Ecole Supérieure en Sciences Appliquées de Tlemcen

Analyse de données

Chapitre 4: Analyse en composantes principales

Partie 1

Présentée par:

Dr Imane NEDJAR

Introduction

Analyses de données

Les modèles statistiques

Sont utilisés pour nettoyer les données au début par l'élimination des valeurs aberrantes, et aussi de visualiser les données, afin de construire l'ensemble initial d'exemples.

Les modèles factorielles

Cherchent à réduire le nombre de variables en les résumant par un petit nombre de composantes synthétiques en utilisant essentiellement des outils de l’algèbre linéaire.

Les modèles classification

Analyse en composantes principales (ACP)

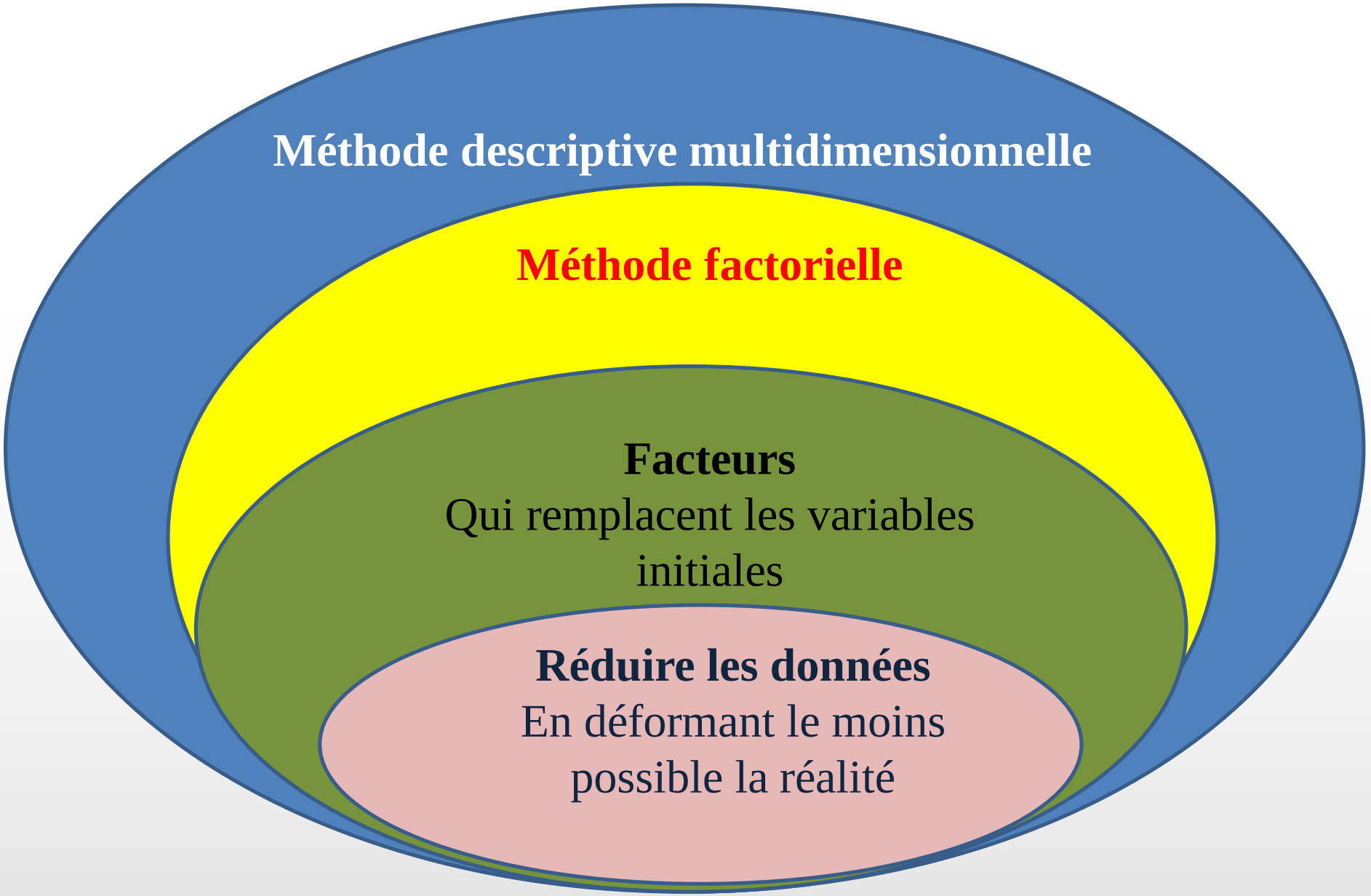
•ACP est l'une des méthodes d'analyse de données **Multivariées**. Permettant d'explorer des données **Multidimensionnels** constitués de variables **Quantitatives**.

•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)



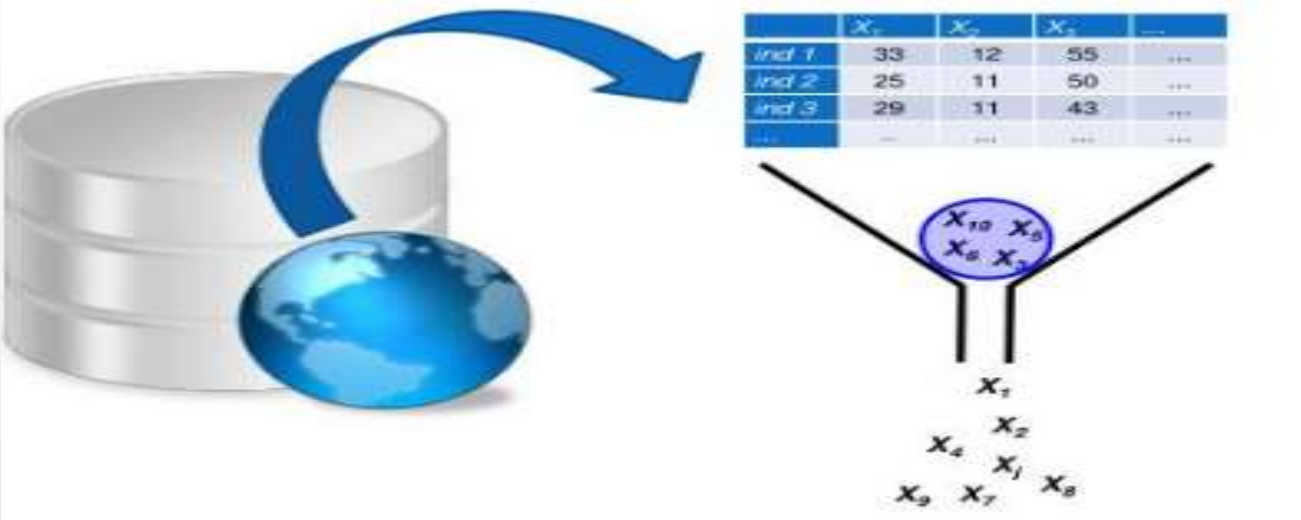
Analyse en composantes principales (ACP)

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



Application de l'ACP

➢**Analyse de données :**

•Réduction du nombre de variables explicatives (Xj) avant modélisation

• Obtention de nouvelles variables explicatives non corrélées

Illustration graphique de l'ACP

Soit le tableau génétique suivant :

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1

Si on mesure seulement gène 1 ➔ les données sont projetées sur un seul axe

Illustration graphique de l'ACP

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1

…Les souris 1,2 et 3 ont relativement des valeurs élevées

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1

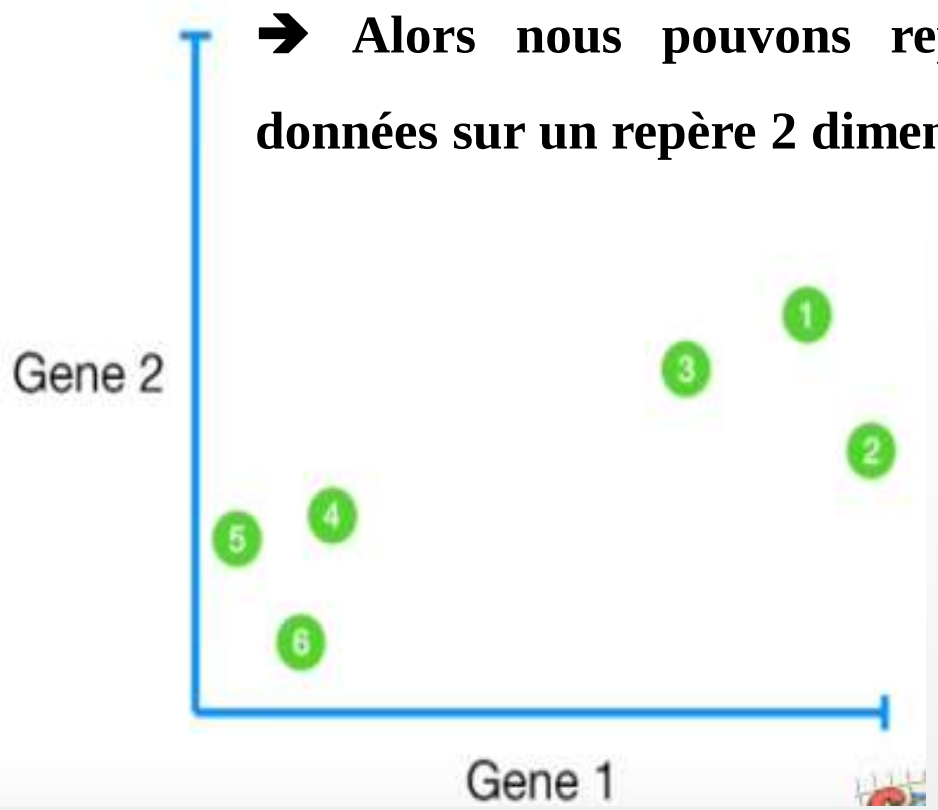
…Les souris 4,5 et 6 ont relativement des faibles valeurs

➔même s'il s'agit d'un simple graphique, il nous montre que les souris 1, 2 et 3 sont plus semblables les unes aux autres que les souris 4, 5 et 6.

Illustration graphique de l'ACP

- Si nous mesurons 2 gènes ?

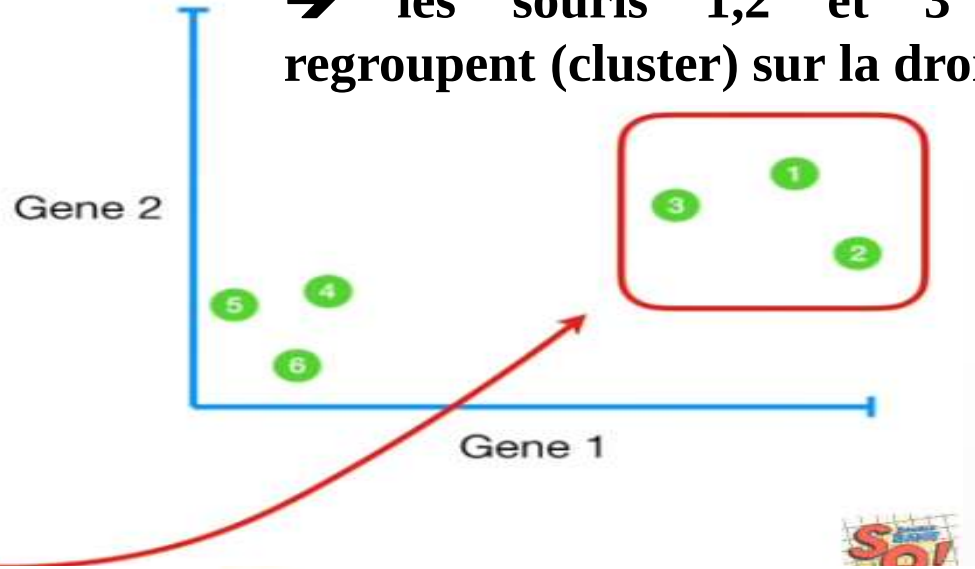
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



➔ **Alors nous pouvons représenter les données sur un repère 2 dimensions X/Y**

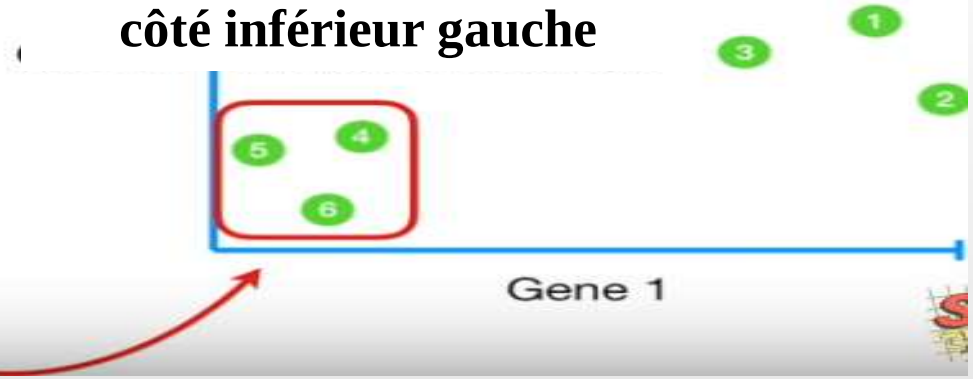
Illustration graphique de l'ACP

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



➔ les souris 1,2 et 3 se regroupent (cluster) sur la droite

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



➔ les souris 4,5 et 6 se regroupent (cluster) sur le côté inférieur gauche

Illustration graphique de l'ACP

- Si nous mesurons 3 gènes ?

