Chapitre I : Analyse Descriptive : (Apprentissage Non Supervisé :Réduction de Dimensionnalité)

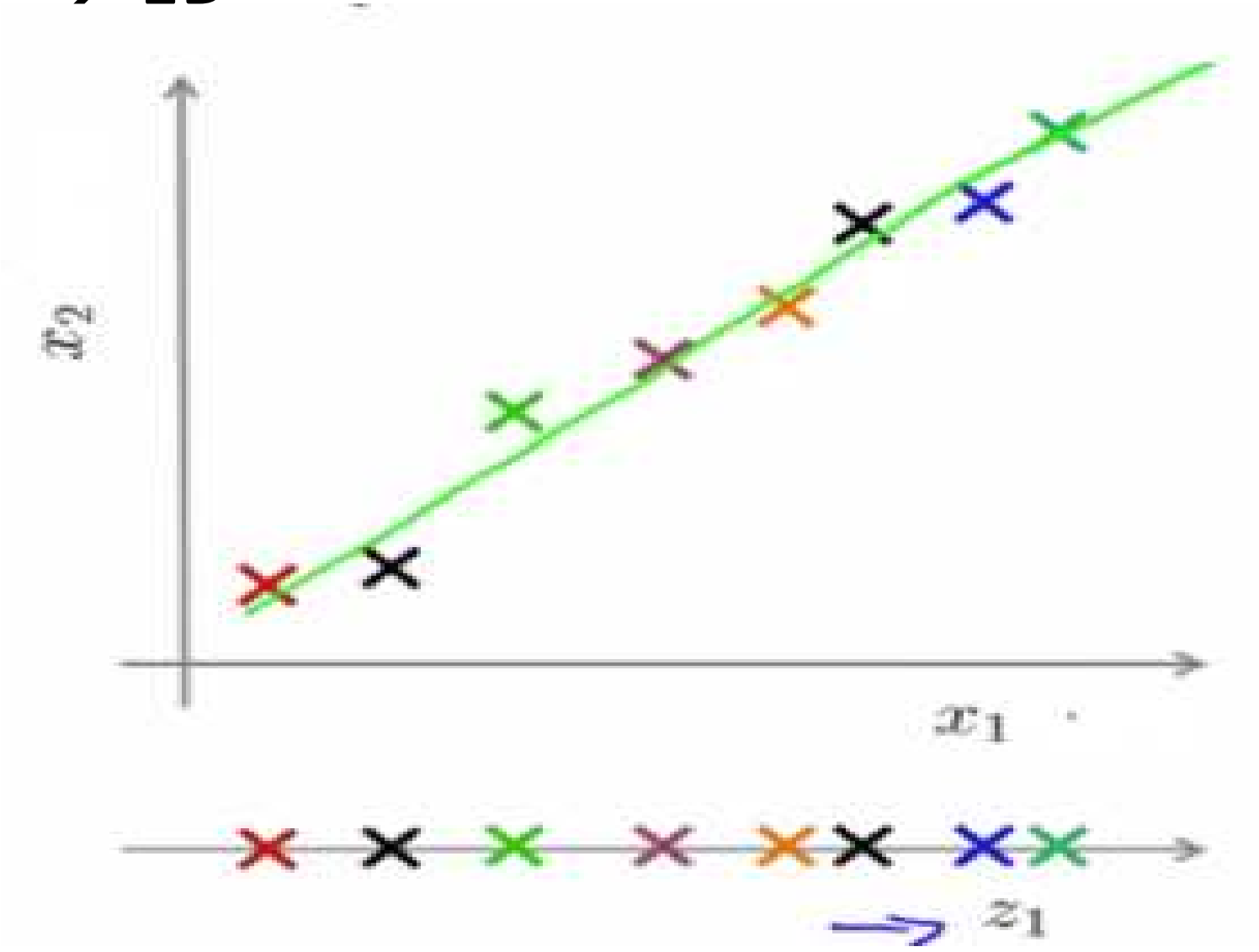
I-1- Analyse en Composantes Principales

ACP

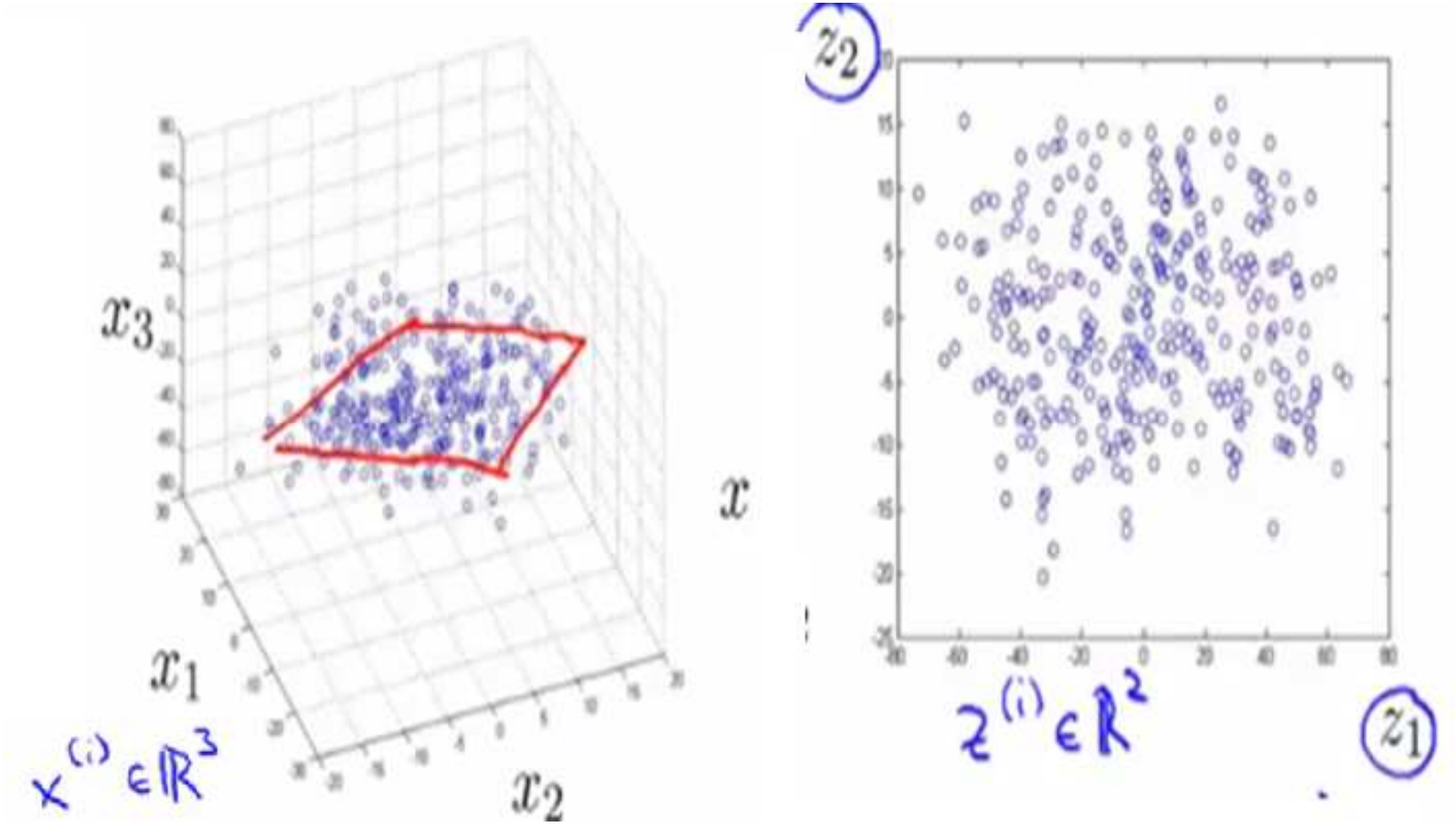
PLAQUETTE COMMERCIALE

contoso

2D ---→ 1D



3D --→ 2D



De la même façon on réduit la dimension 1000D a 100D,

Réduction de Dimension:

- A pour but de **résumer** les données en leur assignant une nouvelle représentation en faisant ressortir ce qui est dissimulé par le volume.
- Elle consiste à transformer un ensemble de variables initiales dont certaines sont corrélées en un ensemble de variables non corrélées appelées composantes principales, en préservant autant que possible la variance des données c.a.d l'essentiel de l'information.

Pourquoi Réduire la dimension des données ?

1- Compression de données :

- Gagner de l'espace mémoire
- Rendre l'algorithme d'apprentissage plus rapide.

2-Amélioration des performances des algorithmes.

3- Visualisation des données.

