BASES D'ALGEBRE LINEAIRE POUR LA MODELISATION EN SANTE

I. POURQUOI MODELISER EN SANTE

Modéliser permet de reproduire **informatiquement** une situation dans le but de tester des scénarios, des configurations.

Dans le contexte de Big Data (données massives) c'est-à-dire de données de différentes natures, hétérogènes … il est possible de les structurer dans des grands tableaux (matrices) et d'effectuer des opérations mathématiques sur ces matrices (évolution, distances entre profils…).

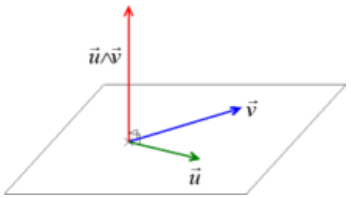
Par exemple, estimer l'effet d'une augmentation du prix des cigarettes sur le système de santé (maladies, coût pour la société…), on est typiquement dans un ***problème multifactoriel*** *et de Big data (il faut un grand nombre d'individus fumeurs ou non-fumeurs pour évaluer cet impact)*

Le calcul matriciel est précieux dans le domaine des **statistiques multivariées**.

II. DEFINITION DE L'ALGEBRE LINEAIRE

L'algèbre linéaire est le domaine des mathématiques qui étudie les transformations linéaires et les espaces vectoriels.

Un **espace vectoriel** est une **structure stable** par addition de vecteurs et par multiplication par un scalaire. Autrement dit, on peut ajouter deux éléments d'un tel espace, ou les multiplier par un nombre, le résultat appartiendra encore à l'espace de départ.



Les transformations linéaires et les espaces vectoriels sont des outils pour l'analyse **multivariée de données** dans un contexte de BIG DATA

A. BASES DU CALCUL MATRICIEL :

Une matrice est un tableau de nombres à n lignes et p colonnes : A (n, p). On aurait par exemple n individus mesurés selon p variables (avec n et p≥1).

- Si p=1, on parle de matrice **univariée** (matrice colonne). Ex : $\mathbf{b} = \begin{bmatrix} 9 \\ 2 \end{bmatrix}$
- Si p≥2, on parle de matrice **multivariée**. Ex : $\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 1 & -1 \end{bmatrix}$
- Si **n=p**, on parle d'une matrice **carrée**, avec autant de colonnes que de lignes.
- Pour calculer le **produit** de deux matrices A et B, il faut **que le nombre de lignes de la deuxième matrice soit égale au nombre de colonnes de la première matrice**. Ainsi A(n,p)*B(p, m) = C(n, m)

Tut'Astuce : Pour calculer le produit A*b, place tes matrices ainsi, et vérifie qu'elles forment un petit carré. Pour A*b ça marche, tu peux donc calculer ce produit.

$\begin{bmatrix} 9 \\ 2 \end{bmatrix}$ $\begin{bmatrix} 2 & 3 \\ 1 & -1 \end{bmatrix}$

$\begin{bmatrix} 2 & 3 \\ 1 & -1 \end{bmatrix}$ $\begin{bmatrix} 9 \\ 2 \end{bmatrix}$

Attention, ici comme tu le vois, tu n'as pas de carré, le produit b*A n'est donc pas calculable !

- **Les puissances d'une matrice n'existent que pour des matrices carrées**, en effet pour effectuer un produit de matrices il faut que le nombre de lignes de la deuxième matrice soit égal au nombre de colonnes de la première matrice, or dans ce cas B=A et donc n=p.

Le tutorat est gratuit. Toute vente ou reproduction est interdite.

B. TRANSPOSEE, INVERSE ET DETERMINANT :

La **transposée** d'une matrice revient à présenter l'information qui est en colonnes en lignes. Ainsi la transposée de A(n,p), notée ${}^tA$ est une matrice à p lignes et n colonnes. Autrement dit, on s'intéresse à **p variables déclinées selon n individus**.

Ex : si $A = \begin{bmatrix}1 & 2\\3 & 4\\5 & 6\end{bmatrix}$, *alors* $tA = \begin{bmatrix}1 & 3 & 5\\2 & 4 & 6\end{bmatrix}$

- Lorsqu'on effectue le **produit tA*A**, on obtient une **matrice carrée d'ordre p** (le nombre de ligne de tA= nombre de colonnes de A). Cette matrice est alors symétrique (par rapport à la diagonale). *Ex :* $\begin{bmatrix}35 & 44\\44 & 56\end{bmatrix}$
- Une matrice est dite **nilpotente** d'ordre n lorsque $A^n=0$ et $A^{n-1}\neq 0$

Point tut' : La définition de matrice nilpotente c'est « matrice dont il existe une puissance égale à la matrice nulle ». Je ne sais pas pourquoi, mais pour le prof, il faut aussi que la puissance corresponde au nombre de lignes de la matrice.
On parle alors juste de matrice « d'ordre n », n correspondant au nombre de ligne ET de colonnes.