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2

➔**Repère avec 3 dimension**

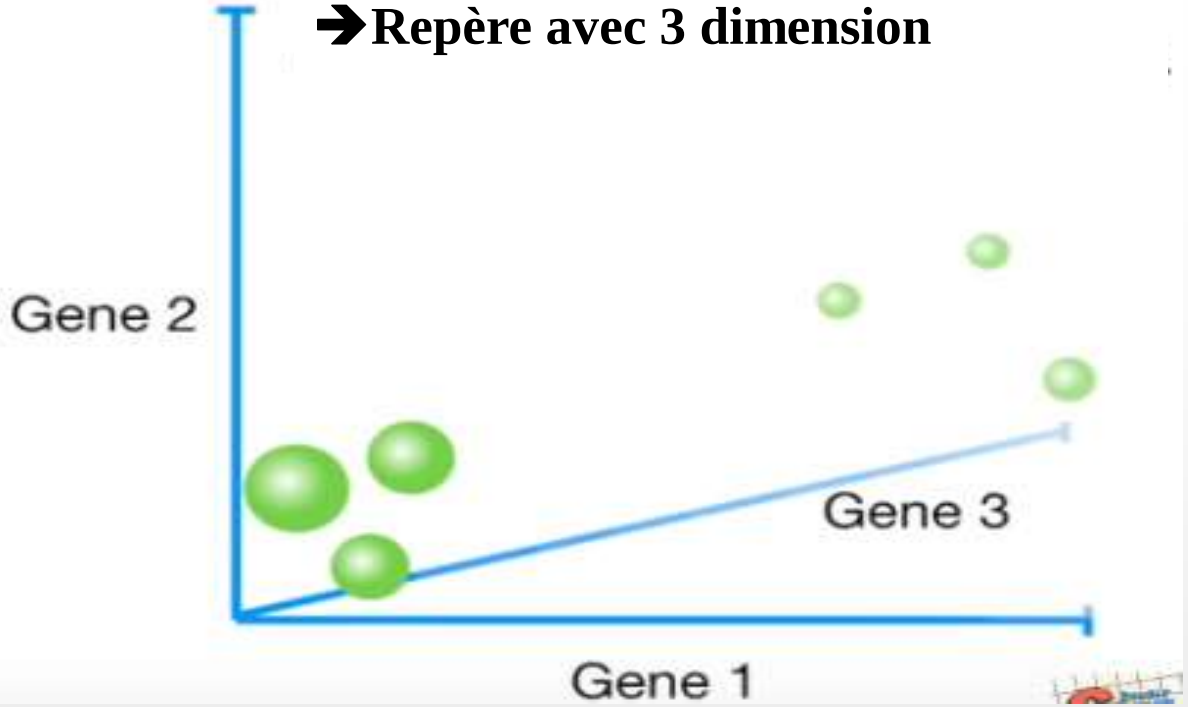
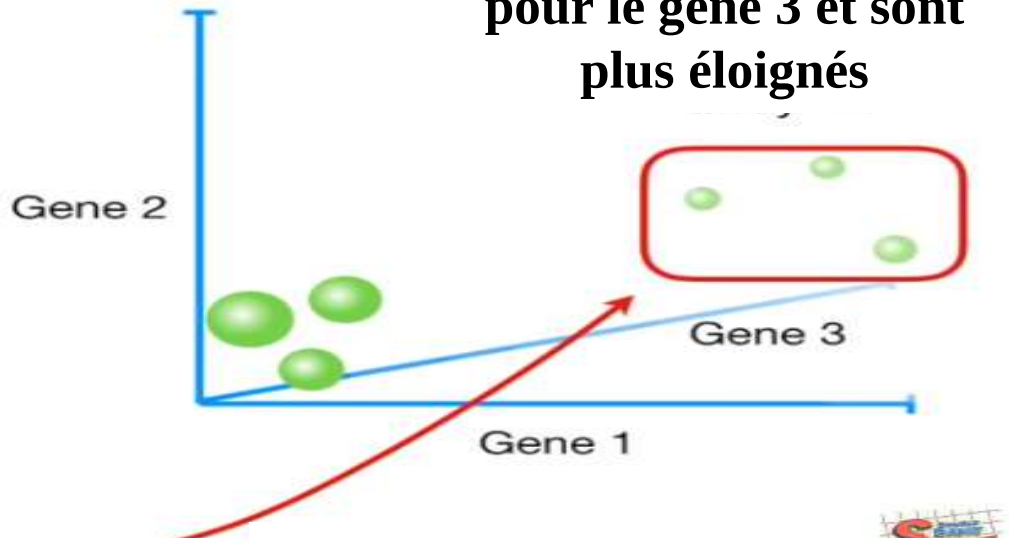


Illustration graphique de l'ACP

Les points plus petits ont des valeurs plus grandes pour le gène 3 et sont plus éloignés

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2



Les grands points ont des valeurs plus petites pour le gène 3 et sont plus proches

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2

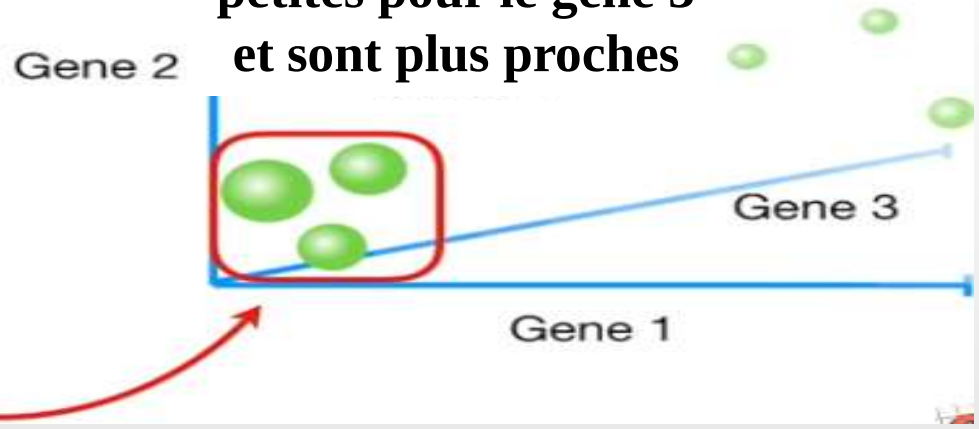


Illustration graphique de l'ACP

- Si nous mesurons 4 gènes ?

➔ Impossible de représenter les données sous forme de graphique nuage de points (4 dimensions !)

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

Illustration graphique de l'ACP

ACP peut prendre 4 variables de type gènes ou plus +

➔ par la suite établir une représentation graphique ACP en 2-D

➔ Cette représentation permet de voir que des souris similaires se regroupent.

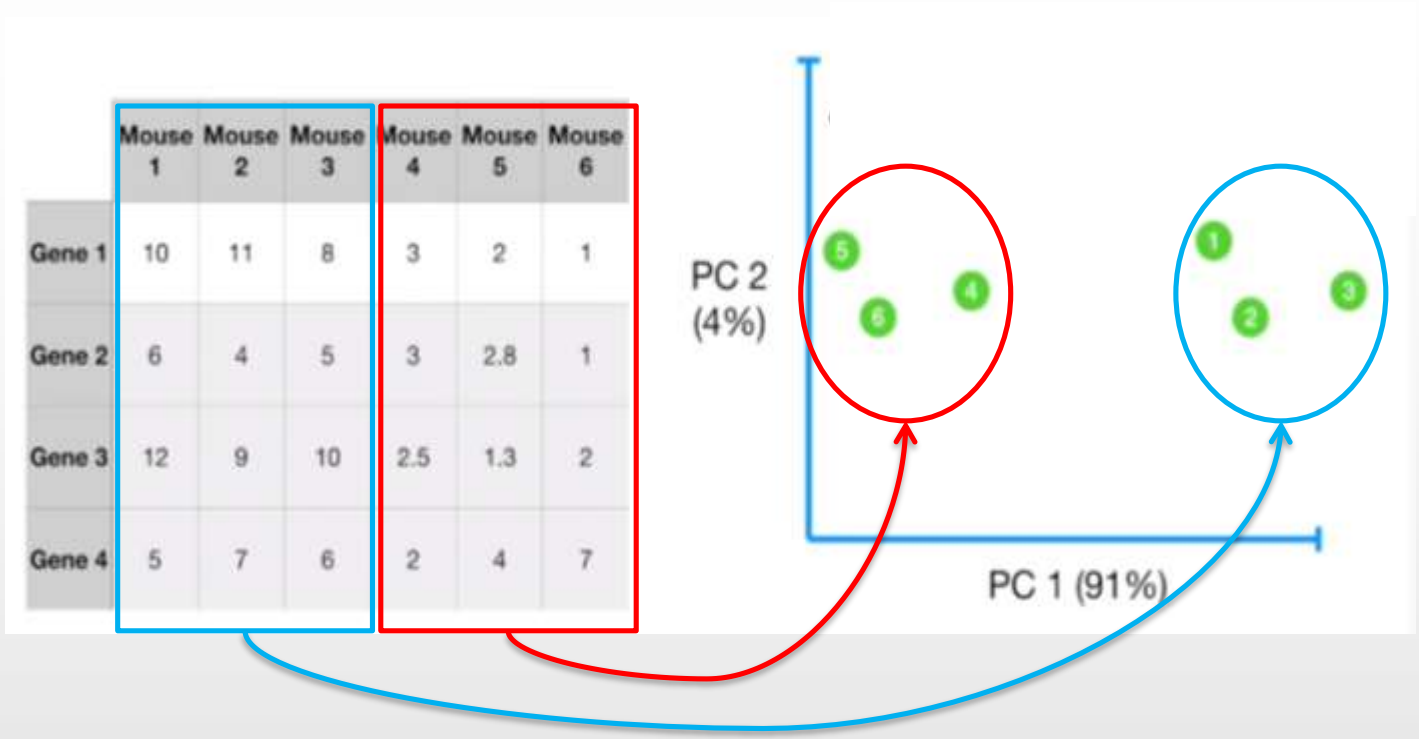
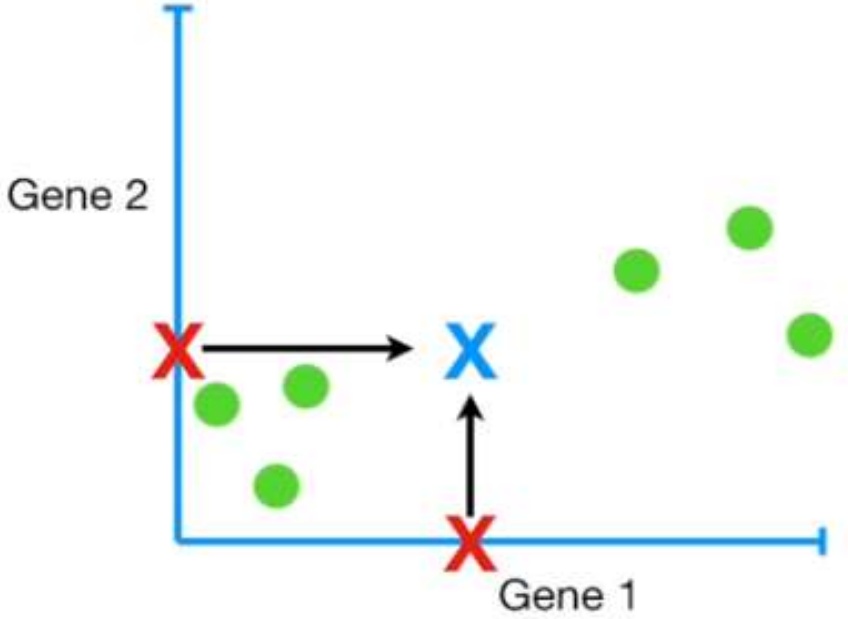