Exemple:

Il s'agit d'analyser un tableau de plusieurs pays par rapport à 50 variables comme indiqué dans le tableau ➔ P**roblème:**

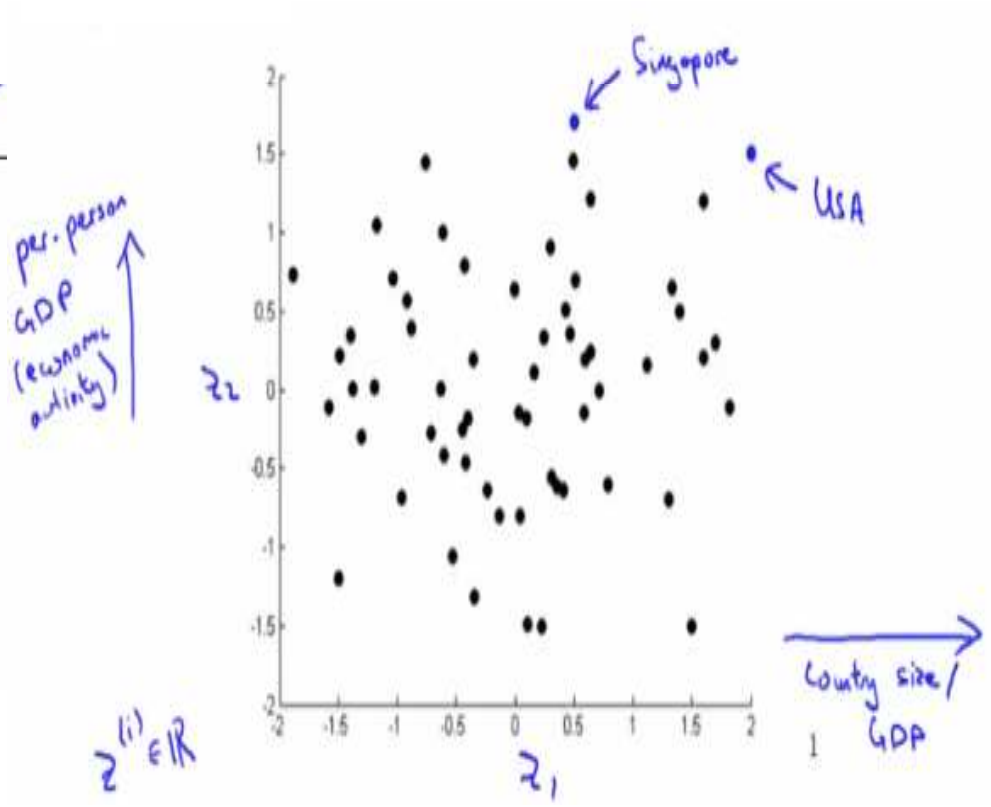
Bcp de Données ➔ on ne voit pas l'information?

Country	X_1 GDP (trillions of US$)	X_2 Per capita GDP (thousands of intl. $)	X_3 Human Develop-ment Index	X_4 Life expectancy	X_5 Poverty Index (Gini as percentage)	X_6 Mean household income (thousands of US$)	...
⇒Canada	1.577	39.17	0.908	80.7	32.6	67.293	...
China	5.878	7.54	0.687	73	46.9	10.22	...
India	1.632	3.41	0.547	64.7	36.8	0.735	...
Russia	1.48	19.84	0.755	65.5	39.9	0.72	...
Singapore	0.223	56.69	0.866	80	42.5	67.1	...
USA	14.527	46.86	0.91	78.3	40.8	84.3	...

Exemple:

- ✓Bcp de Données, on ne peut les analyser
- ✓On ne peut les visualiser pour donner des descriptions
- ➔ **On réduit l'espace de 50D à 2D**

Country	z_1	z_2
Canada	1.6	1.2
China	1.7	0.3
India	1.6	0.2
Russia	1.4	0.5
Singapore	0.5	1.7
USA	2	1.5



I-1-Analyse en Composantes Principales : ACP

- l'ACP est la méthode la plus ancienne parmi les méthodes de réduction de dimension.
- Elle traite des données consignées dans des tableaux individus-variables où les « n » variables sont quantitatives, continues, homogènes ou non à priori corrélées entre elles.
- Elle décrie à l'aide de k < n composantes un maximum de variabilité ce qui permet :
  - Une réduction des données à k nouveaux descripteurs.