L'**inverse** d'une matrice n'existe que pour des matrices carrées sous condition que (**detA ≠0**). L'inverse est défini comme A^{-1} tel que $A.A^{-1}=I$ où I est la matrice identité c'est-à-dire avec tous les coefficients diagonaux égaux à 1 et tous les autres coefficients nuls. (Le déterminant servira justement à calculer l'inverse.)

Ex : $I = \begin{bmatrix}1 & 0 & 0\\0 & 1 & 0\\0 & 0 & 1\end{bmatrix}$, matrice identité d'ordre 3

Le déterminant d'une matrice (detA) est donné par la formule suivante :

- **Det** $\begin{bmatrix}a & b\\c & d\end{bmatrix}$ **= a*d – b*c**
- Det $\begin{bmatrix}a & b & c\\d & e & f\\g & h & i\end{bmatrix}$ = a*Det$\begin{bmatrix}e & f\\h & i\end{bmatrix}$ – b*Det$\begin{bmatrix}d & f\\g & i\end{bmatrix}$ – c*Det $\begin{bmatrix}d & e\\g & h\end{bmatrix}$

= formule trop moche mais c'est visuel…

Ex : Det$\begin{bmatrix}1 & 2 & 3\\4 & 5 & 6\\7 & 8 & 9\end{bmatrix}$ = 1*Det$\begin{bmatrix}5 & 6\\8 & 9\end{bmatrix}$ – 2*Det$\begin{bmatrix}4 & 6\\7 & 9\end{bmatrix}$ – 3*Det $\begin{bmatrix}4 & 5\\7 & 8\end{bmatrix}$

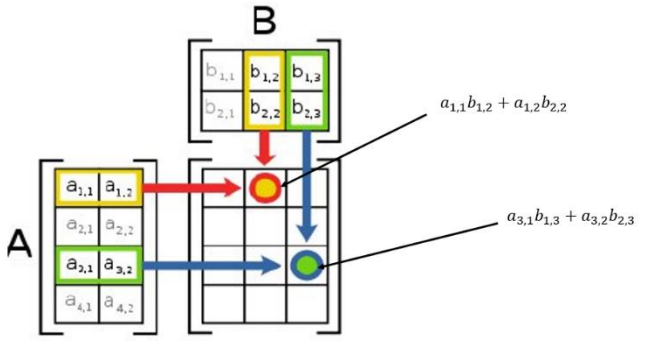
= 1*(5*9 - 6*8) – 2*(4*9 – 7*6) – 3*(4*8- 7*5)

= -3 – 2*(-6) – 3*(-3) = -3 + 12 + 9 = 18

- Soit A $\begin{bmatrix}a & b\\c & d\end{bmatrix}$ une matrice carrée d'ordre 2,

Si Det(A)≠0, alors $A^{-1} = \frac{1}{ad-bc} \times \begin{pmatrix}d & -b\\-c & a\end{pmatrix}$

C. CALCUL MATRICIEL, PRODUIT DE MATRICES :



Ex : Soit A=$\begin{bmatrix}3 & 1\\9 & 12\end{bmatrix}$ *et B=* $\begin{bmatrix}6 & 2\\7 & 5\end{bmatrix}$*, on a alors A*B = C =* $\begin{bmatrix}25 & 11\\138 & 78\end{bmatrix}$ *. C est une matrice carrée d'ordre 2 de coefficients* $c_{i,i}$*. Pour trouver C, on a fait :*

Le tutorat est gratuit. Toute vente ou reproduction est interdite.

- *$c_{1,1}$.(coefficient 1ère ligne 1ère colonne) = 3*6 + 1*7 = 25*
- *$c_{1,2}$ (1ère ligne 2ème colonne) = 3*2 + 1*5 = 11*
- *$c_{2,1}$ (2e ligne 1ère colonne) = 9*6 + 12*7 = 138*
- *$c_{2,2}$ (2e ligne 2e colonne) = 9*2 + 12*5 = 78*

Dans le cas général, **le produit de matrices AB est différent de BA**. Parfois l'un des produits n'existe même pas (*notamment quand le nombre de lignes de la 2e matrice ne correspond pas au nombre de colonnes de la 1ère, cf début du cours*).