Illustration graphique de l'ACP

- Pour comprendre l'ACP, on va reprendre le tableau 2-D
- Avec les moyennes de variables, nous calculons le centre de gravité du nuage = le point « moyen »

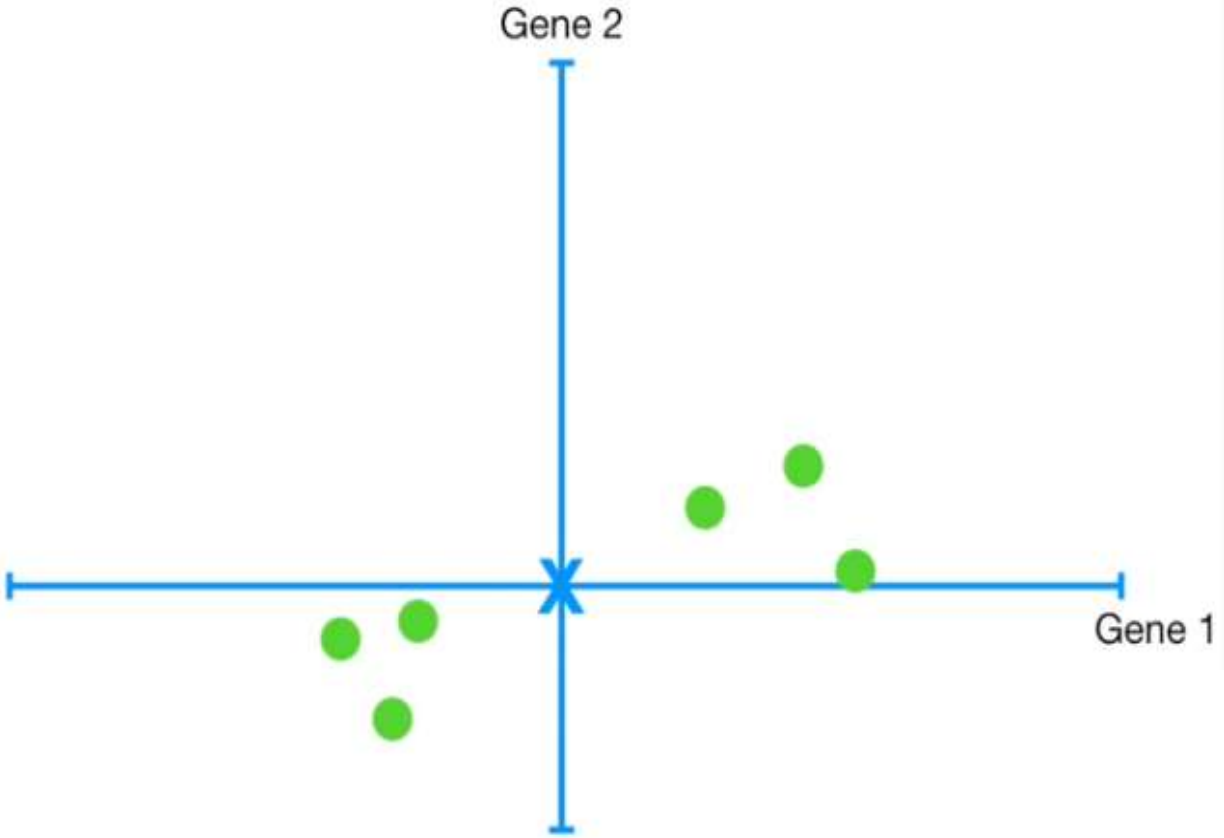
  ➔ maintenant nous utiliserons pas le tableau initial

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



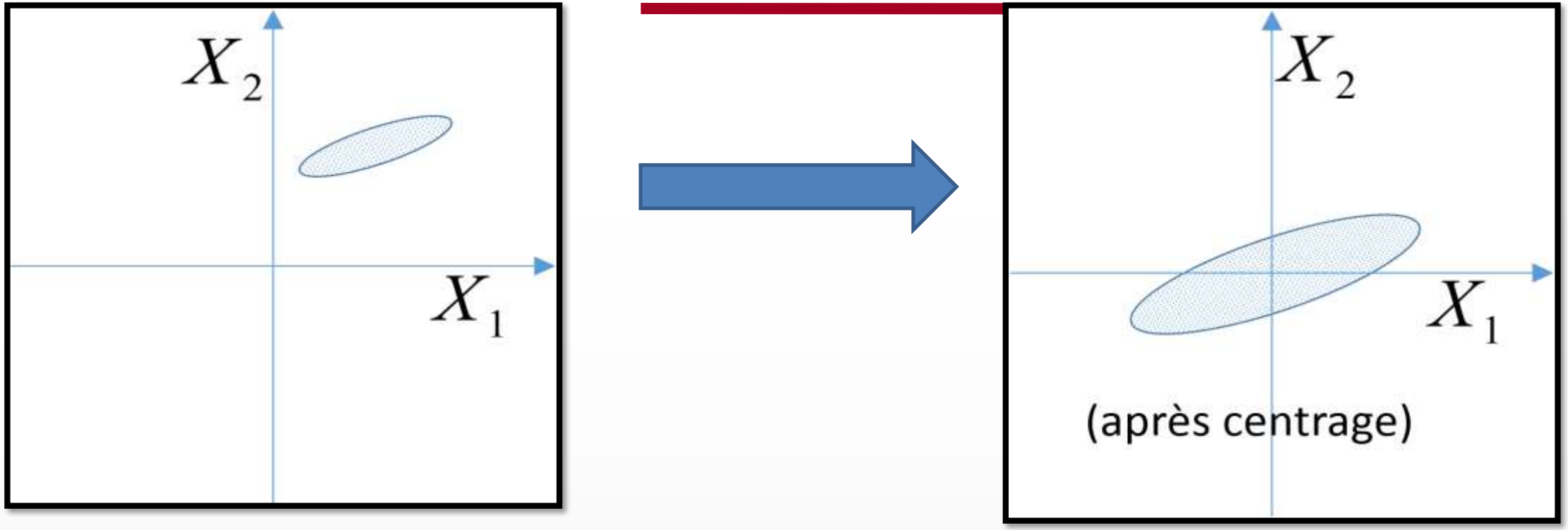
Prétraitement = Centrer les données ne modifie pas la forme du nuage

Illustration graphique de l'ACP



Centrer les données revient juste à translater le nuage

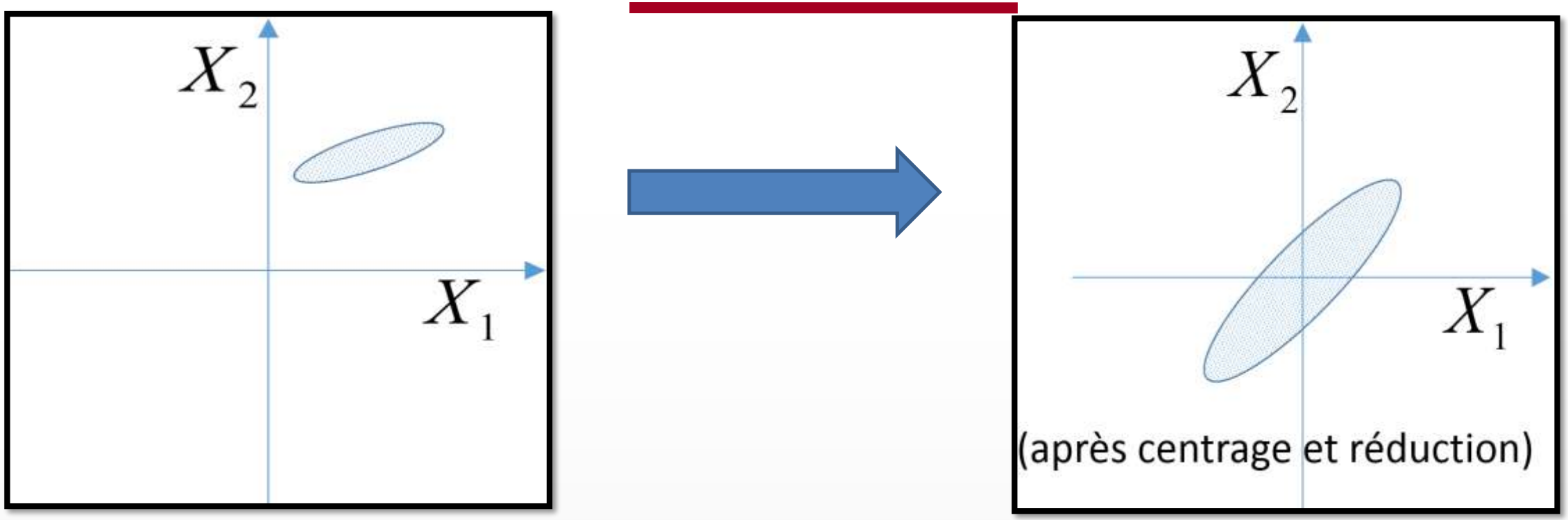
Illustration graphique de l'ACP



ACP centrée

Centrage préalable des variables. Cela revient à s'intéresser à la *forme* du nuage d'individus par rapport à son centre de gravité. Cette variante est utilisée lorsque les variables initiales sont directement comparables (de même nature, intervalles de variation comparables).

Illustration graphique de l'ACP



ACP centrée et réduite

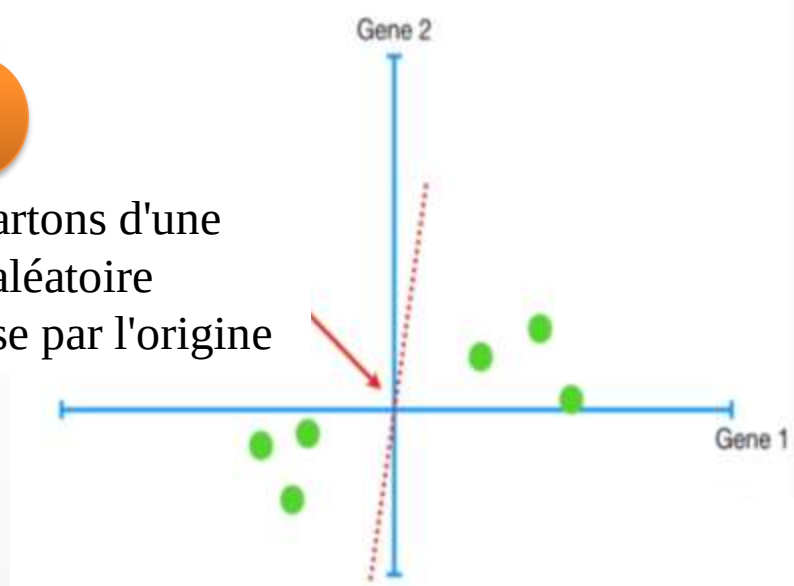
Réduction préalable des variables. On s'intéresse donc à la forme du nuage d'individus après centrage et réduction des variables. Cette normalisation est employée lorsque les variables (toutes quantitatives) sont de nature différente ou présentent des intervalles de variation très différents.

Illustration graphique de l'ACP

- Après centrage des données ➔ essayer d'ajustez une droite dessus

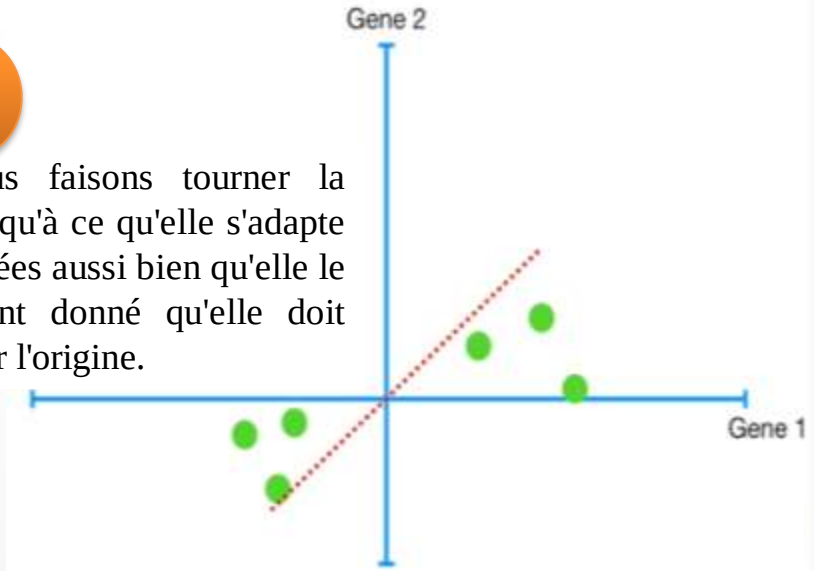
1

Nous partons d'une droite aléatoire qui passe par l'origine



2

Puis nous faisons tourner la droite jusqu'à ce qu'elle s'adapte aux données aussi bien qu'elle le peut, étant donné qu'elle doit passer par l'origine.



