Ecole Supérieure en Sciences Appliquées de Tlemcen

Analyse de données

Chapitre 4: Analyse en composantes principales

Partie 2

Présentée par:

Dr Imane NEDJAR

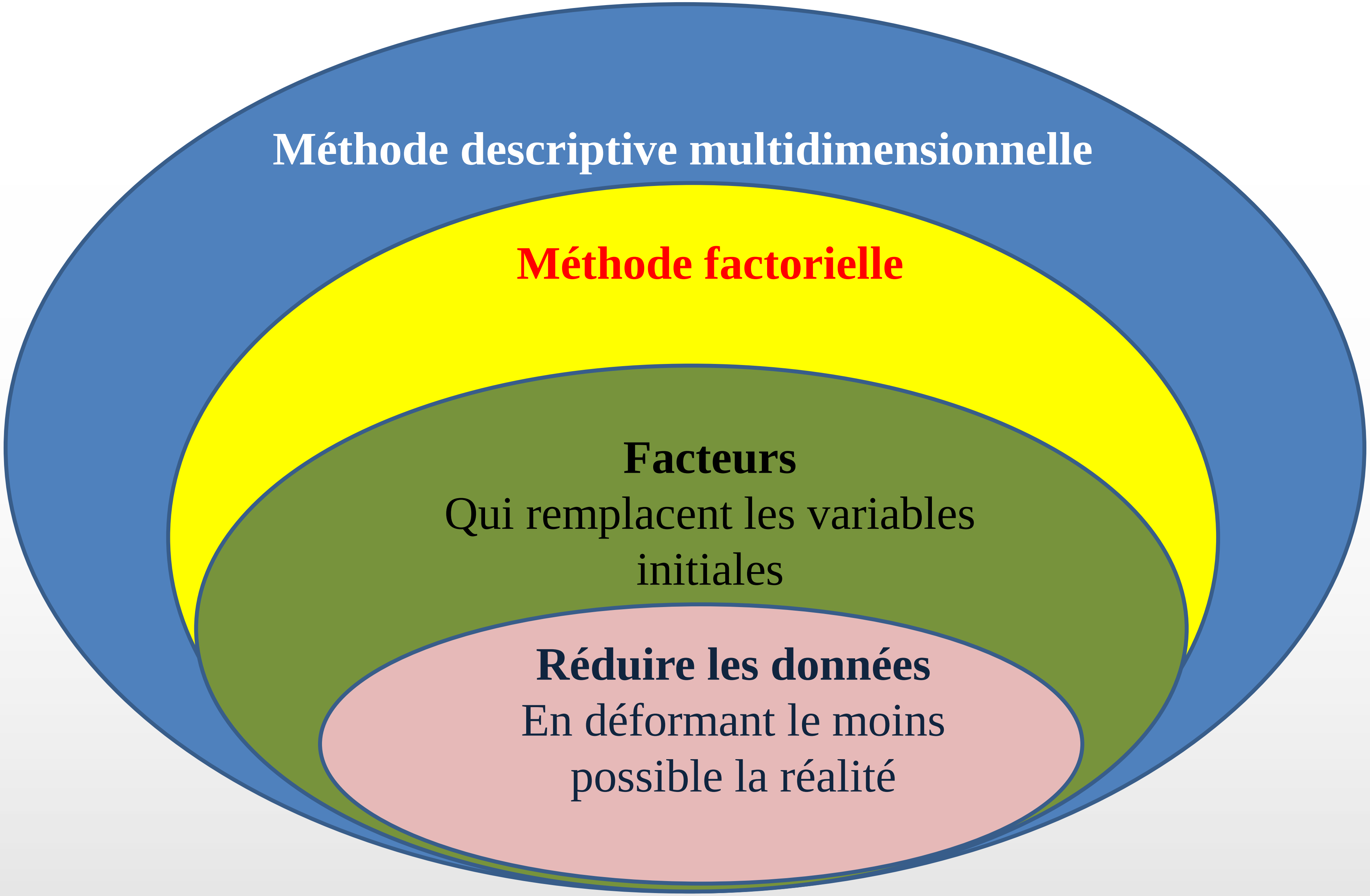
Analyse en composantes principales (ACP)

- ACP est l'une des méthodes d'analyse de données **Multivariées**. Permettant d’explorer des données **Multidimensionnels** constitués de variables **Quantitatives**.

- Il convertit un ensemble d'observations de variables éventuellement corrélées en un ensemble de valeurs de variables linéairement non corrélées appelées **Composantes principales**

- ACP est une procédure statistique utilisée pour réduire la dimensionnalité.