- **Si AB=BA**, on dit que les matrices **commutent**.
- On peut aussi avoir un produit de matrices nul, sans que l'une des matrices soit nulle.

Ex : A=$\begin{bmatrix}1 & 0\\0 & 0\end{bmatrix}$ *et B=* $\begin{bmatrix}0 & 0\\0 & 1\end{bmatrix}$*, alors A*B=*$\begin{bmatrix}0 & 0\\0 & 0\end{bmatrix}$

III. Calcul matriciel pour LES ANALYSES MULTIVARIEES (Analyse factorielles) – modélisation en santé

Le terme d'analyses factorielles désigne un ensemble de techniques d'ajustement linéaire dont le but est de **résumer** l'essentiel de l'information contenue dans des gros tableaux de données (plusieurs dizaines, centaines d'individus observés selon un grand nombre de variables (supérieur à 100 par exemple).

Le procédé consiste à passer d'un espace de grandes dimensions à un espace plus petit (factorisation du tableau de données) avec une **perte d'information minimale et contrôlée.**

2 techniques principales d'analyse factorielle :

- L'analyse en **composantes principales** ou **ACP** : la plus ancienne des deux (1933) mais véritablement développée avec l'informatique, employée dans le cadre de variables **quantitatives**, homogènes ou pas.
- L'analyse **factorielle des correspondances** ou **AFC** (années 70), pour l'étude de tableaux de contingence (données **qualitatives**)

D'un point de vue méthodologique, le fonctionnement des deux méthodes est le même, aussi on ne décrira ici que l'ACP.

A. ACP : intérêt et domaines d'application :

L'intérêt de l'ACP est de :

- **Extraire** le maximum d'informations sous une forme **simple** et **cohérente** à partir d'un ensemble très important de données (description synthétique).
- Mettre en évidence 1) Des **interrelations** entre variables (redondance)
  2) Des **ressemblances** et/ou des **oppositions** entre individus (profils)

Les résultats se présentent sous forme de combinaisons linéaires de variables différenciant les individus statistiques.

L'ACP s'applique uniquement sur des variables **QUANTITATIVES** qui peuvent être exprimées soit :

- Dans une même unité (*ex :% de cas Covid19 dans les passages aux urgences*)
- Dans des unités différentes (*ex : mort infantile, revenu par habitant…)*

Le tableau de données est constitué de n individus statistiques (unités spatiales, individus…) caractérisés par p variables quantitatives. Ce tableau de données D constitue une matrice d'informations (n,p) (avec toujours n lignes et p colonnes).

- Chaque ligne = vecteur ligne, décrit un individu selon p variables
- Chaque colonne = vecteur colonne décrit un indicateur selon n individus

Le tutorat est gratuit. Toute vente ou reproduction est interdite.

Ex à petite échelle :

Caractères étudiés = p colonnes

Individus = n lignes

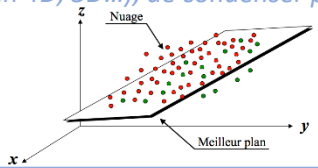
	taille (m)	poids (kg)	Imc	tour de tête (cm)	âge (années)	moyenne au bac	…
patient 1	1,86	85	24,569	57	23	11,5	…
patient 2	1,53	44	18,796	59	72	14	…
…	…	…	…	…	…	…	…

Donc le but de l'ACP est de prendre l'information contenue dans ce tableau de dimension importante et de la représenter sous forme simplifiée.

B. ACP - méthode :

L'ACP consiste à réduire la taille du nuage de points multi-dimensionnels en un nuage de points en 3-4 dimensions.

Tut'Explication : Quand vous analysez 1 variable, vous n'avez besoin pour la représenter que d'une dimension : représentation axiale. Pour 2 variables, 2 dimensions : nuage de points. Par exemple la taille et le poids d'individus, avec le poids en ordonné, la taille en abscisse. Si vous voulez rajouter une donnée (comme la moyenne au bac), il faudrait pour que votre graphe soit clair, passer à une représentation 3D, avec un 3e axe (faisable sur ordi). Au-delà, on utilisera l'ACP, qui permet (à défaut de pouvoir passer en 4D, 5D…), de condenser plusieurs types d'informations dans les 1ers axes.



Pour cela on fait une projection selon des axes (axes factoriels ou facteurs Fi). Ces axes sont des combinaisons linéaires de variables.