3

Finalement, cette droite est plus appropriée.

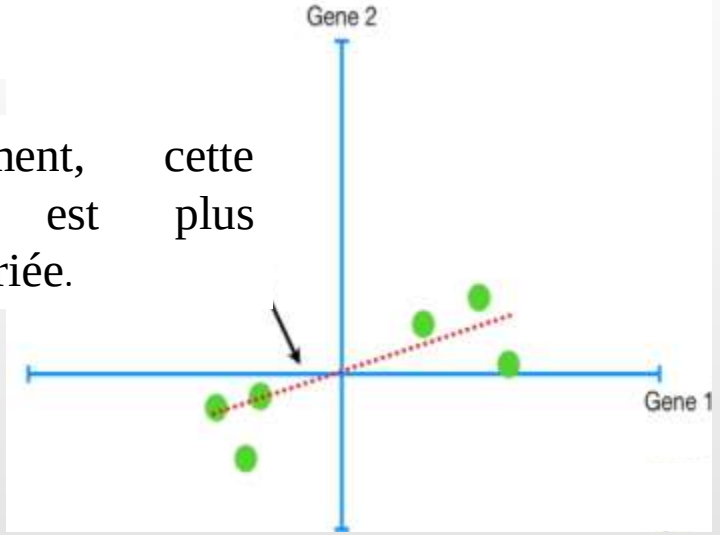
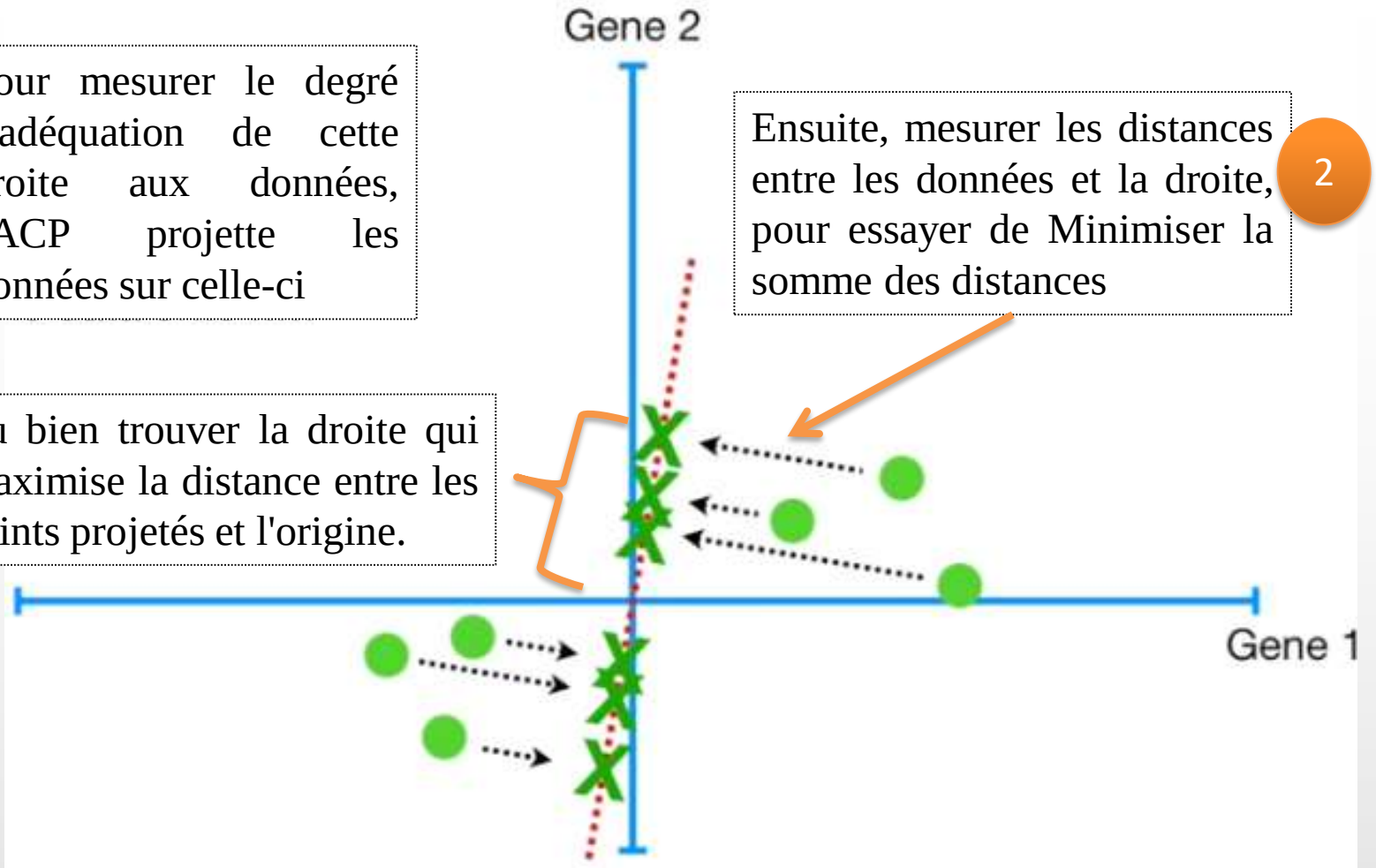


Illustration graphique de l'ACP

- Comment l'ACP trouve la meilleure droite ?

1 Pour mesurer le degré d'adéquation de cette droite aux données, l'ACP projette les données sur celle-ci

2 Ou bien trouver la droite qui Maximise la distance entre les points projetés et l'origine.



2 Ensuite, mesurer les distances entre les données et la droite, pour essayer de Minimiser la somme des distances

Illustration graphique de l'ACP

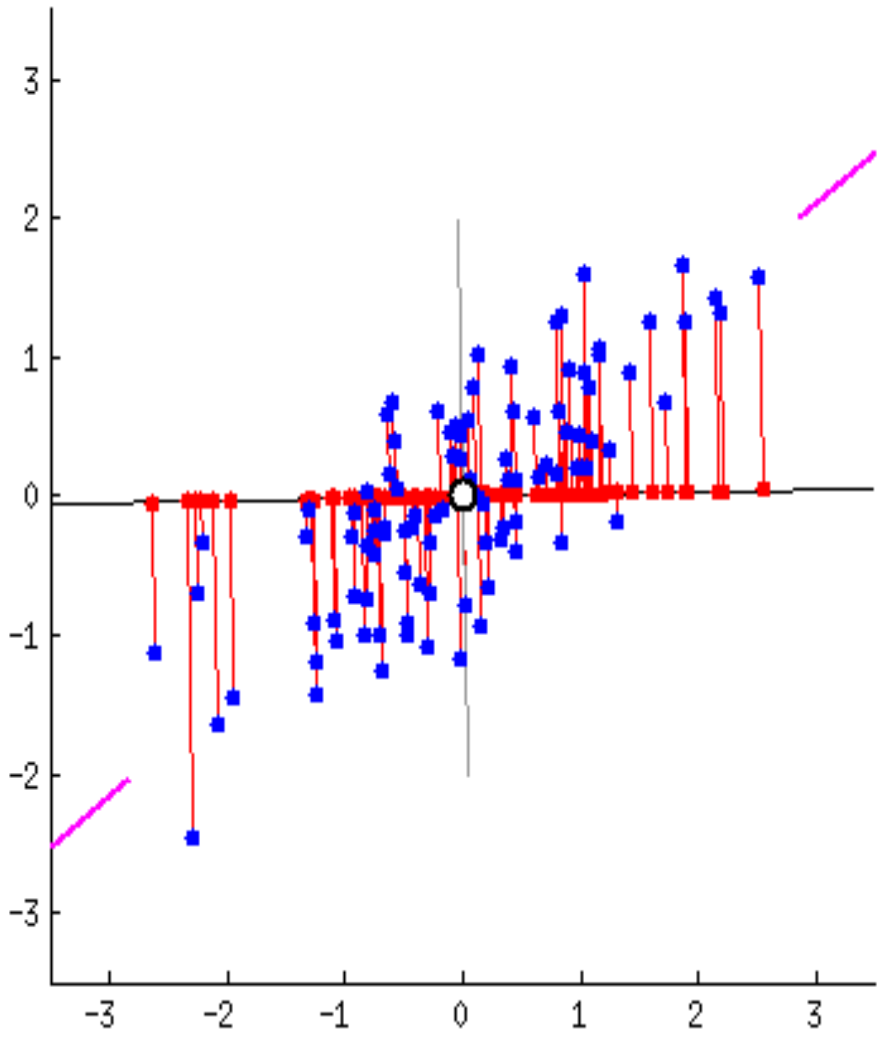
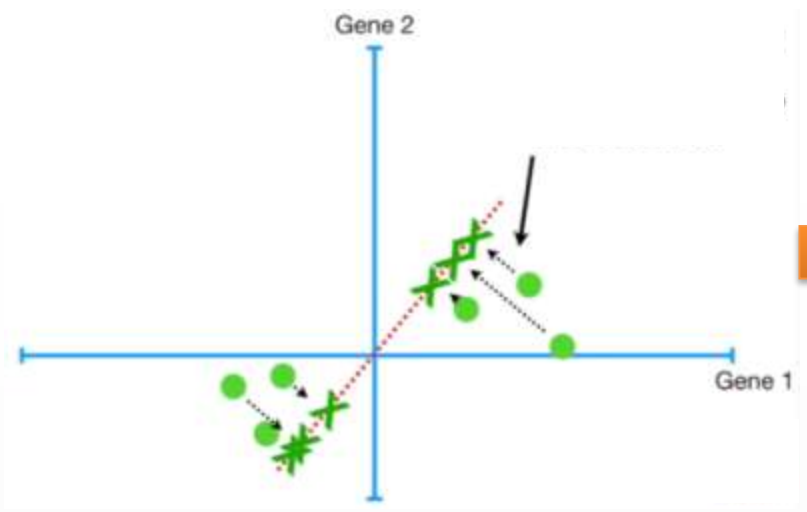
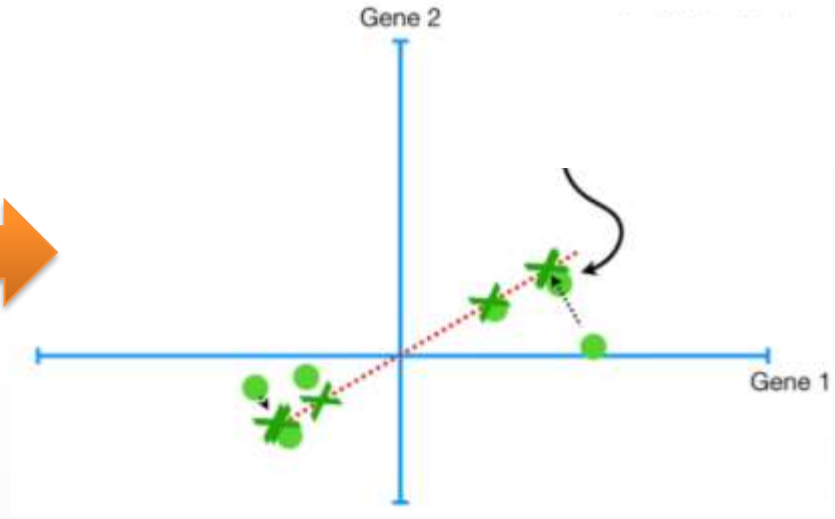
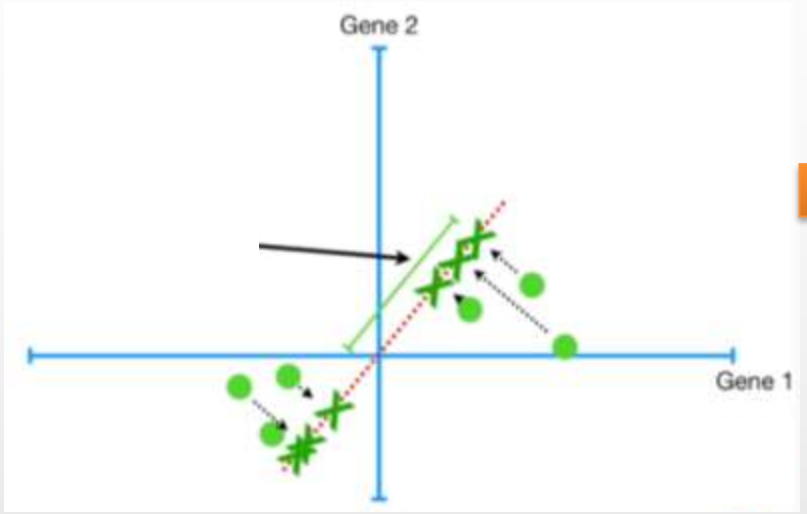