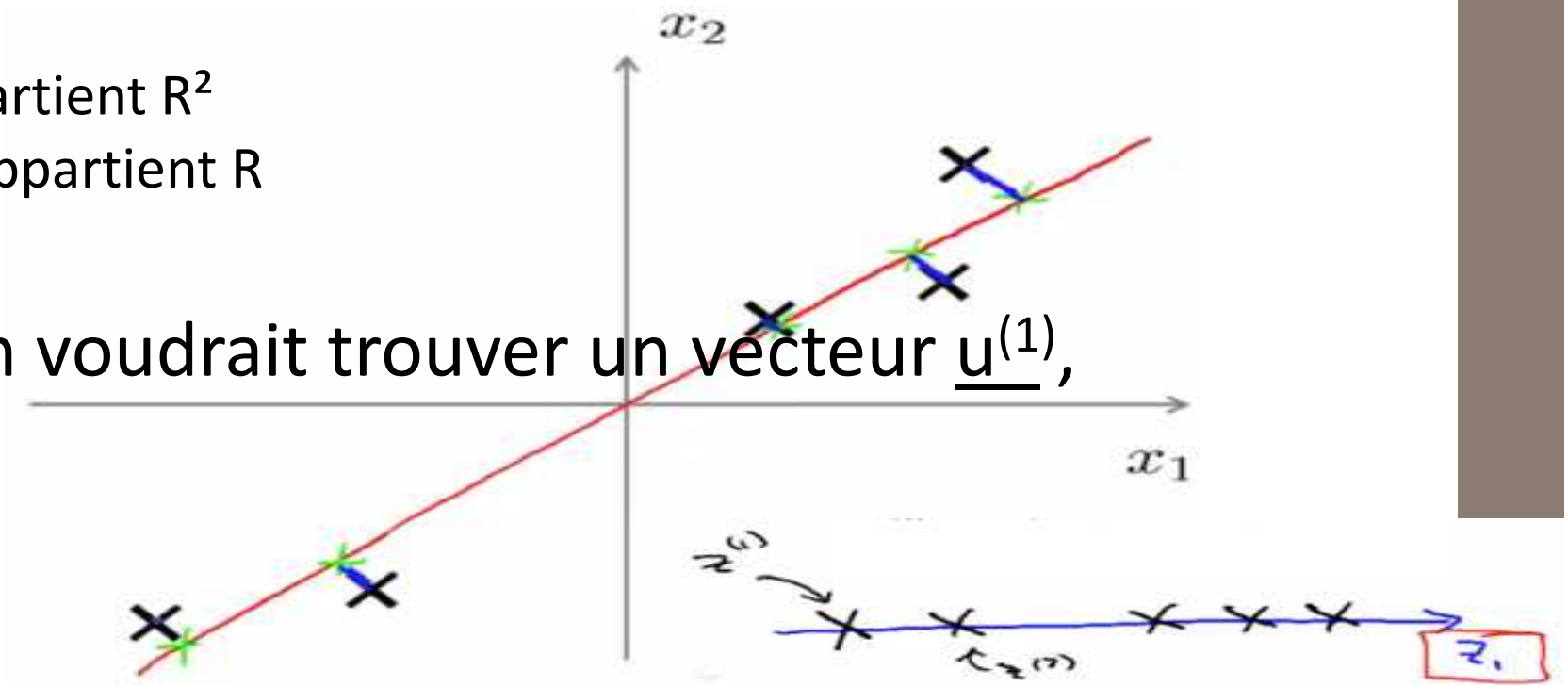
Analyse en Composantes Principales :

Formulation du problème :

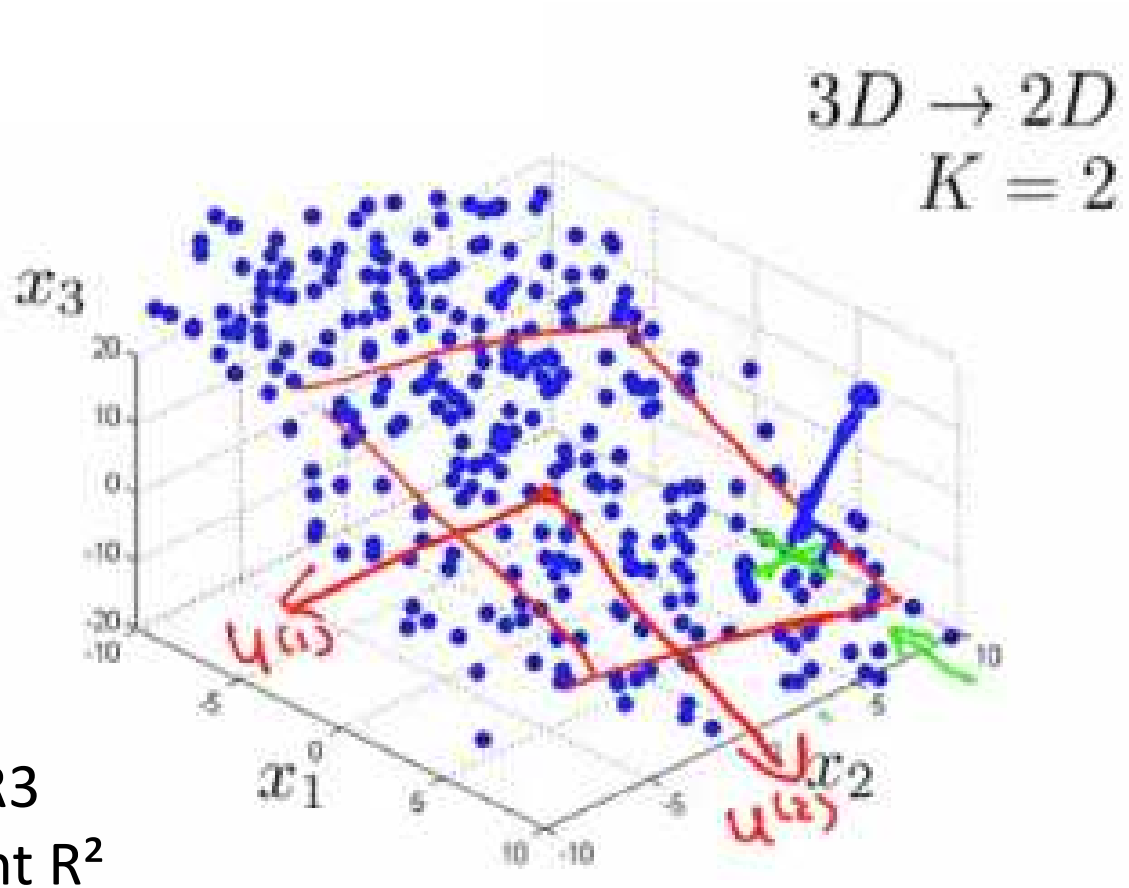
Réduire la Dimension de 2D à 1D (k=1) = Trouver une droite sur laquelle projeter les données de sorte que la distance entre chaque point et le point projeté **est assez petite.**