Analyse en composantes principales (ACP)



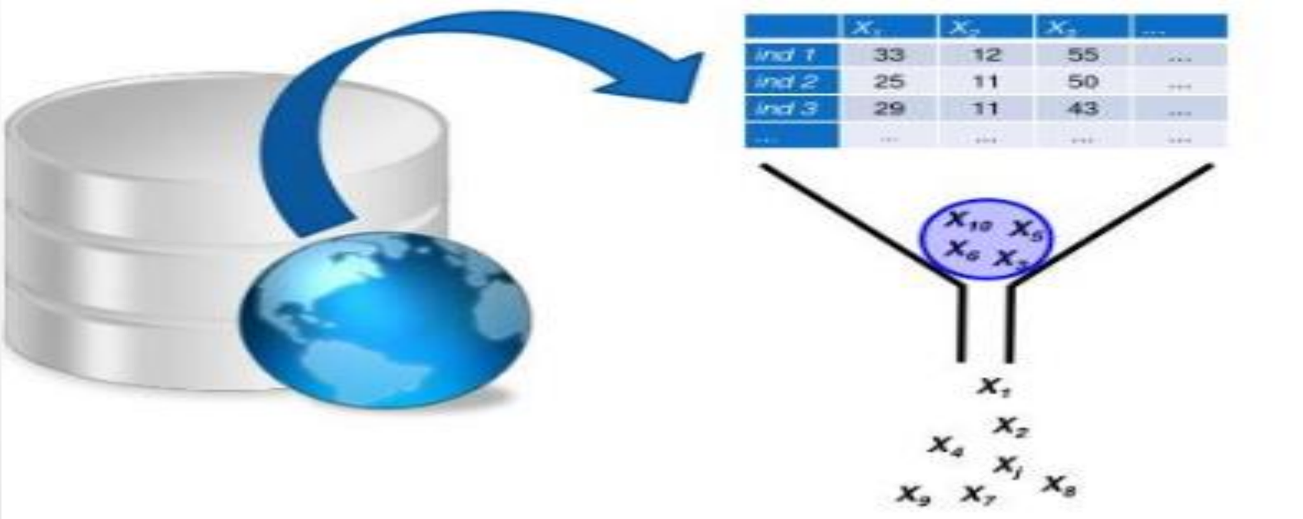
Analyse en composantes principales (ACP)

Représenter les données au mieux dans un espace plus réduit des observations issues d’un espace plus grand en nombres de dimensions (Xj variables) afin de :

•Simplification de la réalité

•Concentration d’une information de départ diluée

•Description du maximum de variabilité dans un espace réduit



Détails des Calculs de l’ACP

- L’ACP s’intéresse à des tableaux de données rectangulaires avec des **Individus** en lignes et des **variables Quantitatives** en colonnes.

$$
\underset{(n,p)}{X} = \begin{array}{c} \\ 1 \\ \vdots \\ i \\ \vdots \\ n \end{array}
\begin{array}{c}
\begin{array}{ccccc} 1 & \cdots & j & \cdots & p \end{array} \\
\begin{bmatrix} & & \vdots & & \\ & & \vdots & & \\ \cdots & \cdots & \boxed{x_{i,j}} & \cdots & \cdots \\ & & \vdots & & \\ & & \vdots & & \end{bmatrix}
\end{array}
$$

p variables **métriques**

n individus

valeur de la variable *j* prise par l’individu *i*

Projeter les observations depuis l'espace à P dimensions des P variables vers un espace à K dimensions (K < P) tel qu'un maximum d'information soit conservée

Détails des Calculs de l’ACP

Variables $\qquad\qquad\qquad\qquad$ Composantes

$$\underset{(n,p)}{X} = \text{Individus}\;\begin{array}{cc} & 1 \quad \cdots \quad j \quad \cdots \quad p \\ \begin{array}{c} 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} & \begin{bmatrix} & & \vdots & & \\ & & \vdots & & \\ \cdots & \cdots & x_{i,j} & \cdots & \cdots \\ & & \vdots & & \\ & & \vdots & & \end{bmatrix} \end{array} \Longrightarrow \begin{array}{cc} & 1 \quad \cdots \quad j \quad \cdots \quad p \\ \begin{array}{c} 1 \\ \vdots \\ i \\ \vdots \\ n \end{array} & \begin{bmatrix} & & \vdots & & \\ & & \vdots & & \\ \cdots & \cdots & u_{i,j} & \cdots & \cdots \\ & & \vdots & & \\ & & \vdots & & \end{bmatrix} \end{array}$$

Projeter les observations depuis l'espace à P dimensions des P variables vers un espace à K dimensions (K < P) tel qu'un maximum d'information soit conservée