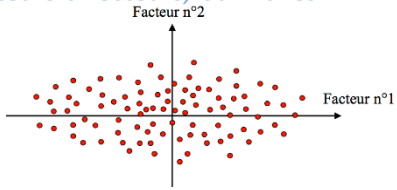
$F_i=A_1X_1+A_2X_2+.....A_pX_p$ *avec la plupart des* $A_i=0$

Les **coefficients Ai** permettent de mesurer l'intensité de relation de chaque variable avec le facteur considéré (maximum à nul selon les composantes). Ces coefficients changent d'un facteur à l'autre.

Tut' Explication : les coefficients sont les données dans les cases de la matrice

Les facteurs sont hiérarchisés : **l'axe 1** compte le **maximum d'informations**, c'est l'axe de plus grande dispersion du nuage de points, mais il laisse de côté les résidus. C'est le 2e axe qui prend en compte le maximum d'informations résiduelles et ainsi de suite pour les axes suivants.

Par construction, **tous les axes (facteurs) sont non corrélés, ils forment donc des angles droits 2 à 2**. *(Tut' intervention : Détail pas mentionné par le prof, mais lorsque les facteurs sont corrélés ils forment des angles aigues ou obtus. Important aussi pour comprendre : un plan est formé par 2 vecteurs directeurs, ou 2 axes ayant la même origine)*



C. ACP - Calcul des axes factoriels :

L'idée est de transformer la matrice d'information en une matrice de projection des individus statistiques sur les axes.

Etape importante, Centrer-réduire les données :

Centrer-réduire les données permet de gommer les effets taille. Si les données sont assez **homogènes**, on pratique un simple **centrage**. Si elles sont assez **hétérogènes**, le **centrer-réduire est obligatoire** (on ramène la moyenne à 0 et l'écart-type à 1), ce qui donne une **ACP normée**.

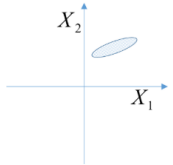


Fig. 7 ACP générale

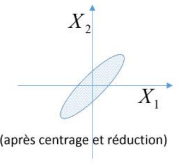


Fig. 9 ACP normée

Le tutorat est gratuit. Toute vente ou reproduction est interdite.

Une ACP **normée** consiste en :

- Variables centrées-réduites
- Projections orthogonales
- Méthode des moindres carrés

Détermination des axes factoriels :

Tut'Recap : Pour essayer de comprendre…

Les axes = facteurs, sont définis **séquentiellement** : On détermine l'axe (premier axe factoriel) sur lequel le nuage se déforme le moins possible en projection. On cherche ensuite un second axe, sur lequel le nuage se déforme le moins en projection, après le premier axe, tout en étant **orthogonal au premier.** On réitère jusqu'à l'obtention de p axes.

Le meilleur axe (premier axe factoriel) sera celui sur lequel le nuage de points projeté est de dispersion maximale, c.a.d tel que le nuage de points projeté est **d'inertie** maximale. D'où la « **matrice d'inertie** » dans la suite du cours…

La matrice de données **D** (individus, variables) est une matrice à n lignes et p colonnes. On lui associe une matrice **D'**, la **matrice transposée de D**.

En faisant le produit de **D'*D,** on obtient une matrice carrée, symétrique d'ordre p, appelée **matrice d'inertie**, notée **T**.

Toute l'ACP repose sur du calcul matriciel. Les Axes sont définis par des **vecteurs propres et valeurs propres** :

- Un **vecteur propre** (ici V) est un vecteur tel que **T.V= µV** où **µ est une valeur propre.**

Ex : Soit T= $\begin{bmatrix}1 & 2\\ 2 & 1\end{bmatrix}$*. Pour* ***trouver la valeur propre µ****, on résout le système*

T.V = µV => $T - \mu I = \begin{pmatrix}1 & 2\\ 2 & 1\end{pmatrix} - \mu\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix} = \begin{pmatrix}1-\mu & 2\\ 2 & 1-\mu\end{pmatrix} = 0$

(Quand on divise par V des 2 côtés de l'égalité, comme c'est du calcul matriciel on ne peut pas écrire T= µ, on est obligé de garder une forme matricielle à gauche comme à droite, donc on multiplie par I, la matrice identité d'ordre 2.)