xi appartient R^2

→ zi appartient R



On voudrait trouver un vecteur $u^{(1)}$,



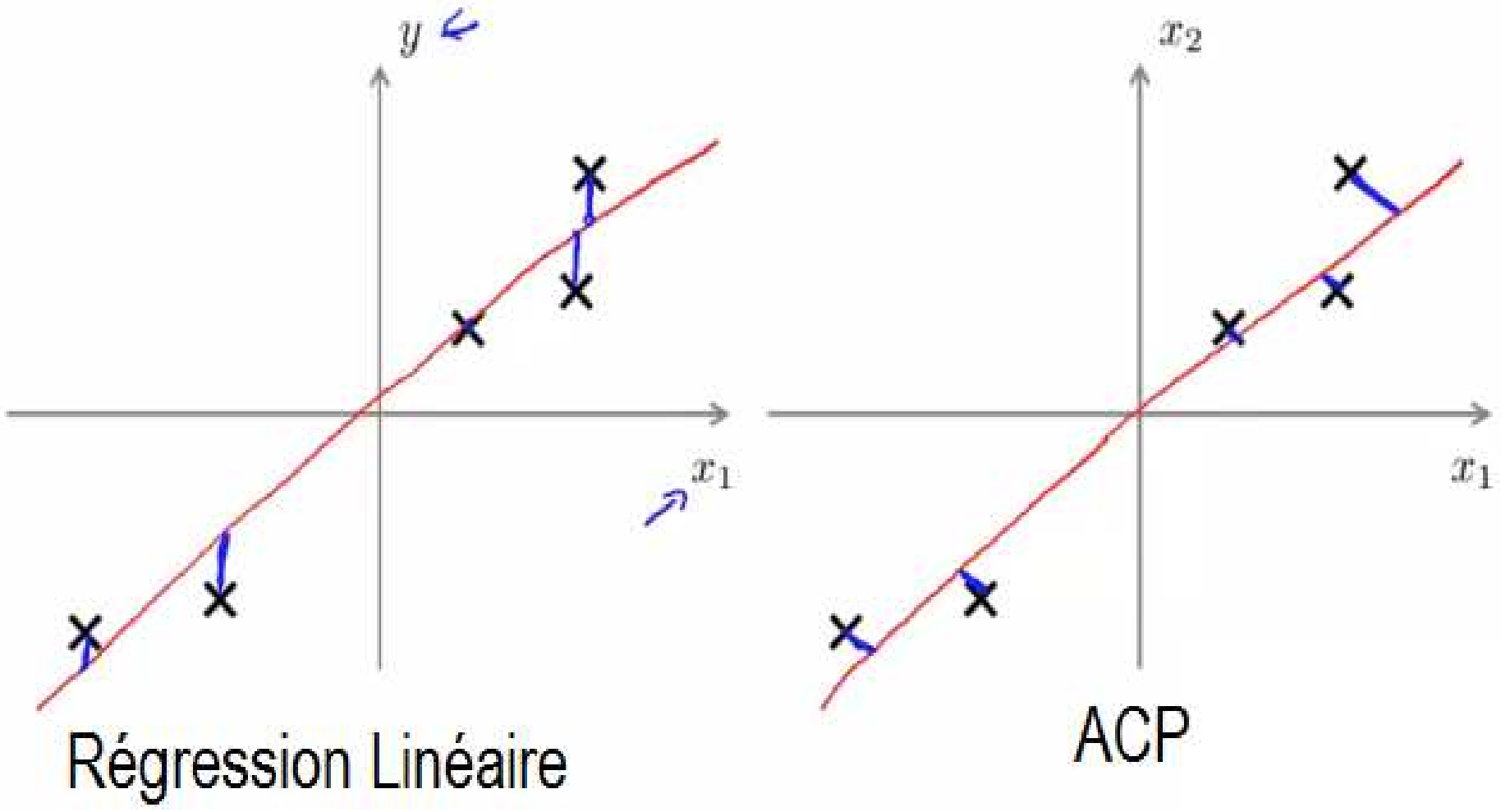
xi appartient R3

→ zi appartient R^2

Généralement on doit trouver k vecteurs $u^{(k)}$ sur lesquelles nous allons projeter les points de façon à réduire l'erreur de projection.

REMARQUE : D'un point de vue esthétique vous allez dire que l'ACP est une régression linéaire.

NON ; l'ACP n'est pas une régression linéaire :



Régression Linéaire	ACP
il y a des variables xi et une variable spéciale Y qu'on voudrait prédire depuis toutes les valeurs des xi,	toutes les variables sont traitées de la même façon il n'y pas de variable spéciale ni de variable qu'on voudrait prédire.
la projection des points est **diagonale** sur l'axe des X	la projection est quelconque le plus important est **qu'elle doit se faire au plus proche point**.

Se sont deux algorithmes totalement différents

Rappel :

Définition 1.

Une décomposition en valeurs singulières (SVD Singular Value Decomposition) est une factorisation $A^{(mxn)} = U\Sigma V$ **où**

- $U \in R^{(mxm)}$ est la matrice des vecteurs propres de AA^t
- $\Sigma \in R^{(mxn)}$ est Diagonale et les coefficients (Racine des valeurs propres):

$\sigma 1 \geq \sigma 2 \geq \ldots \geq \sigma p$, où $p=\min(m,n)$

- $V \in R^{(nxn)}$ est la matrice des vecteurs propres de A^tA

Rappel :

Théorème 1. Existence et unicité de SVD.
Toute matrice $A \in R^{(mxn)}$ possède ***une SVD***. Les valeurs singulières $\{\sigma i\}$ sont déterminées de façon unique.

Théorème 2. Les valeurs singulières d'une matrice A sont les racines carrées des valeurs propres non-nulles de A^tA et AA^t