Détails des Calculs de l’ACP

Étape 1:

standardisez l'ensemble de données

Étape 2:

Calculez la matrice de covariance des entités du jeu de données

Étape 3:

Calculez les valeurs propres et les vecteurs propres de la matrice de covariance

Étape 4:

Trier les valeurs propres et leurs vecteurs propres correspondants

Étape 5:

Choisissez k valeurs propres et formez une matrice de vecteurs propres

Étape 6:

Transformez la matrice d'origine

(Matrice de caractéristiques * k principaux vecteurs propres = données transformées)

Détails des Calculs de l'ACP

Soit le tableau suivant :

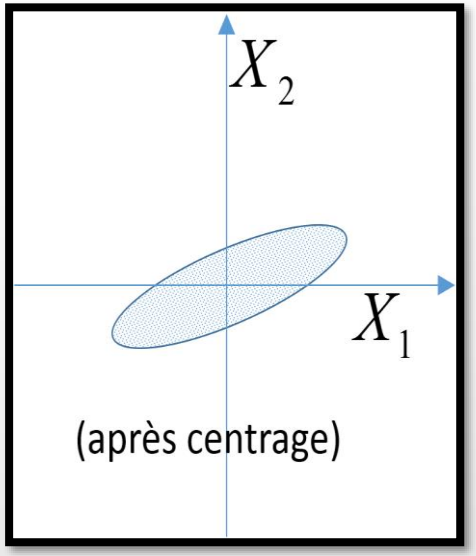
f1	f2	f3	f4
1	2	3	4
5	5	6	7
1	4	2	3
5	3	2	1
8	1	2	2

Détails des Calculs de l'ACP

Étape 1: standardisez l'ensemble de données

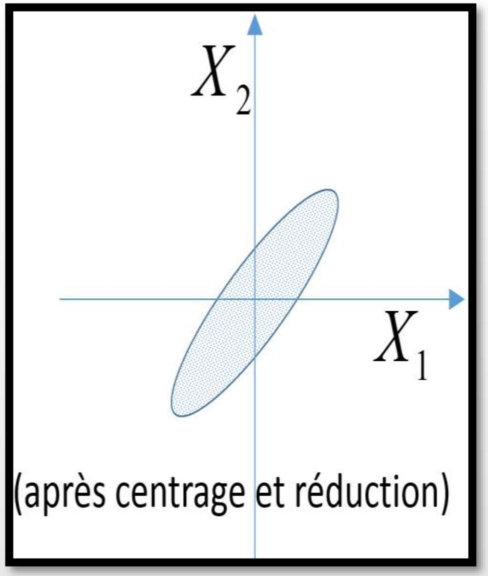
ACP centrée

$$X_{new} = X - \mu$$



ACP centrée et réduite

$$X_{new} = \frac{X - \mu}{\sigma}$$



Détails des Calculs de l'ACP

Étape 1: standardisez l'ensemble de données

	f1	f2	f3	f4
μ =	4	3	3	3.4
σ =	3	1.58114	1.73205	2.30217

f1	f2	f3	f4
-1	-0.63246	0	0.26062
0.33333	1.26491	1.73205	1.56374
-1	0.63246	-0.57735	-0.17375
0.33333	0	-0.57735	-1.04249
1.33333	-1.26491	-0.57735	-0.60812

Détails des Calculs de l'ACP

Étape 2: Calculez la matrice de covariance/Corrélation des entités du jeu de données

Les données centrées

$$C=\begin{pmatrix} \mathrm{var}(X_1) & \mathrm{cov}(X_1,X_2) & \ldots & \mathrm{cov}(X_1,X_j) & \ldots & \mathrm{cov}(X_1,X_p) \\ \mathrm{cov}(X_2,X_1) & \mathrm{var}(X_2) & \ldots & \mathrm{cov}(X_2,X_j) & \ldots & \mathrm{cov}(X_2,X_p) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \mathrm{cov}(X_j,X_1) & \mathrm{cov}(X_j,X_2) & \ldots & \mathrm{var}(X_j) & \ldots & \mathrm{cov}(X_j,X_p) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \mathrm{cov}(X_p,X_1) & \mathrm{cov}(X_p,X_2) & \ldots & \mathrm{cov}(X_p,X_j) & \ldots & \mathrm{var}(X_p) \end{pmatrix}$$

$$\mathrm{cov}(x_1,x_2)=\frac{1}{n-1}\left(\sum_{i=1}^{n}(x_{i,1}-\bar{x}_1)(x_{i,2}-\bar{x}_2)\right)$$