Pour trouver une solution à ce type d'équation, on cherche à ce que le déterminant de la matrice T-µI soit égal à 0. On aura donc…

Det $(T - \mu I) = (1-\mu)^2 - 4 = (-1-\mu)(3-\mu) = 0$
On a alors 2 possibilités : µ=-1 et µ=3.
On résout alors les systèmes T.V=-V et T.V=3V (le prof ne détaille que le 2nd)

$\Rightarrow \begin{pmatrix}1 & 2\\ 2 & 1\end{pmatrix} * V = 3V$

$\Rightarrow \begin{pmatrix}1 & 2\\ 2 & 1\end{pmatrix} * \begin{pmatrix}V_1\\ V_2\end{pmatrix} = 3\begin{pmatrix}V_1\\ V_2\end{pmatrix}$

$\Rightarrow \begin{pmatrix}1V_1 + 2V_2\\ 2V_1 + 1V_2\end{pmatrix} = \begin{pmatrix}3V_1\\ 3V_2\end{pmatrix}$

$\Rightarrow \begin{cases}1V_1 + 2V_2 = 3V_1\\ 2V_1 + 1V_2 = 3V_2\end{cases}$

$\Rightarrow V_1 = V_2$ *donc V=* $\begin{pmatrix}\frac{1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}$ ★ *car les vecteurs sont normés.*

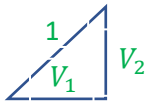
De même, l'autre équation amène à $V_3 = -V_4$ *et* $V = \begin{pmatrix}\frac{-1}{\sqrt{2}}\\ \frac{1}{\sqrt{2}}\end{pmatrix}$ ★★

Remarquons que nos 2vecteurs obtenus sont bien orthogonaux (leur produit scalaire est nul : les axes sont donc bien décorrélés) ★★★

Le tutorat est gratuit. Toute vente ou reproduction est interdite.

Point Tut' Explication :

★ Le prof ne le précise pas, mais on cherche un vecteur **unitaire**. Un vecteur unitaire est un vecteur **dont la norme est égale à 1** (c'est surement ce qu'il appelle un vecteur « normé »). La norme c'est la « longueur » du vecteur. Pour la calculer, on visualise un triangle rectangle :



On a donc d'après le théorème de Pythagore $1= {V_1}^2 + {V_2}^2$,

et comme $V_1 = V_2$, *ça nous fait* $1=2{V_1}^2$ donc $V_1 = V_2 = \sqrt{\frac{1}{2}} = \frac{1}{\sqrt{2}}$

★★ Le raisonnement est le même que pour l'autre système :

$$\Rightarrow \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} * \begin{pmatrix} V_3 \\ V_4 \end{pmatrix} = -\begin{pmatrix} V_3 \\ V_4 \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} 1V_3 + 2V_4 \\ 2V_3 + 1V_4 \end{pmatrix} = \begin{pmatrix} -V_3 \\ -V_4 \end{pmatrix} \quad \Rightarrow \quad V_3 = -V_4$$

★★★ *soit les vecteurs $\vec{u}(x,y)$ et $\vec{v}(x',y')$, pour voir s'ils sont orthogonaux (perpendiculaires), on calcule leur* ***produit scalaire =*** $xx' + yy'$
Dans notre ex ci-dessus, nous avons le produit scalaire =
$V_1 \times V_3 + V_2 \times V_4 = \frac{1}{\sqrt{2}} \times \left(-\frac{1}{\sqrt{2}}\right) + \frac{1}{\sqrt{2}} \times \frac{1}{\sqrt{2}} = 0$ *donc ils sont* ***orthogonaux****.*

D. ACP - part d'explication des axes factoriels :

La **part d'explication** d'un axe est donné par la formule suivante :

$$\mu_i\% = \frac{\mu_i}{\sum \mu_i} * 100$$

Dans le cas à n dimensions, deux situations peuvent se produire

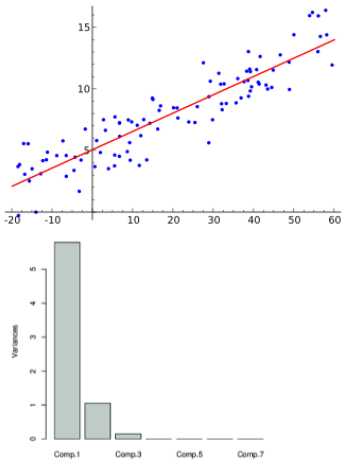
- Soit **l'histogramme des valeurs propres** est assez droit, le nuage de points est plutôt **arrondi sans axe d'allongement** véritablement marqué : on peut en déduire que les interrelations entre les variables sont sans doute faibles et qu'il ne se dégage pas de combinaisons simples de l'ensemble des données.
- Soit **l'histogramme des valeurs propres** est assez **concentré**, les valeurs propres sont très **différenciées** et on perçoit l'existence de deux **axes d'allongement très marqués**, on peut s'attendre alors à ce qu'il ressorte une structure de différenciation forte (profils marqués)

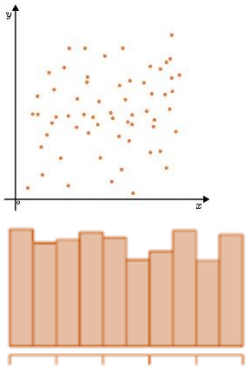
Le tutorat est gratuit. Toute vente ou reproduction est interdite.