L'Algorithme ACP :

1-Prétraitement de données : Centrer et réduire

- Ensemble de données : $x^{(1)}, x^{(2)}, \ldots\ldots x^{(m)}$:

  i=1..m exemples et j=1..n variables

-➔Remplacer chaque $xj^{(i)}$; **xj :=(xj- μj)/σj**

Tel que **μj :** Moyenne de la variable j**; σj** : Ecart type

2- Calcul de la matrice de Covariance :Sigma= $\frac{1}{m} X^T X$

3- Calcul des Vecteurs Propres:

[U,S,V] = SVD(Sigma)

4- Ureduce=U(1 :k)

5- Z=Ureduce^T*X

Fin

$$U = \begin{bmatrix} | & | & & | \\ u^{(1)} & u^{(2)} & \cdots & u^{(n)} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n \times n}$$

k

ACP:

Reconstruction des données originales depuis la représentation compressée :

$$Z = U_{reduce}^{T} * X$$

$$== > U_{reduce} * Z = U_{reduce} * U_{reduce}^{T} * X$$

$$== > U_{reduce} * Z = X_{approx}$$

ACP:

Choix de k: le Nbre des Composantes Principales :

- ✓ [U,S,V]=SVD(sigma):
- ✓ S est une matrice carrée Diagonale.
- ✓ On choisie le plus petit k pour lequel la condition suivante est vérifiée :

Pour k donné : $\frac{\sum_{i=1}^{k} sii}{\sum_{i=1}^{n} sii} <= 0.01$

c.a.d 99% de la variance retenue

ACP:

Visualisation des données :

Afin de visualiser les données dans le nouvel espace il faut :

- ✓Calcul des coordonnées des points individus $\Psi = XU$
- ✓Calcul des coordonnées des points variables : $\Phi = \sqrt{\lambda i}$ Ui
- ✓Projeter les points et faire une interprétation.

ACP:

Interprétation du Graphique :

- ✓Vérifier le % de variance expliquée;
- ✓ Nommer les axes par rapport aux variables qui se rapprochent de chaque axe;
- ✓Vérifier la dépendance de variables;
- ✓Expliquer la dispersion des individus par rapport aux variables.

