGL	HIST	BIOL	EDMU	ARTS	TECH	EPS	GEO	EXPO
ORTH	1.00	0.53	0.51	-0.07	0.38	0.65	0.54	0.35	0.18	0.31	-0.29	-0.04	0.16	0.14
GRAM	0.53	1.00	0.35	-0.10	0.34	0.40	0.30	-0.02	0.21	0.15	-0.17	-0.09	-0.07	0.05
EXPR	0.51	0.35	1.00	0.24	0.59	0.54	0.59	0.40	0.63	0.72	0.00	0.03	0.39	0.31
RECI	-0.07	-0.10	0.24	1.00	0.18	-0.04	-0.06	0.13	0.41	0.32	0.34	0.24	-0.11	0.33
MATH	0.38	0.34	0.59	0.18	1.00	0.70	0.64	0.12	0.28	0.37	0.21	-0.06	0.33	0.41
ANGL	0.65	0.40	0.54	-0.04	0.70	1.00	0.59	0.17	0.00	0.29	-0.23	-0.27	0.13	0.25
HIST	0.54	0.30	0.59	-0.06	0.64	0.59	1.00	0.26	0.22	0.26	-0.01	0.07	0.33	0.26
BIOL	0.35	-0.02	0.40	0.13	0.12	0.17	0.26	1.00	0.46	0.58	-0.17	0.10	0.46	0.56
EDMU	0.18	0.21	0.63	0.41	0.28	0.00	0.22	0.46	1.00	0.70	0.18	0.29	0.32	0.43
ARTS	0.31	0.15	0.72	0.32	0.37	0.29	0.26	0.58	0.70	1.00	-0.05	0.07	0.24	0.57
TECH	-0.29	-0.17	0.00	0.34	0.21	-0.23	-0.01	-0.17	0.18	-0.05	1.00	0.07	0.15	0.29
EPS	-0.04	-0.09	0.03	0.24	-0.06	-0.27	0.07	0.10	0.29	0.07	0.07	1.00	0.35	-0.22
GEO	0.16	-0.07	0.39	-0.11	0.33	0.13	0.33	0.46	0.32	0.24	0.15	0.35	1.00	0.09
EXPO	0.14	0.05	0.31	0.33	0.41	0.25	0.26	0.56	0.43	0.57	0.29	-0.22	0.09	1.00

p-value

	ORTH	GRAM	EXPR	RECI	MATH	ANGL	HIST	BIOL	EDMU	ARTS	TECH	EPS	GEO	EXPO
ORTH		0.0043	0.0063	0.7442	0.0523	0.0002	0.0033	0.0732	0.3599	0.1188	0.1371	0.8445	0.4182	0.4777
GRAM	0.0043		0.0700	0.6056	0.0816	0.0368	0.1351	0.9195	0.3046	0.4699	0.4072	0.6384	0.7221	0.7937
EXPR	0.0063	0.0700		0.2356	0.0012	0.0038	0.0012	0.0378	0.0004	0.0000	0.9857	0.8961	0.0415	0.1099
RECI	0.7442	0.6056	0.2356		0.3675	0.8375	0.7753	0.5032	0.0329	0.0982	0.0819	0.2350	0.5819	0.0929
MATH	0.0523	0.0816	0.0012	0.3675		0.0000	0.0003	0.5562	0.1643	0.0566	0.2933	0.7653	0.0968	0.0332
ANGL	0.0002	0.0368	0.0038	0.8375	0.0000		0.0011	0.3922	0.9837	0.1468	0.2568	0.1676	0.5041	0.2018
HIST	0.0033	0.1351	0.0012	0.7753	0.0003	0.0011		0.1915	0.2673	0.1875	0.9418	0.7387	0.0939	0.1931
BIOL	0.0732	0.9195	0.0378	0.5032	0.5562	0.3922	0.1915		0.0169	0.0014	0.4107	0.6266	0.0151	0.0023
EDMU	0.3599	0.3046	0.0004	0.0329	0.1643	0.9837	0.2673	0.0169		0.0000	0.3798	0.1389	0.1014	0.0241
ARTS	0.1188	0.4699	0.0000	0.0982	0.0566	0.1468	0.1875	0.0014	0.0000		0.7893	0.7454	0.2287	0.0019
TECH	0.1371	0.4072	0.9857	0.0819	0.2933	0.2568	0.9418	0.4107	0.3798	0.7893		0.7298	0.4414	0.1441
EPS	0.8445	0.6384	0.8961	0.2350	0.7653	0.1676	0.7387	0.6266	0.1389	0.7454	0.7298		0.0761	0.2737
GEO	0.4182	0.7221	0.0415	0.5819	0.0968	0.5041	0.0939	0.0151	0.1014	0.2287	0.4414	0.0761		0.6395
EXPO	0.4777	0.7937	0.1099	0.0929	0.0332	0.2018	0.1931	0.0023	0.0241	0.0019	0.1441	0.2737	0.6395	