Illustration graphique de l'ACP



min





max

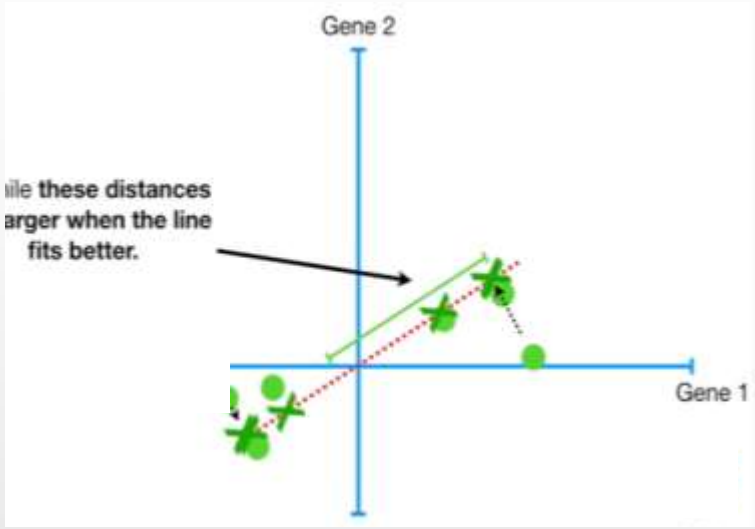
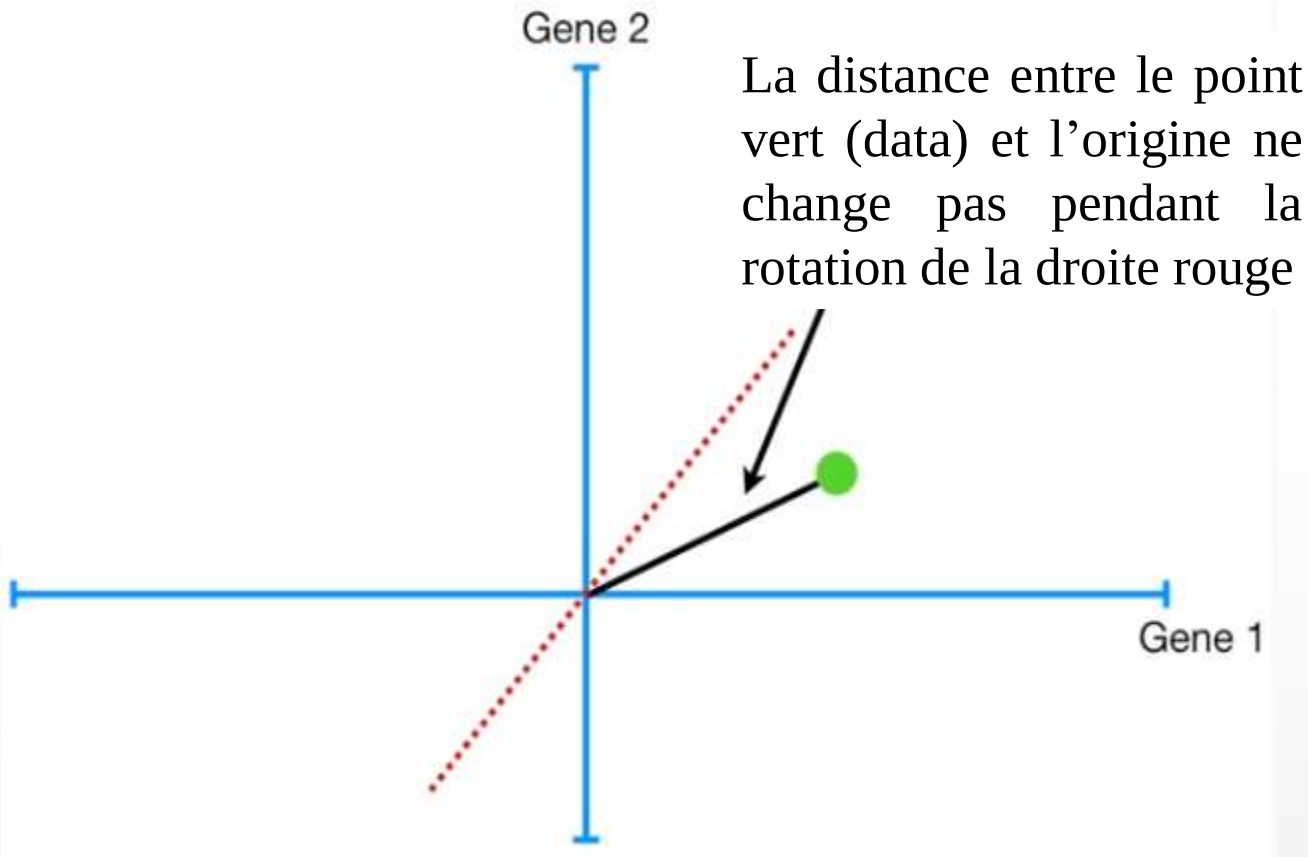


Illustration graphique de l'ACP

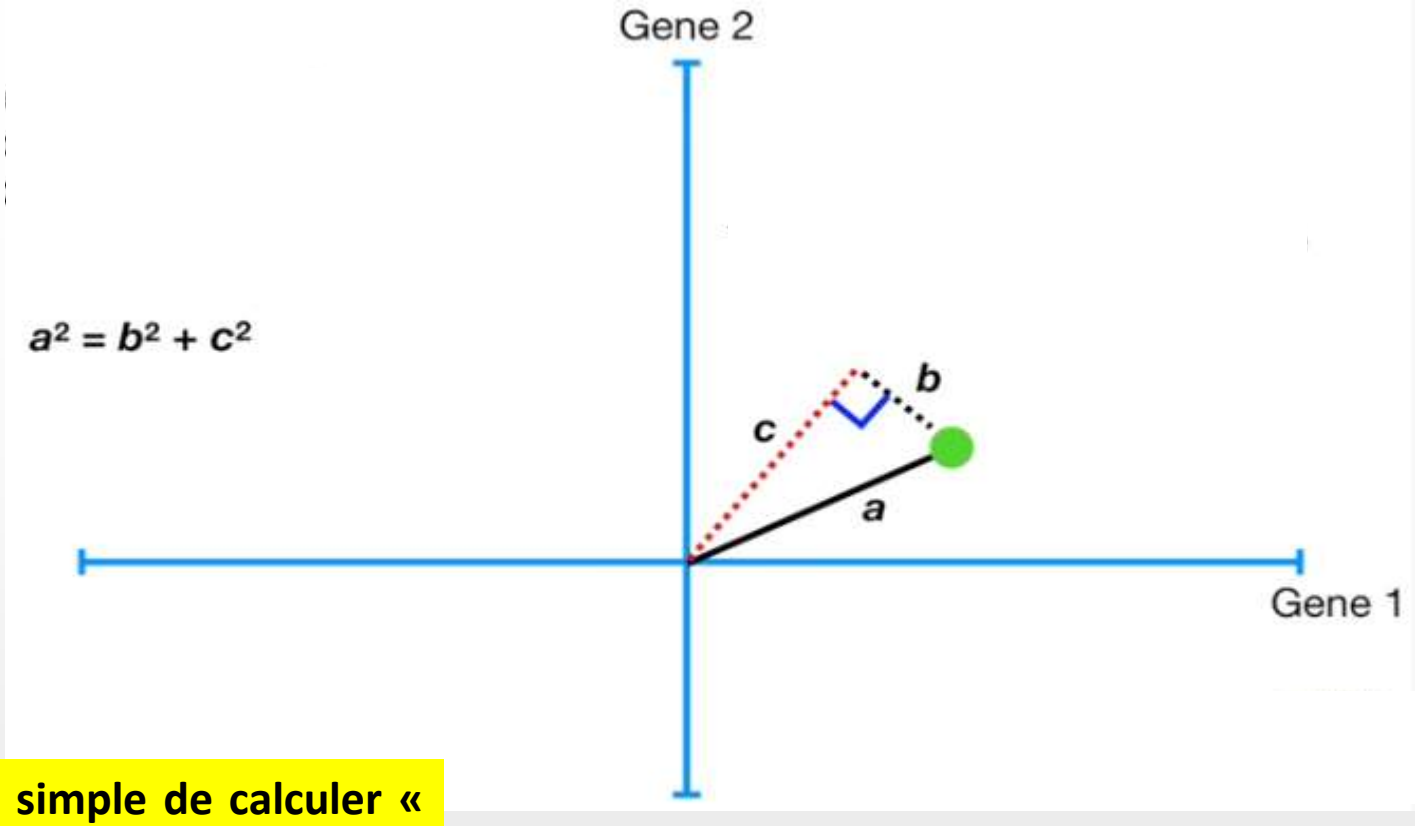


La distance entre le point vert (data) et l'origine ne change pas pendant la rotation de la droite rouge

Illustration graphique de l'ACP

- Utilisant le théorème de Pythagore pour montrer la relation inverse entre « b » et « c »

➔ l'ACP peut minimiser la distance « b » ou bien maximiser la distance « c »



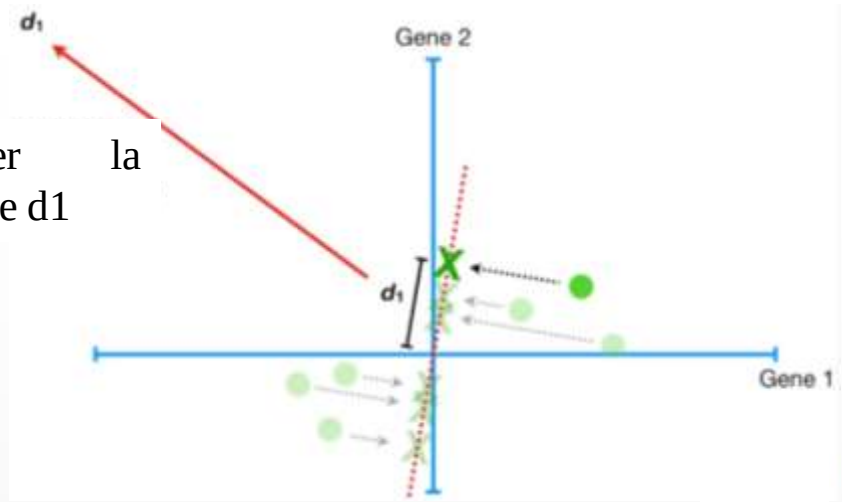
Il est plus simple de calculer « c » que « b »

Illustration graphique de l'ACP

- Généralement l'ACP maximise la somme des distances quadratiques des points projetées à l'origine ➔ «c »