Exemple 1: Ei= Elève n°1..9

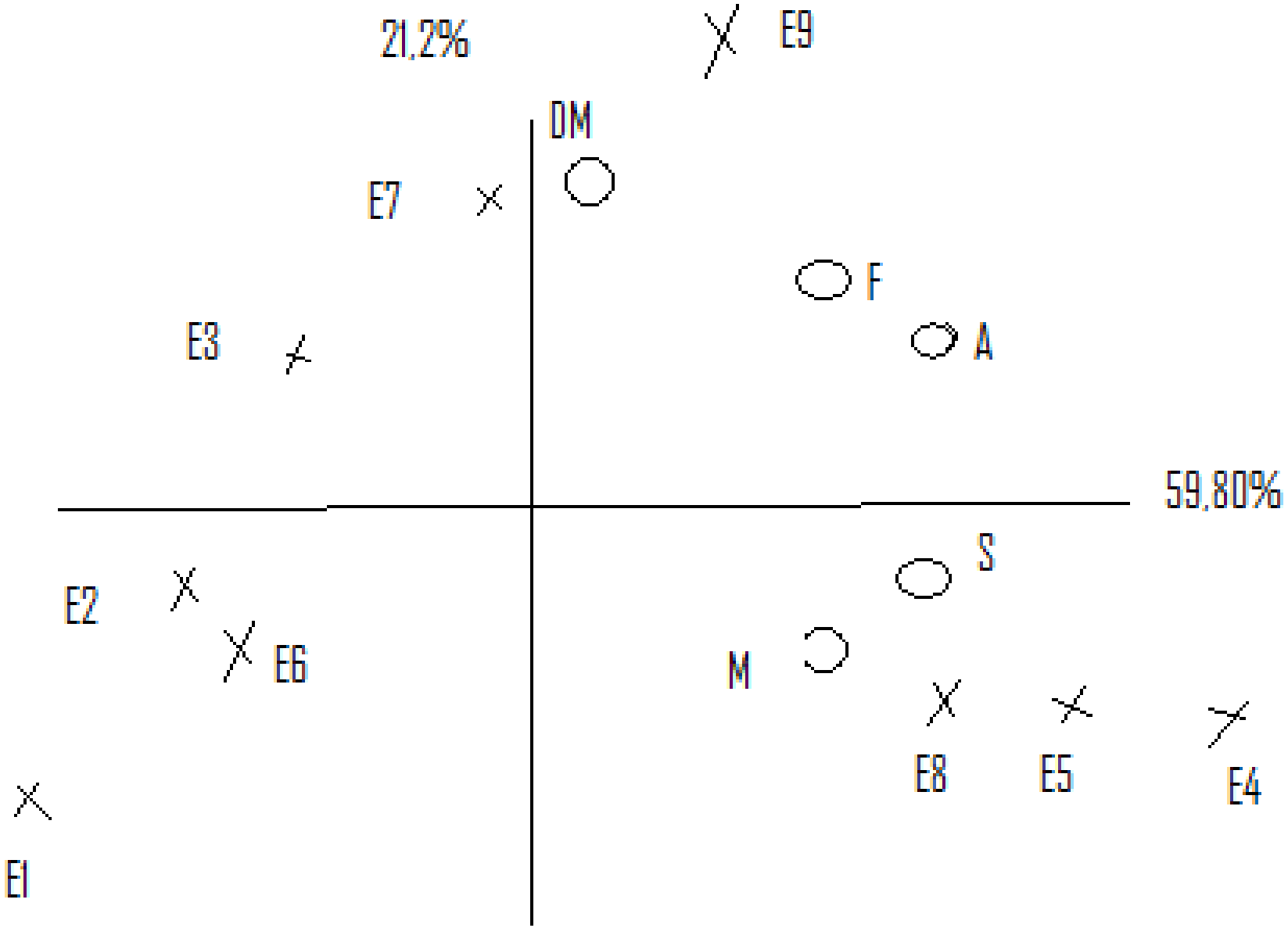
M=Math ; S=science ; F=Français ; A=Arabe ;

MD= Musique Dessin.

Sachant que le tableau de départ X est :

	M	S	F	A	DM
E1	6	6	5	6	8
E2	8	8	8	8	9
E3	6	7	11	10	11
E4	15	15	16	15	8
E5	14	14	12	13	10
E6	11	10	6	7	13
E7	7	7	14	12	10
E8	13	13	9	10	12
E9	9	10	13	12	18