Question 3 :

En prenant une marge d'erreur de 5%, nous pouvons relever les coefficients de corrélation les plus importants grâce à la matrice des p-value. En effet, nous devons prendre en compte les corrélations correspondant à des p-values inférieures à 0.05.

Ainsi, nous retenons les p-value entre les matières suivantes :

- l'expression écrite et l'art plastique (p-value de 0.0000)
- l'art plastique et l'éducation musicale (p-value de 0.0000)
- les mathématiques et l'anglais (p-value de 0.0000)
- l'orthographe et l'anglais (p-value de 0.0002)
- les mathématiques et l'histoire (p-value de 0.0003)
- l'expression écrite et l'éducation musicale (p-value de 0.0004)
- l'expression écrite et l'histoire (p-value de 0.0012)
- les mathématiques et l'expression écrite (p-value de 0.0012)
- l'anglais et l'histoire (p-value de 0.0011)

En conséquence, les coefficients de corrélation les plus importantes sont entre :

- l'expression écrite et l'art plastique (coefficient de corrélation de 0,72)

- l'art plastique et l'éducation musicale (coefficient de corrélation de 0,70)
- les mathématiques et l'anglais (coefficient de corrélation de 0,70)
- l'orthographe et l'anglais (coefficient de corrélation de 0,65)
- les mathématiques et l'histoire (coefficient de corrélation de 0,64)
- l'expression écrite et l'éducation musicale (coefficient de corrélation de 0,63)
- l'expression écrite et l'histoire (coefficient de corrélation de 0,59)
- les mathématiques et l'expression écrite (coefficient de corrélation de 0,59)
- l'anglais et l'histoire (coefficient de corrélation de 0,59)

Cela implique que si un élève a de bons résultats dans l'une des matières, ils semblent avoir de bons résultats dans l'autre matière. De même, si un élève a de mauvais résultats dans l'une des matières, ils semblent avoir de mauvais résultats dans l'autre matière. Plus le coefficient de corrélation est proche de 1, plus cette relation est importante.

De ce fait, nous pouvons affirmer que l'expression écrite et l'art plastique, l'art plastique et l'éducation musicale ainsi que les mathématiques et l'anglais ont une corrélation forte, c'est-à-dire que nous pouvons dire que

- si un élève a de bons résultats en expression écrite, il semble en avoir de bons en art plastique et inversement s'il a de bons résultats en art plastique il semble en avoir de bons en expression écrite. De même, si un élève a de mauvais résultats en expression écrite, il semble en avoir de mauvais en art plastique et inversement s'il a de mauvais résultats en art plastique il semble en avoir de mauvais en expression écrite
- si un élève a de bons résultats en art plastique, il semble en avoir de bons en éducation musicale et inversement s'il a de bons résultats en éducation musicale il semble en avoir de bons en art plastique. De même, si un élève a de mauvais résultats en art plastique, il semble en avoir de mauvais en éducation musicale et inversement s'il a de mauvais résultats en éducation musicale il semble en avoir de mauvais en art plastique
- si un élève a de bons résultats en art plastique, il semble en avoir de bons en mathématiques et inversement s'il a de bons résultats en mathématiques il semble en avoir de bons en anglais. De même, si un élève a de mauvais résultats en anglais, il semble en avoir de mauvais en mathématiques et inversement s'il a de mauvais résultats en mathématiques il semble en avoir de mauvais en anglais

Les autres matières deux à deux citées ont une corrélation légèrement moins haute, nous pouvons alors dire les résultats des matières sont dépendantes deux à deux, mais que leurs dépendances est moins forte que pour les trois couples précédemment cités.