Point Tut' Explication :

La valeur propre d'un axe permet de caractériser l'importance de cet axe par rapport aux valeurs de notre nuage de points.

Par exemple, ce nuage de points dans un plan 2D peut être caractérisé de manière très significative par la droite rouge, qui est alors un facteur (=axe factoriel). On peut supposer que ce facteur aura alors une valeur propre µ et une part d'explication (implication dans la forme du nuage) importante ! On pourrait alors avoir un histogramme assez concentré avec un axe d'allongement au niveau de µ1 très marqué, comme si contre.





Au contraire, sur ce 2e type de nuage de points, on n'observe aucun axe d'allongement marqué, les valeurs sont dispersées. L'histogramme des valeurs propres est alors assez droit, aucune ne prend le dessus sur les autres. On peut en déduire que les interrelations entre les variables sont faibles

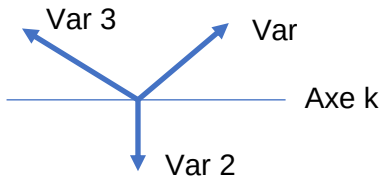
Bonus pour la suite du cours : l'inertie mesure la dispersion totale du nuage. C'est la somme des variances de chaque variable. G le centre de gravité est une sorte de moyenne de tous les points de notre nuage.

E. ACP – interprétation d'une analyse factorielle :

Plusieurs données doivent être analysées :

- Les **coordonnées** sur les axes factoriels : elles donnent la **position** des individus par rapport aux axes factoriels. On peut alors mettre en évidence des **oppositions entre groupe d'individus** par rapport aux combinaisons de variables définies par les axes.

- La **qualité de représentation** des individus sur les axes : deux points distincts peuvent avoir la même projection sur l'axe factoriel, mais l'un sera mieux représenté que l'autre (angle plus petit).



Ex : Nos variables 1 et 3 forment un angle aigue avec l'axe k, donc elles sont bien représentées. En revanche, la variable 2 est orthogonale donc n'a pas une bonne qualité de représentation.

- La **contribution des individus dans la formation de l'axe** : les individus contribuent plus ou moins à déterminer la direction des différents axes d'allongement du nuage. Elle est mesurée par la **part de l'individu dans la variance**.
- La **part de l'individu dans l'inertie totale** du nuage = **INR** : elle est proportionnelle à sa distance au centre de gravité G. Elle donne une idée de la spécificité de l'individu par rapport à la moyenne.

Le tutorat est gratuit. Toute vente ou reproduction est interdite.

Voilà pour ce tout nouveau cours, et pas des plus simples. J'ai essayé de vous expliquer au maximum tout ce que je pouvais, mais je ne suis malheureusement pas prof agrégée (lol).

Comme vous l'avez peut-être vu si vous avez regardé la diapo du prof, ses qcm de fin de cours ne traitent que de la partie MATRICES. On vous dira au plus vite si le reste est aussi susceptible de tomber à l'examen ou non. En attendant la réponse du prof, essayez au moins de comprendre un maximum la 2e partie.

A noter, hyper important : **tous les encadrés en bleu sont des AIDES à la compréhension et viennent de mes recherches PERSONNELLES**. Ce n'est donc pas à apprendre, et il se peut que je ne sois pas toujours la plus exacte (mais j'ai fait au mieux promis).

Cette fiche est donc susceptible de changer si le prof apporte quelques précisions.

Dernier point important : **faites les QCM de la diapo** !!!! Les matrices ça vient en s'exerçant, vous aurez quelques dm d'entraînement aussi, mais là c'est un cadeau qu'il vous fait de vous donner tout ça !

Je suis de tout cœur avec vous, des gros matheux jusqu'aux plus réfractaires. COURAGE, encore une fois ne lâchez rien. Fred'.

Le tutorat est gratuit. Toute vente ou reproduction est interdite.