ACP: Visualisation des Données



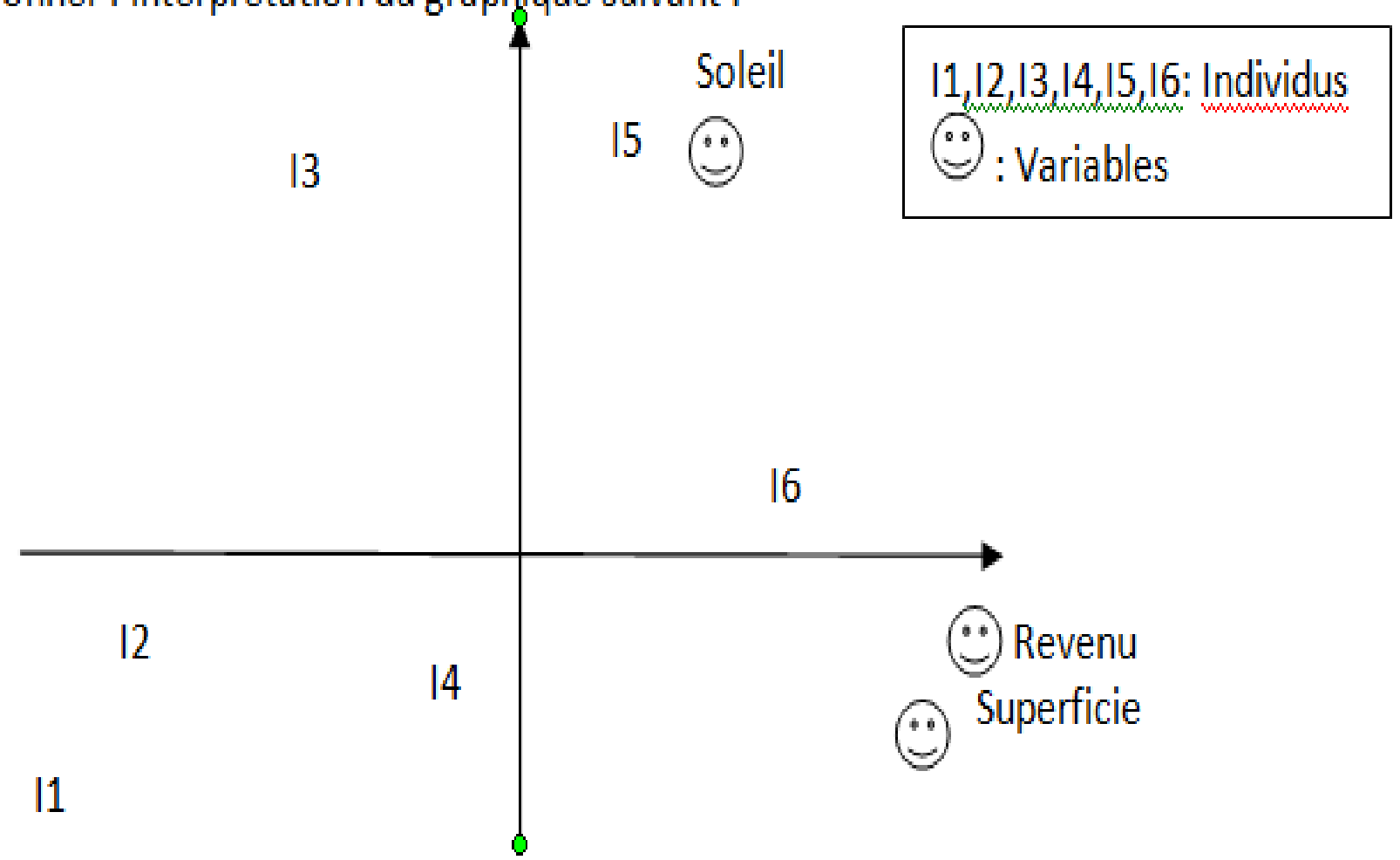
Exemple 2:

Un constructeur de piscine veut effectuer une étude de marché pour trouver les facteurs essentiels qui interviennent dans la décision d'achat de piscine qui sont revenu, et superficie jardin.
Son responsable marketing lui a confirmé qu'un autre facteur intervient dans cette décision qui est le soleil. Après recueil d'info sur 6 individus nous avons trouvé le tableau suivant

	Revenu	Superficie	Soleil
I1	9.3	30.2	3
I2	9.7	35.3	6
I3	10.3	40.1	10
I4	14.2	50.1	5
I5	18.6	60.2	10
I6	22 .7	68.4	7

ACP : Visualisation des Données

- Donner l'interprétation du graphique suivant :



ACP:

Remarque

- l'ACP peut très bien être utilisée afin de rendre rapide les algorithmes d'apprentissage supervisé : **MAIS**
- ceci ne doit pas être effectué tout le temps, il faut travailler normalement avant d'appliquer ACP et voir les résultats.

ACP:

Remarque :

- il ne faut pas utiliser l'ACP pour prévenir ou éviter le sur-apprentissage, même si elle réduit le nombre de variables,
- il faut utiliser la régularisation pour éviter l'over-fitting.