3 - Variances des axes (valeurs propres)

Variance du nuage (trace de la matrice) : 14

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.77570045	34.1121461	34.11215
comp 2	2.39238450	17.0884607	51.20061
comp 3	1.49934217	10.7095870	61.91019
comp 4	1.34084173	9.5774409	71.48763
comp 5	1.12670192	8.0478709	79.53551
comp 6	0.72038480	5.1456057	84.68111
comp 7	0.59919071	4.2799337	88.96104
comp 8	0.42748490	3.0534636	92.01451
comp 9	0.39040787	2.7886277	94.80314
comp 10	0.25931678	1.8522627	96.65540
comp 11	0.18210022	1.3007158	97.95611
comp 12	0.12552548	0.8966105	98.85273
comp 13	0.09728799	0.6949142	99.54764
comp 14	0.06333049	0.4523606	100.00000

Question 4 :

La somme des valeurs propres et donc la trace de la matrice vaut 14. Le nombre de variables actives est égal à la trace de la matrice, c'est-à-dire 14.

Selon le critère de Kaiser, on ne retient que les 5 premières composantes puisque ce sont les seuls à avoir une valeur propre supérieure à 1.

En respectant ce critère, la somme des pourcentages des variances avoisine bien les 80% mais ne les dépasse pas. Afin d'obtenir au moins 80% d'information, on sélectionne les 6 premiers axes.

4 - Variables actives

Contribution

\$contrib	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
ORTH	8.857178434	10.147700841	2.011115774	0.819262759	1.9692093	1.08436118
GRAM	3.647913177	9.862949216	0.343291721	0.008243911	22.7076131	27.51108242
EXPR	16.001163011	0.009919583	0.146062754	0.142529916	1.9664615	0.39444515
RECI	1.280583878	14.800939040	9.201063052	0.030321946	13.6369489	27.46853062
MATH	11.258878062	1.249038115	6.282449446	13.644601594	0.6409718	0.82326303
ANGL	9.147755175	14.164029878	2.131100334	0.304427329	0.8831389	9.62778805
HIST	10.205943479	3.985966555	0.352172894	8.812289759	1.4405132	2.54906139
BIOL	7.493655757	3.591472838	7.035864125	14.725203802	10.7481175	0.45216960
EDMU	8.707419780	11.968896845	0.632205474	0.643621144	7.9649991	10.04618126
ARTS	12.033404324	4.583971428	0.002441541	9.735102286	0.5733860	0.08672036
TECH	0.001890561	11.091355234	14.496417584	23.845091334	1.3943743	8.96598313
EPS	0.039310859	7.801449997	24.155785266	10.697922436	11.1753449	6.70986985
GEO	4.189905868	2.037057095	19.492996119	10.971312784	15.2269386	3.77273356
EXPO	7.134997636	4.705253336	13.717033917	5.620069002	9.6719828	0.50781041