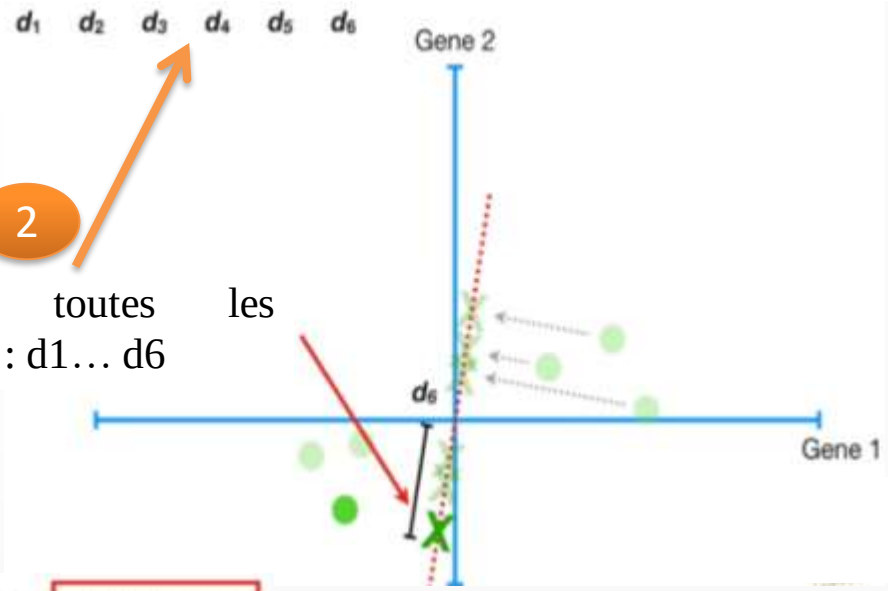
1

Calculer la distance d1



2

Calculer toutes les distances : d1… d6



3

Distances Quadratiques

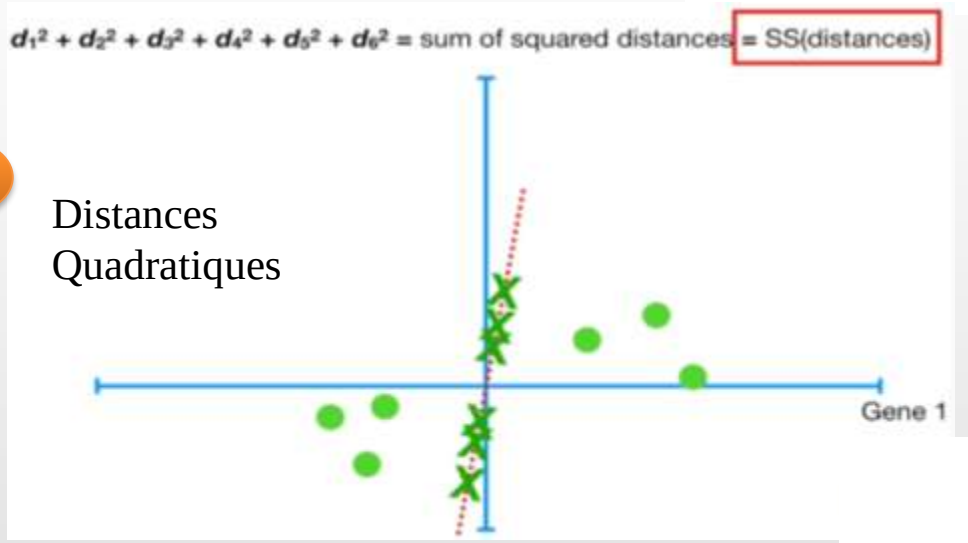


Illustration graphique de l'ACP

$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ =somme des carrés des distances (Sum of squared distance)

SS(distances)

4

Nous arrivons finalement à cette droite qui maximise SS distance

Appelée : **Composante Principale 1 (PC1)**

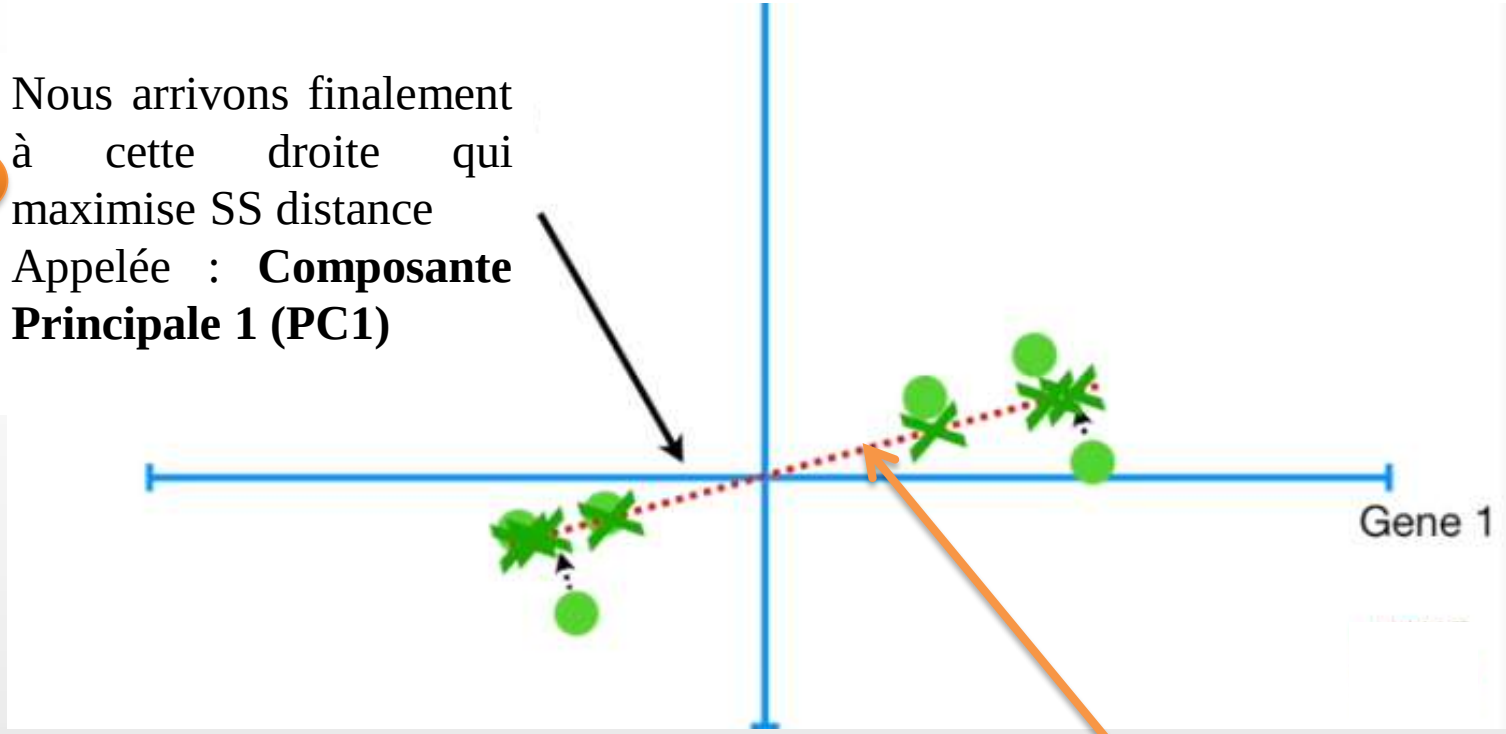
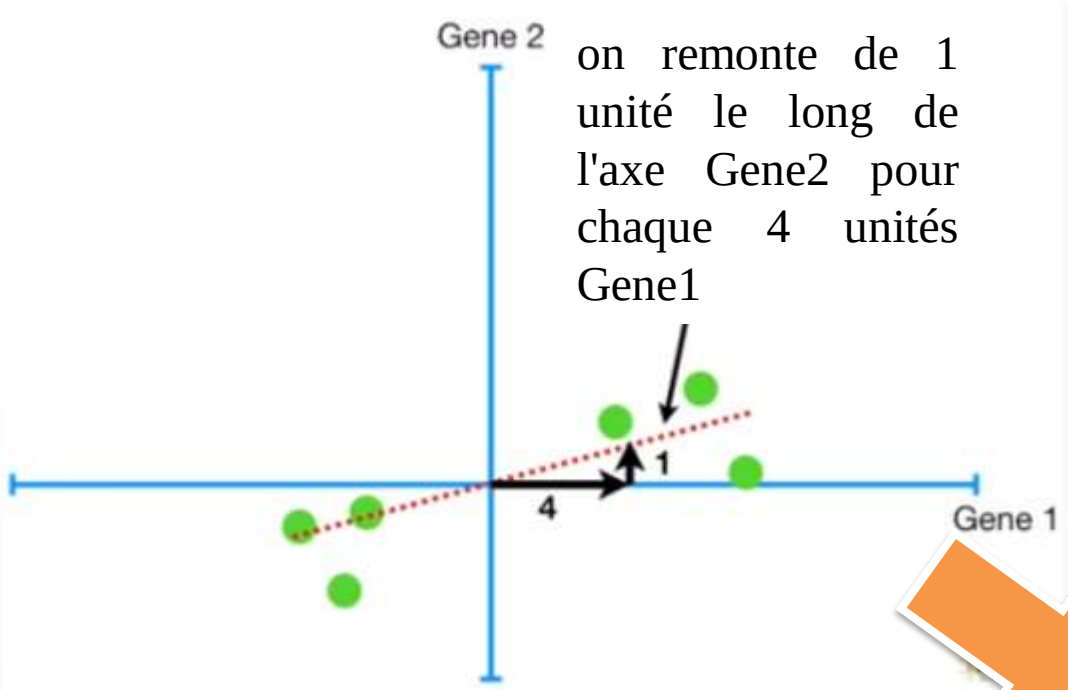