Les données centrées et réduites

$$C=\begin{pmatrix} 1 & \rho(X_1,X_2) & \ldots & \rho(X_1,X_j) & \ldots & \rho(X_1,X_p) \\ \rho(X_2,X_1) & 1 & \ldots & \rho(X_2,X_j) & \ldots & \rho(X_2,X_p) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho(X_j,X_1) & \rho(X_j,X_2) & \ldots & 1 & \ldots & \rho(X_j,X_p) \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho(X_p,X_1) & \rho(X_p,X_2) & \ldots & \rho(X_p,X_j) & \ldots & 1 \end{pmatrix}$$

$$corr(x_1,x_2)=\rho(x_1,x_2)=\frac{\mathrm{cov}(X_1,X_2)}{\sigma(X_1)\sigma(X_2)}$$

Détails des Calculs de l'ACP

Étape 2: Calculez la matrice de covariance des entités du jeu de données avec (n=5)

	f1	f2	f3	f4
f1	var(f1)	cov(f1,f2)	cov(f1,f3)	cov(f1,f4)
f2	cov(f2,f1)	var(f2)	cov(f2,f3)	cov(f2,f4)
f3	cov(f3,f1)	cov(f3,f2)	var(f3)	cov(f3,f4)
f4	cov(f4,f1)	cov(f4,f2)	cov(f4,f3)	var(f4)

	f1	f2	f3	f4
f1	0.8	-0.25298	0.03849	-0.14479
f2	-0.25298	0.8	0.51121	0.4945
f3	0.03849	0.51121	0.8	0.75236
f4	-0.14479	0.4945	0.75236	0.8

Détails des Calculs de l'ACP

Étape 3: Calculez les valeurs propres et les vecteurs propres de la matrice de covariance

Les vecteurs propres
ce sont les directions dans lesquelles la matrice agit.

Les valeurs propres
c'est le facteur multiplicatif associe a une direction donnée.

Un vecteur **v** de taille p est un vecteur propre d'une matrice **A** de taille p x p s'il existe $\lambda \in \mathbf{C}$ telle que

$$Av = \lambda v$$

Détails des Calculs de l'ACP

Étape 3: Calculez les valeurs propres et les vecteurs propres de la matrice de covariance

$$Av = \lambda v$$

$$Av - \lambda v = 0$$

$$(A - \lambda I)v = 0$$

$$\det(A - \lambda I) = 0$$

Détails des Calculs de l'ACP

Étape 3: Calculez les valeurs propres et les vecteurs propres de la matrice de covariance

$$\det(A - \lambda I) = 0$$

	f1	f2	f3	f4
f1	0.8 - λ	-0.25298	0.03849	-0.14479
f2	-0.25298	0.8- λ	0.51121	0.4945
f3	0.03849	0.51121	0.8 - λ	0.75236
f4	-0.14479	0.4945	0.75236	0.8 - λ

$$\lambda = 2.51579324\ ,\ 1.0652885\ ,\ 0.39388704\ ,\ 0.02503121$$

Détails des Calculs de l'ACP

Étape 3: Calculez les valeurs propres et les vecteurs propres de la matrice de covariance

$$(A-\lambda I)v = 0$$

$$\begin{pmatrix} 0.800000-\lambda & -(0.252982) & 0.038490 & -(0.144791) \\ -(0.252982) & 0.800000-\lambda & 0.511208 & 0.494498 \\ 0.038490 & 0.511208 & 0.800000-\lambda & 0.752355 \\ -(0.144791) & 0.494498 & 0.752355 & 0.800000-\lambda \end{pmatrix} \times \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = 0$$

Détails des Calculs de l'ACP

Étape 3: Calculez les valeurs propres et les vecteurs propres de la matrice de covariance

$$(A - \lambda I)v = 0$$

Pour λ = *2.51579324, les valeurs du vecteur V obtenu sont*

x1 = 0.16195986

x2 = -0.52404813

x3 = -0.58589647

x4 = -0.59654663

Détails des Calculs de l'ACP

Étape 4: Trier les valeurs propres et leurs vecteurs propres correspondants

Vp1	Vp2	Vp3	Vp4
0.161960	-0.917059	-0.307071	0.196162
-0.524048	0.206922	-0.817319	0.120610
-0.585896	-0.320539	0.188250	-0.720099
-0.596547	-0.115935	0.449733	0.654547

Détails des Calculs de l'ACP

Étape 5: Choisissez k valeurs propres et formez une matrice de vecteurs propres

Vp1	Vp2
0.161960	-0.917059
-0.524048	0.206922
-0.585896	-0.320539
-0.596547	-0.115935