cos²

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
ORTH	4.229923e-01	0.2427720217	3.015351e-02	0.0109850169	0.022187119	0.0078115731
GRAM	1.742134e-01	0.2359596680	5.147118e-03	0.0001105378	0.255847113	0.1981856560
EXPR	7.641676e-01	0.0002373146	2.189980e-03	0.0019111006	0.022156159	0.0028415229
RECI	6.115685e-02	0.3540953710	1.379554e-01	0.0004065693	0.153647766	0.1978791194
MATH	5.376903e-01	0.0298817942	9.419541e-02	0.1829525118	0.007221842	0.0059306617
ANGL	4.368694e-01	0.3388580549	3.195249e-02	0.0040818887	0.009950343	0.0693571217
HIST	4.874053e-01	0.0953596459	5.280277e-03	0.1181588583	0.016230290	0.0183630508
BIOL	3.578746e-01	0.0859218394	1.054917e-01	0.1974416771	0.121099246	0.0032573611
EDMU	4.158403e-01	0.2863420326	9.478923e-03	0.0086299409	0.089741798	0.0723711628
ARTS	5.746793e-01	0.1096662218	3.660706e-05	0.1305323137	0.006460351	0.0006247203
TECH	9.028755e-05	0.2653478631	2.173509e-01	0.3197249347	0.015710442	0.0645895797
EPS	1.877369e-03	0.1866406803	3.621779e-01	0.1434422081	0.125912825	0.0483368825
GEO	2.000974e-01	0.0487342381	2.922667e-01	0.1471079399	0.171562210	0.0271781991
EXPO	3.407461e-01	0.1125677514	2.056653e-01	0.0753562303	0.108974416	0.0036581890

Question 5 :

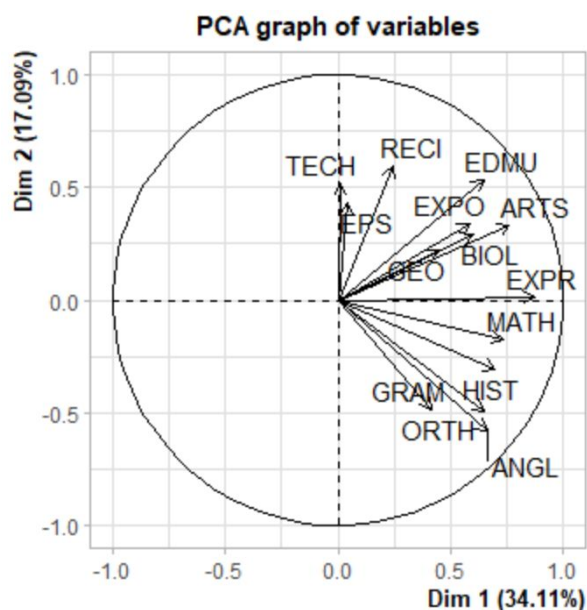
La contribution moyenne théorique est d'environ 7,14. On retient donc pour chaque axe les contributions supérieures à cette contribution moyenne théorique et ceux qui ont un \cos^2 élevé.

De ce fait, on voit que pour l'axe 1, l'expression écrite est très bien représentée (confirmé par la valeur du \cos^2), suivie par l'art plastique, les mathématiques, l'histoire, l'anglais, l'orthographe, l'éducation musicale et la biologie ont une contribution significative. On note que l'art plastique et les mathématiques ont un \cos^2 élevé.

Pour l'axe 2, la récitation et les mathématiques ont les plus fortes contributions, suivie par l'éducation musicale, la technologie, l'orthographe, la grammaire et l'eps qui ont des bonnes contributions.

Pour l'axe 3, la géographie est très bien représentée, suivie par la géographie puis la technologie, l'exposé et la récitation qui ont de bonnes contributions.

Axe 1



Les matières contribuant le plus à l'axe 1 sont l'expression écrite et l'art plastique suivie de près par les mathématiques.

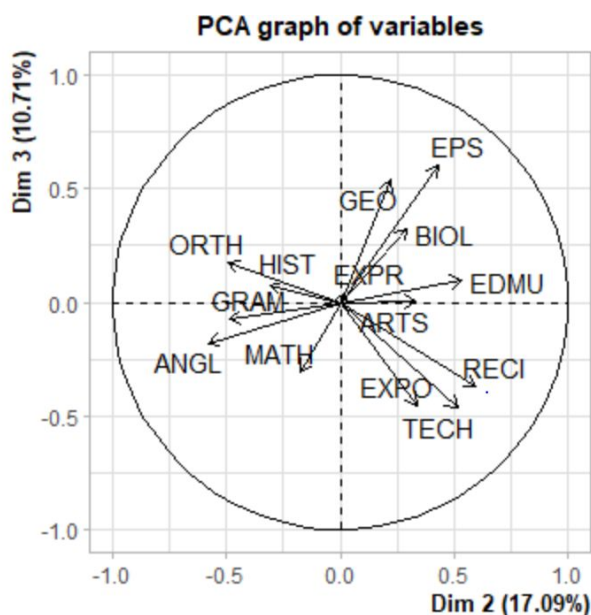
Les matières contribuant un peu moins à l'axe 1 mais d'une façon significative sont l'éducation musicale, l'histoire, l'orthographe, l'exposé et la biologie.