Illustration graphique de l'ACP



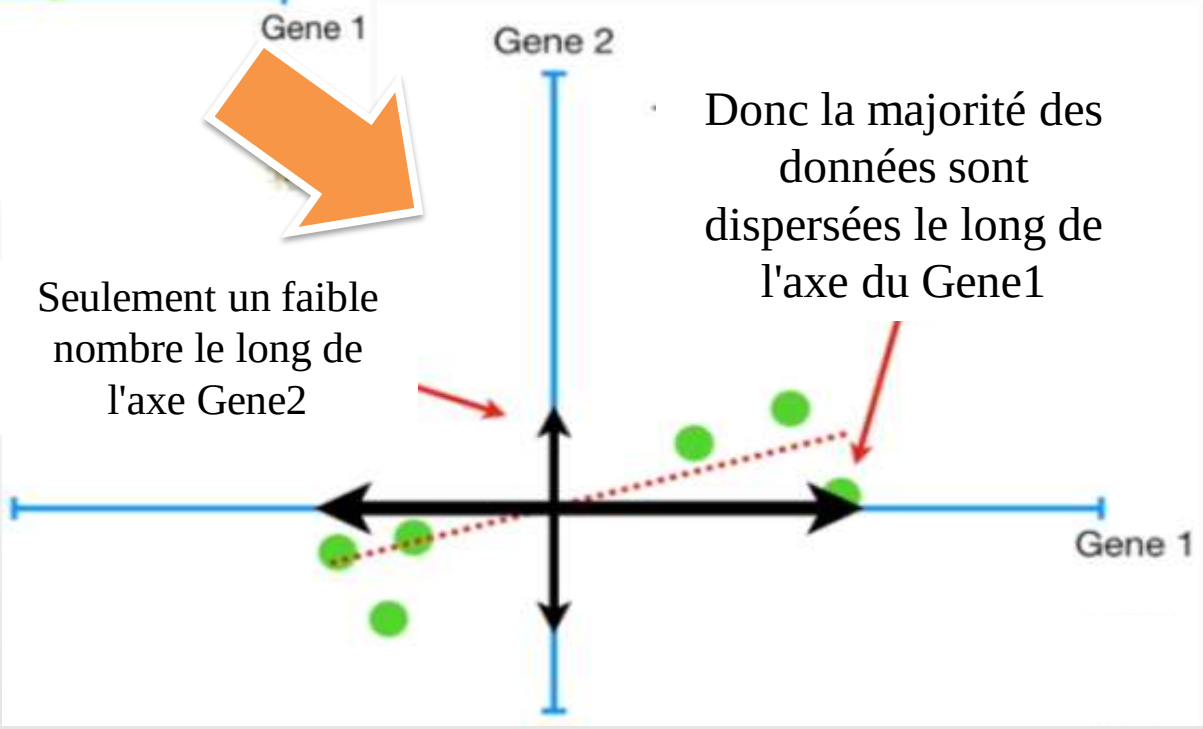
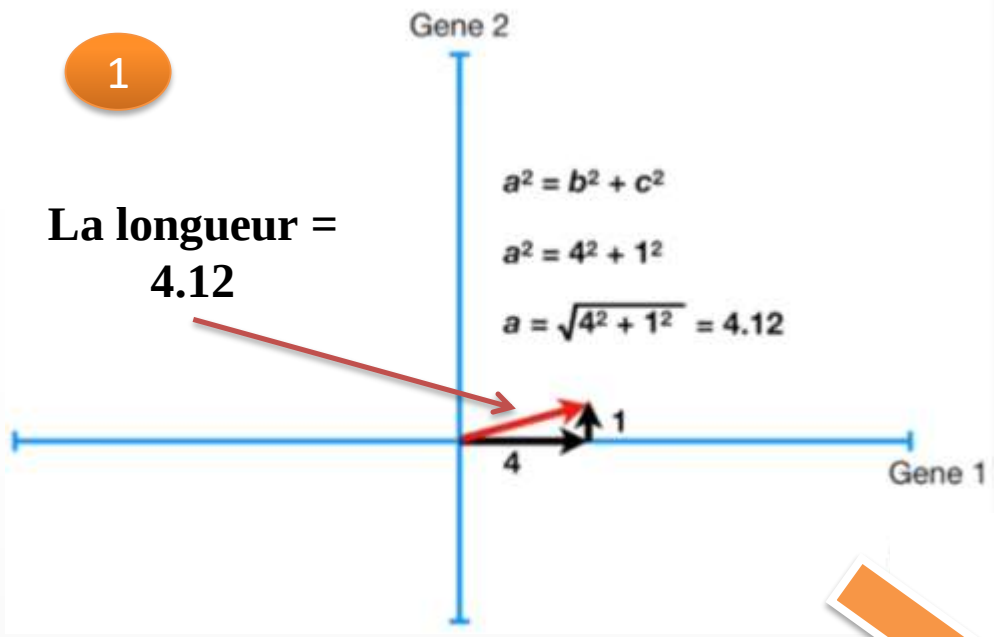
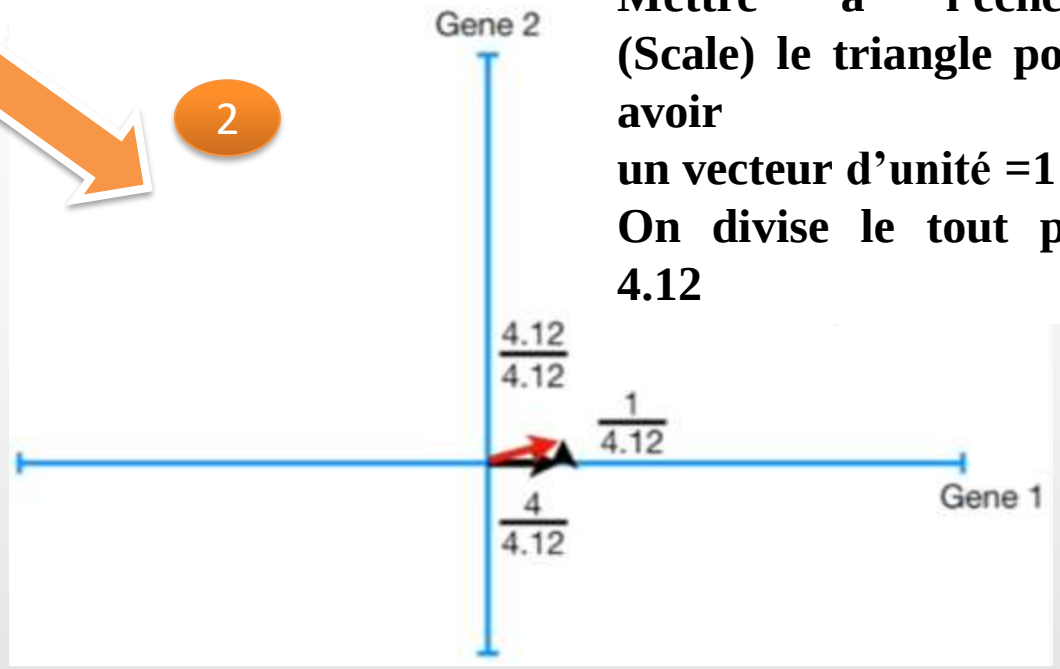


Illustration graphique de l'ACP





Mettre à l'échelle (Scale) le triangle pour avoir
un vecteur d'unité =1
On divise le tout par 4.12

Illustration graphique de l'ACP

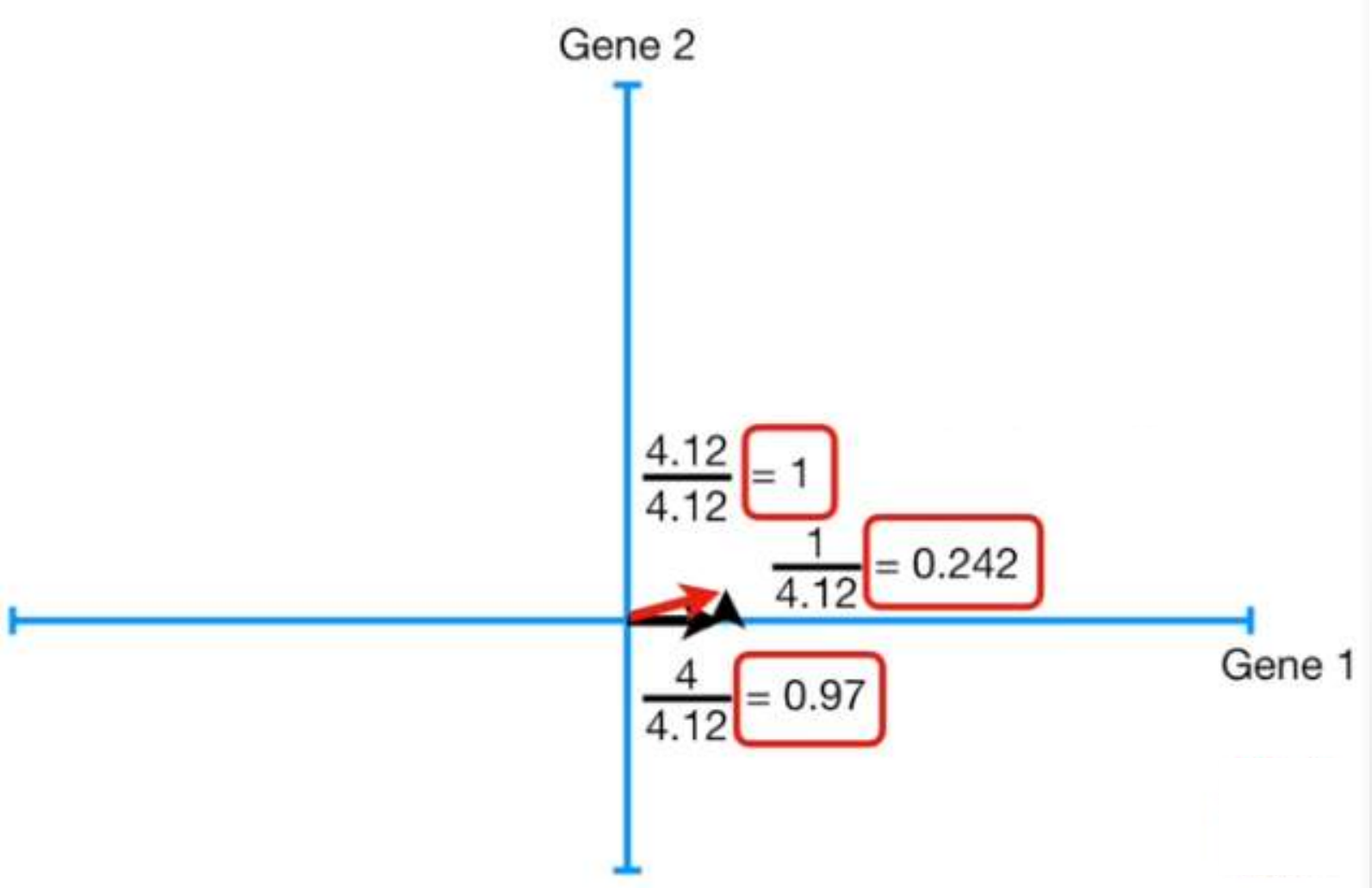
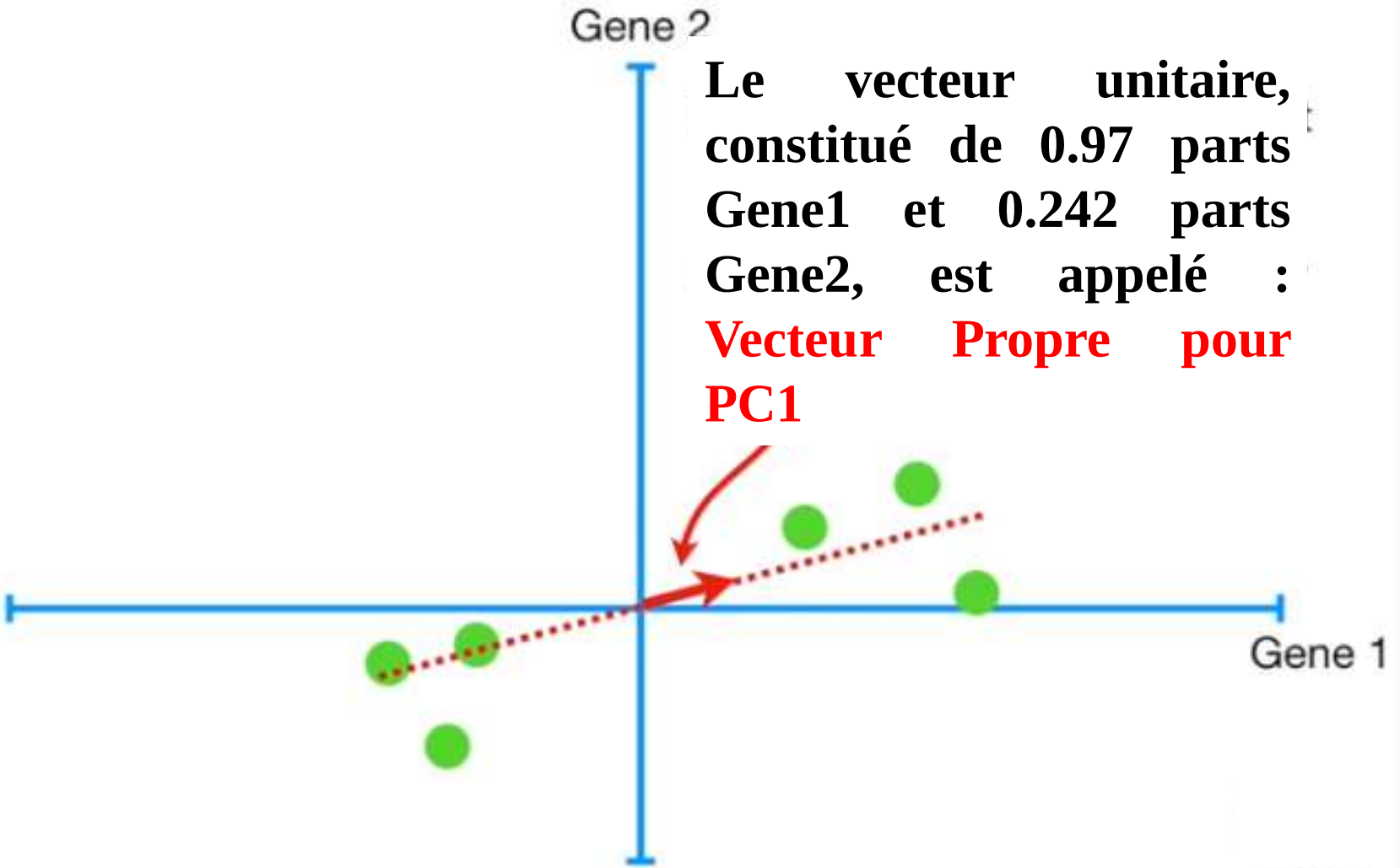


Illustration graphique de l'ACP



Le vecteur unitaire, constitué de 0.97 parts Gene1 et 0.242 parts Gene2, est appelé : Vecteur Propre pour PC1

Illustration graphique de l'ACP

$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ =somme des carrés des distances (Sum of squared distance)
SS(distances)