Détails des Calculs de l’ACP

Étape 6: Transformez la matrice d'origine
(Matrice de caractéristiques * k principaux vecteurs propres = données transformées)

f1	f2	f3	f4
-1.000000	-0.632456	0.000000	0.260623
0.333333	1.264911	1.732051	1.563740
-1.000000	0.632456	-0.577350	-0.173749
0.333333	0.000000	-0.577350	-1.042493
1.333333	-1.264911	-0.577350	-0.608121

(5,4)

*

Vp1	Vp2
0.161960	-0.917059
-0.524048	0.206922
-0.585896	-0.320539
-0.596547	-0.115935

(4,2)

=

ft1	ft2
0.014003	0.755975
-2.556534	-0.780432
-0.051480	1.253135
1.014150	0.000239
1.579861	-1.228917

(5,2)

Propriétés de l'ACP

- Les valeurs propres trouves étant simples, les espaces propres associes aux vecteurs propres seront des droites vectorielles (on les appelles des axes factoriels ou des facteurs).

- U_1 est le vecteur propre unitaire associé à la plus grande valeur propre $\lambda 1$, il vérifie $\mathbf{A\,U_1 = \lambda_1\,U_1}$ et $\mathbf{\|U_1\| = 1}$

- D'un point de vue général, l'ACP nous a permit de traiter un très grand nombre de données (matrice) pour identifier un nombre relativement restreint de données (axes factoriels)

Propriétés de l'ACP

- **Qualité d'un axe factoriel α**

➔ Pourcentage de variation expliqué par l'axe α = la part de toute l'information initiale « visible » sur l'axe α

$$\frac{\lambda_\alpha}{\sum_{1}^{p} \lambda_i}$$

- **Qualité d'un sous espace à q dimensions**

➔ Pourcentage de variation expliqué par le sous-espace à q dimensions = la part de toute l'information initiale « visible » sur le sous-espace à q dimensions

$$\frac{\lambda_1 + \lambda_2 + \ldots + \lambda_q}{\sum_{1}^{p} \lambda_i}$$

Rappel sur les calcules

Exemple de calcul du déterminant d'une matrice 3 x 3

$$
\begin{vmatrix} -1 & 2 & 5 \\ 1 & 2 & 3 \\ -2 & 8 & 1 \end{vmatrix}
\qquad
\begin{array}{c} \begin{matrix} + & - & + \end{matrix} \\ \begin{vmatrix} -1 & 2 & 5 \\ 1 & 2 & 3 \\ -2 & 8 & 1 \end{vmatrix} \end{array}
\qquad
\begin{array}{c} \begin{matrix} + & - & + \end{matrix} \\ \begin{vmatrix} -1 & 2 & 5 \\ 1 & 2 & 3 \\ -2 & 8 & 1 \end{vmatrix} \end{array}
\qquad
\begin{array}{c} \begin{matrix} + & - & + \end{matrix} \\ \begin{vmatrix} -1 & 2 & 5 \\ 1 & 2 & 3 \\ -2 & 8 & 1 \end{vmatrix} \end{array}
$$

$$
\mathbf{-1}\begin{vmatrix} 2 & 3 \\ 8 & 1 \end{vmatrix}
\qquad
\mathbf{-2}\begin{vmatrix} 1 & 3 \\ -2 & 1 \end{vmatrix}
\qquad
\mathbf{+5}\begin{vmatrix} 1 & 2 \\ -2 & 8 \end{vmatrix}
$$

$$-1(2*1-3*8)-2(1*1-3*(-2))+5(1*8-2*(-2))=32$$

Rappel sur les calcules

Définition

- Une équation du second degré est une équation de la forme

$ax^2+bx+c=0$ *où a, b et c sont des réels avec* $a \neq 0$.

Une solution de cette équation s'appelle une racine du trinôme $ax^2+bx+c=0$.

- On appelle discriminant du trinôme $ax^2 + bx + c$, le nombre réel, noté Δ, égal à b^2-4ac .

Rappel sur les calcules

Définition

Propriété : Soit Δ le discriminant du trinôme ax^2 + bx + c .

- Si $\Delta < 0$: L'équation ax^2 + bx + c = 0 n'a pas de solution réelle.

- Si $\Delta = 0$: L'équation ax^2 + bx + c = 0 a une unique solution :

- Si $\Delta > 0$: L'équation ax^2 + bx + c = 0 a deux solutions distinctes :

$$x_1 = \frac{-b-\sqrt{\Delta}}{2a} \qquad x_2 = \frac{-b+\sqrt{\Delta}}{2a}$$