On voit aussi que dans ce plan, l'éducation musicale, l'art plastique, l'expression écrite et l'anglais sont les matières les mieux représentées, suivies de près par la récitation, l'exposé, la biologie, l'histoire et l'orthographe. Toutes les matières ont des coordonnées positives sur l'axe 1.

Questions 6 et 7 :

De ce cercle, on remarque que sur la récitation, l'éducation musicale et la technologie sont opposées à l'anglais, l'orthographe et la grammaire. Cela signifie qu'elles seraient fortement négativement corrélées entre elles.

Nous pouvons en déduire que parmi les élèves de cette classe, ceux qui ont des bonnes notes en récitation, éducation musicale et/ou technologies tendent à en avoir de moins bonnes en anglais, orthographe et grammaire.



Question 5 suite :

Axe 2

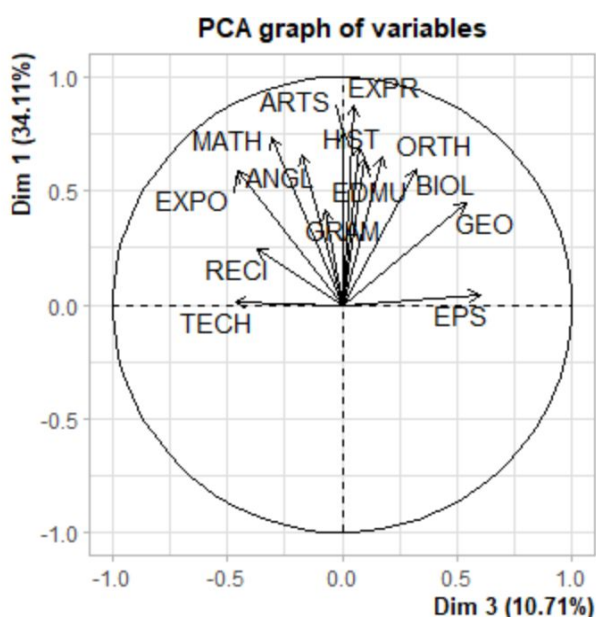
Les matières contribuant positivement à l'axe 2 sont la récitation, l'éducation musicale, la technologie et la grammaire. Celles contribuant négativement à l'axe 2 sont l'orthographe, l'anglais et la grammaire.

Dans ce plan, l'eps, la technologie et la récitation sont les mieux représentées.

Sur l'axe 2, l'orthographe, l'anglais et la grammaire sont opposés à la récitation, l'éducation musicale, la technologie et la grammaire.

Axe 3

La technologie, la récitation et l'exposé sont les matières contribuant négativement à l'axe 3 tandis que l'eps et la géographie contribuent positivement. Dans ce plan, l'expression écrite et l'art plastique sont les mieux représentées. Sur l'axe 3, la technologie, la récitation et l'exposé sont opposés à l'eps et à la géographie.



5 - Individus actifs

Voici les contributions d'une partie des élèves par rapport aux axes 1 à 6

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6
EL01	9.328905241	9.59126708	0.616515806	0.17696256	1.645531171	4.14443581
EL02	0.125275069	3.96118861	0.258445075	12.52737590	1.021099027	0.05991010
EL03	0.090900704	0.16283004	5.087265764	0.32642387	0.428252390	4.10865707
EL04	0.565357272	1.30848854	9.031255314	20.89056129	4.197449434	0.21259158
EL05	1.487781989	1.64284957	0.148851085	0.30166113	0.168538244	0.12252497
EL06	0.707432497	0.57129277	3.707371169	3.71293267	3.200272755	1.64880738
EL07	1.853709107	0.76715711	5.398964876	1.60627596	0.003810331	8.65944717
EL08	0.112046826	2.41932985	0.013041174	0.11265310	5.395428364	7.15016355
EL09	7.407644214	1.52394402	0.009609208	0.49971364	0.855791960	8.43800670
EL10	18.588377979	7.22537004	0.003702659	2.66022614	9.069367389	0.16940238
EL11	0.306502245	1.89876424	1.487091590	1.53197702	0.957152954	0.06602138
EL12	20.985890664	0.00230761	2.387411364	7.84839701	0.460623717	6.98776437

La contribution moyenne théorique est de 3.70%.

Question 8 :

Les élèves 10 et 12 sont les mieux représentés sur l'axe 1 avec des contributions respectives de 18.59% et 20.99%.

Sur l'axe 2, l'élève 10 est bien représenté puisque sa contribution est supérieure à la contribution moyenne théorique. En revanche, l'élève 12 n'est pas bien représenté sur l'axe 2.

Les individus expliquant un axe sont ceux dont la contribution est supérieure à la contribution moyenne, ici 3,70%.

Question 9 :

Axe 1

Les élèves 1, 9, 10, 12, 21, 23 et 24 sont ceux expliquant l'axe 1 avec des contributions respectives de 9.33, 7.41, 18.59, 20.99, 12.58, 4.62 et 11.22.

Au vue de leurs coordonnées, les élèves 12, 24, 1 et 23 contribuent négativement à l'axe 1 tandis que les élèves 10, 21 et 9 contribuent positivement.

Axe 2

Les élèves 1, 2, 10, 13, 16, 17, 19, 21, 22, 23 et 26 sont ceux expliquant l'axe 2 avec des contributions respectives de 9.59, 3.96, 7.22, 14.07, 6.63, 5.86, 5.62, 5.24, 5.84, 16.67 et 5.00.

Au vue de leurs coordonnées, les élèves 23, 13, 1, 10, 17, 21 contribuent négativement à l'axe 2 tandis que les élèves 16, 22, 26, 17 et 2 contribuent positivement.

Axe 3

Les élèves 3, 4, 6, 7, 13, 18, 20, 23, 24 et 25 sont ceux expliquant l'axe 3 avec des contributions respectives de 5.09, 9.03, 3.70, 5.40, 14.45, 3.92, 4.67, 22.14, 15.43, 5.46.

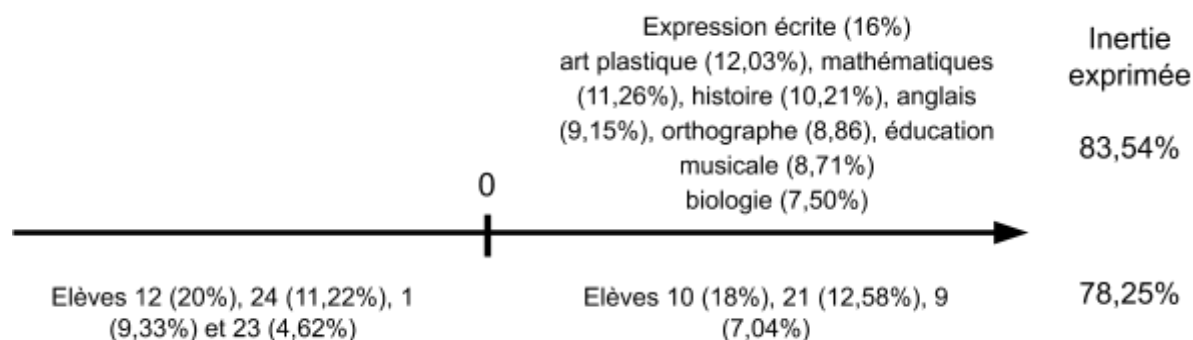
Au vue de leurs coordonnées, les élèves 24, 13, 7 et 6 contribuent négativement à l'axe 3 tandis que les élèves 23, 4, 25, 3, 20 et 18 contribuent positivement.

Conclusion

Question 10 :

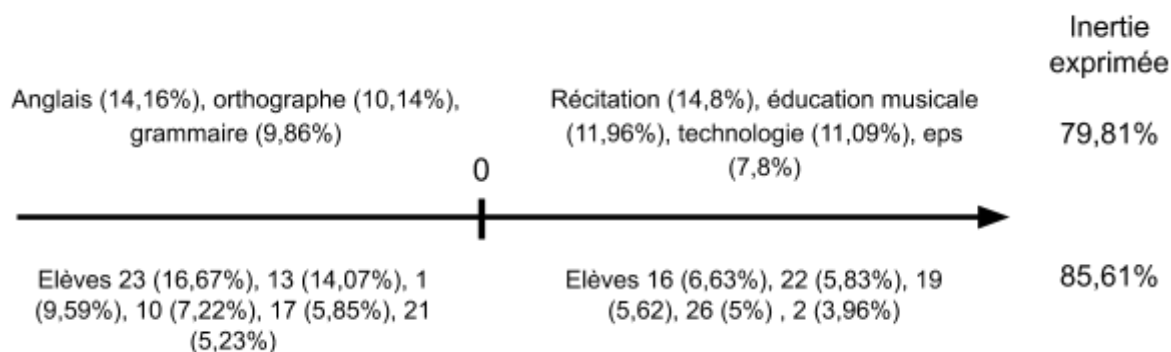
Pour conclure, nous avons vu que nous nous intéressons aux 6 premiers axes pour obtenir au moins 80% de l'information totale. Nous nous sommes penchés plus particulièrement sur les trois premiers axes, nous avons donc cette représentation de tendance.

- **Axe 1 (34,11%)**



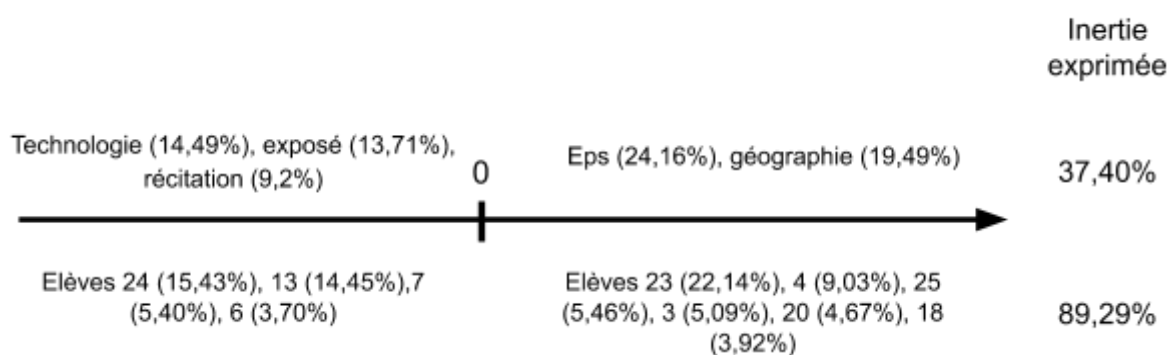
On déduit que les élèves 10, 21 et 9 ont des meilleures notes en expression écrite, art plastique, mathématiques, histoire, anglais, orthographe, éducation musicale et en biologie que les élèves 12, 24, 1 et 23.

- **Axe 2 (17,09%)**



On en déduit que les élèves 23, 13, 1, 10, 17 et 21 ont des meilleures notes en anglais, orthographe et grammaire que les élèves 16, 22, 19, 26 et 2. A l'inverse, les élèves 16, 22, 19, 26 et 2 ont de meilleures notes en rédaction, éducation musicale, technologie et eps que les élèves 23, 13, 1, 10, 17 et 21.

- **Axe 3 (10,71)**



On en déduit que les élèves 23, 13, 7 et 6 ont des meilleures notes en technologie, exposé et récitation que les élèves 23, 4, 25, 3, 20 et 18.

A l'inverse, les élèves 23, 4, 25, 3, 20 et 18 ont de meilleures notes en eps et en géographie que les élèves 23, 13, 7 et 6.