SS(distances de PC1) = Valeur Propre de PC1

EIGENVALUE

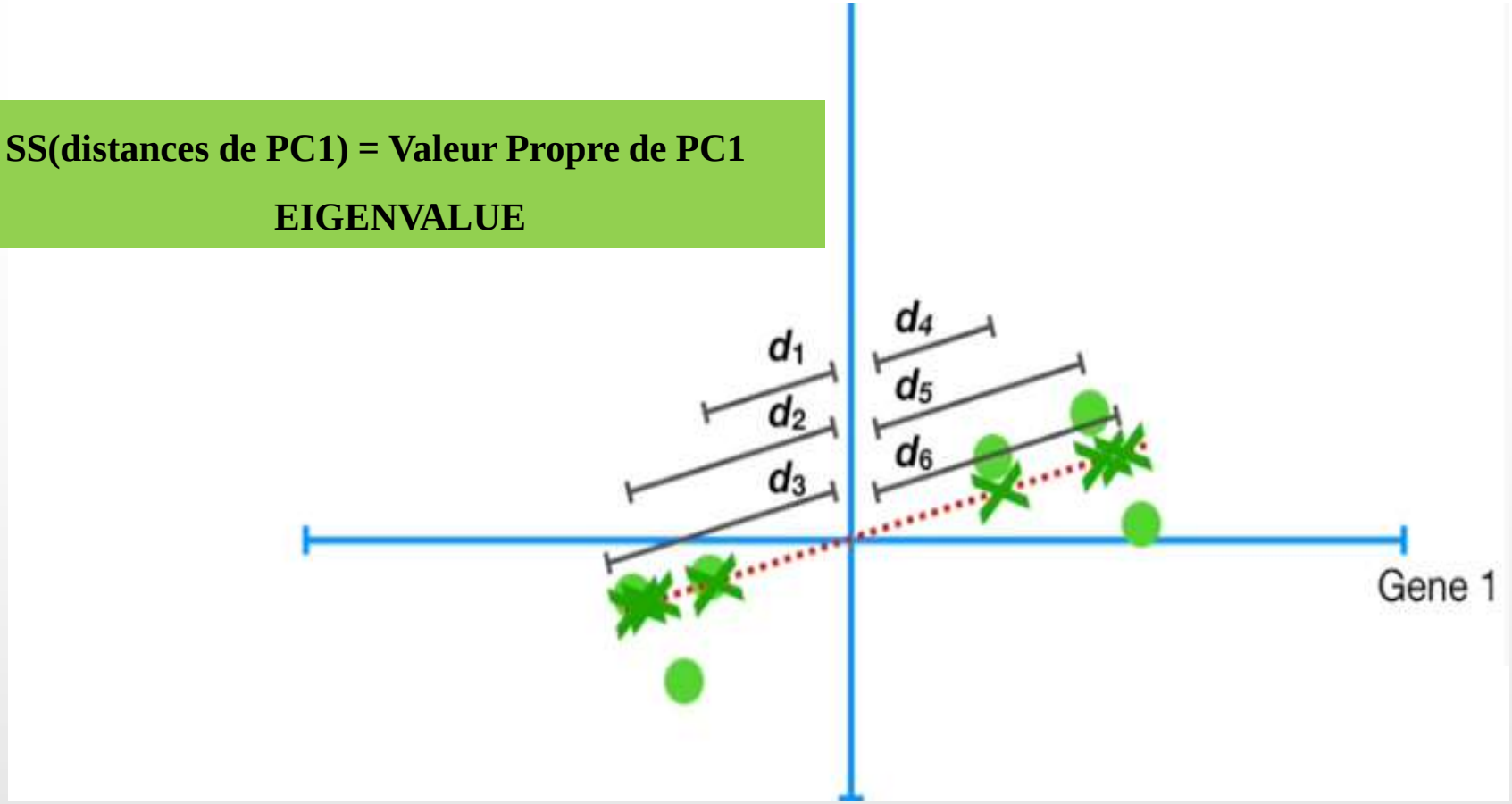


Illustration graphique de l'ACP

Comme il s'agit d'un graph à 2-D seulement ➔ **PC2** est simplement la droite perpendiculaire à **PC1** passant par l'origine (sans aucune optimisation)

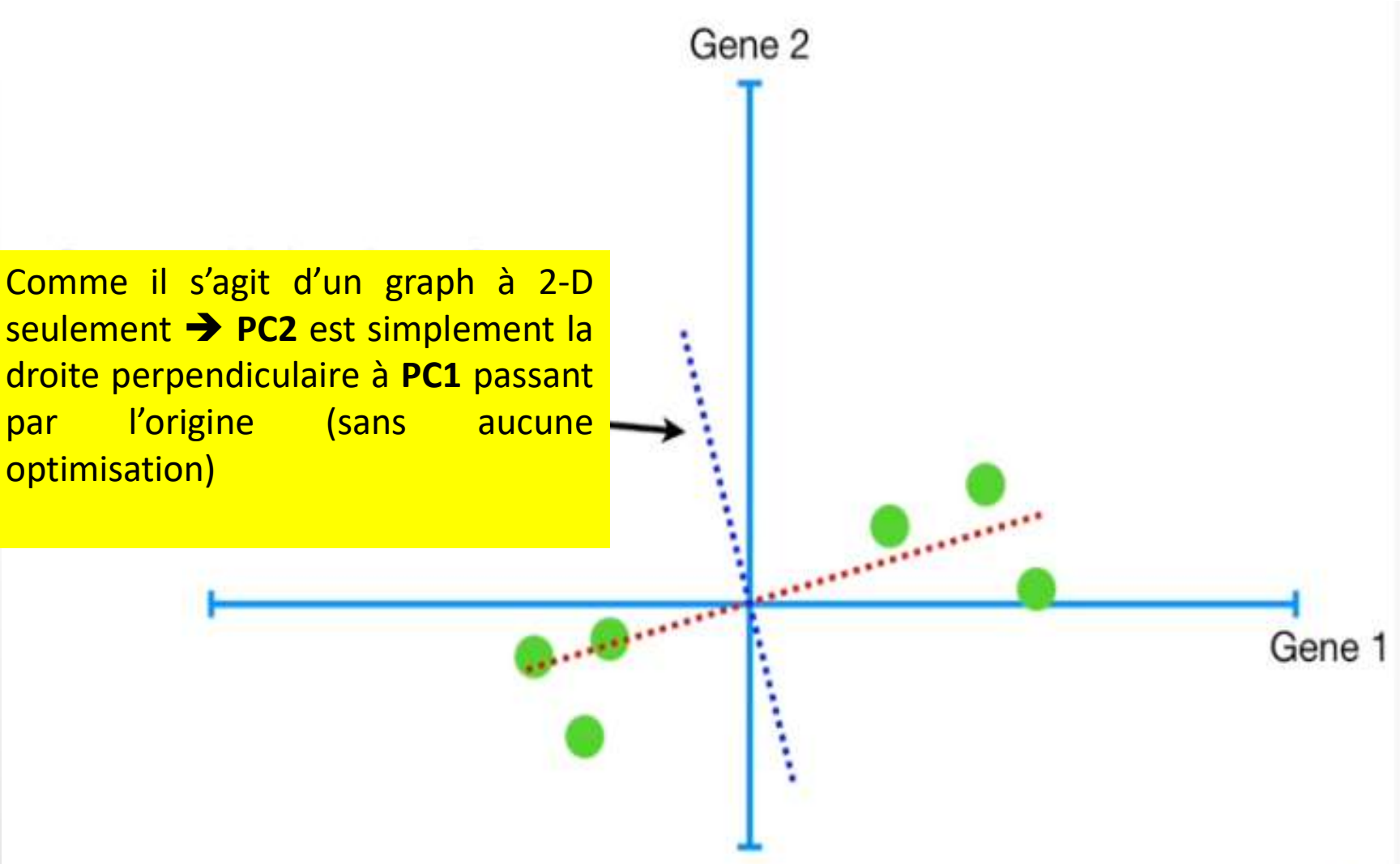
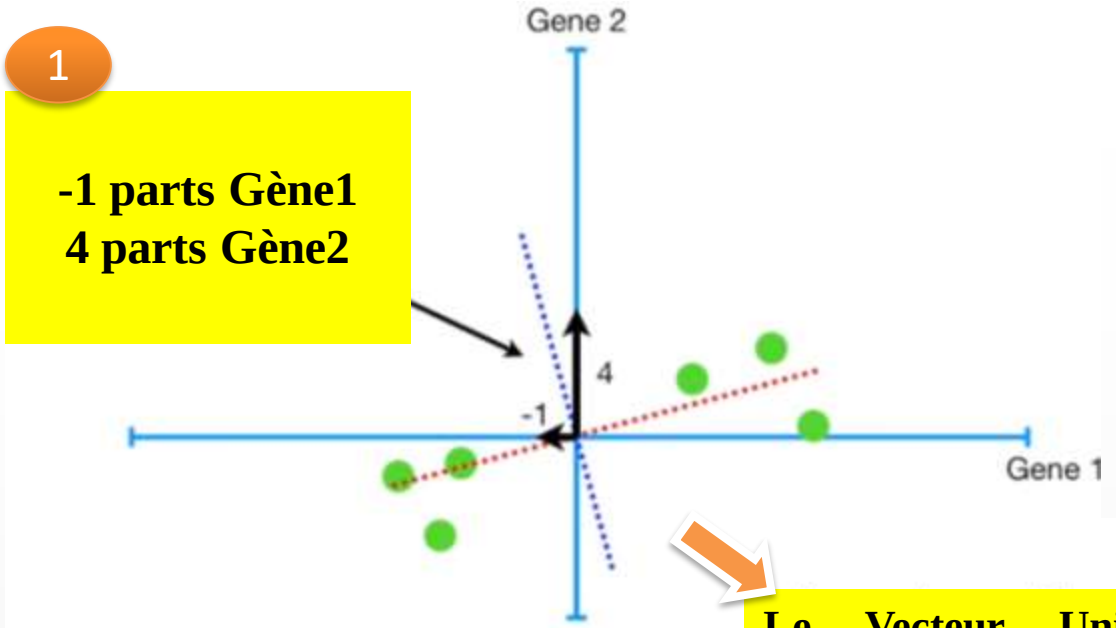


Illustration graphique de l'ACP

1

-1 parts Gène1
4 parts Gène2



2

Le Vecteur Unitaire, constitué de 0.97 parts Gene2 et -0.242 parts Gene1, est appelé : Vecteur Propre pour PC2

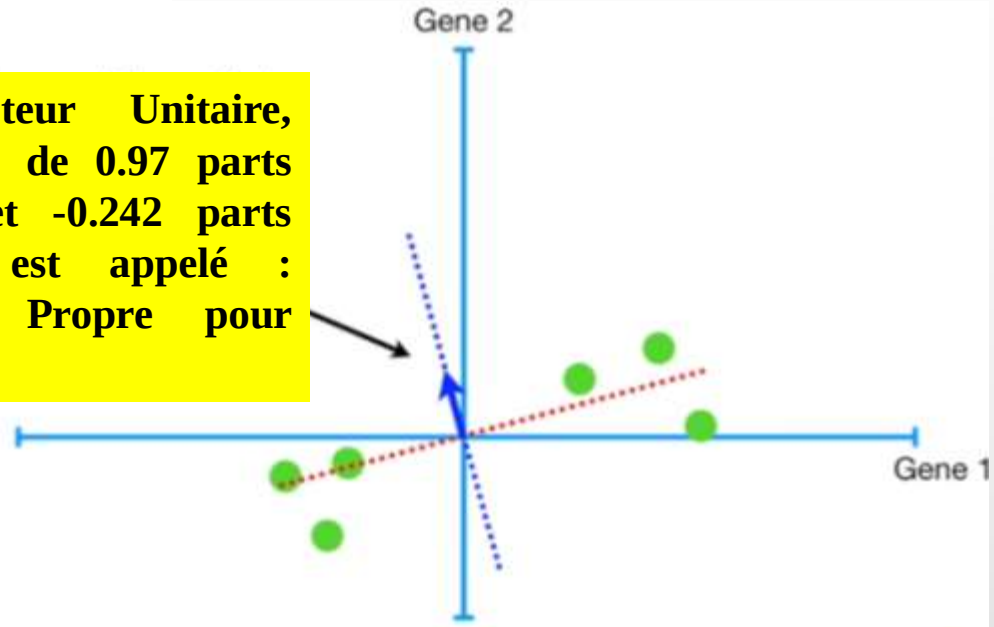


Illustration graphique de l'ACP

$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ =somme des carrés des distances (Sum of squared distance)

SS(distances)

SS(distances de PC2) = Valeur Propre de PC2

La valeur propre pour PC2 est la somme des carrés des distances entre l'origine est les projections

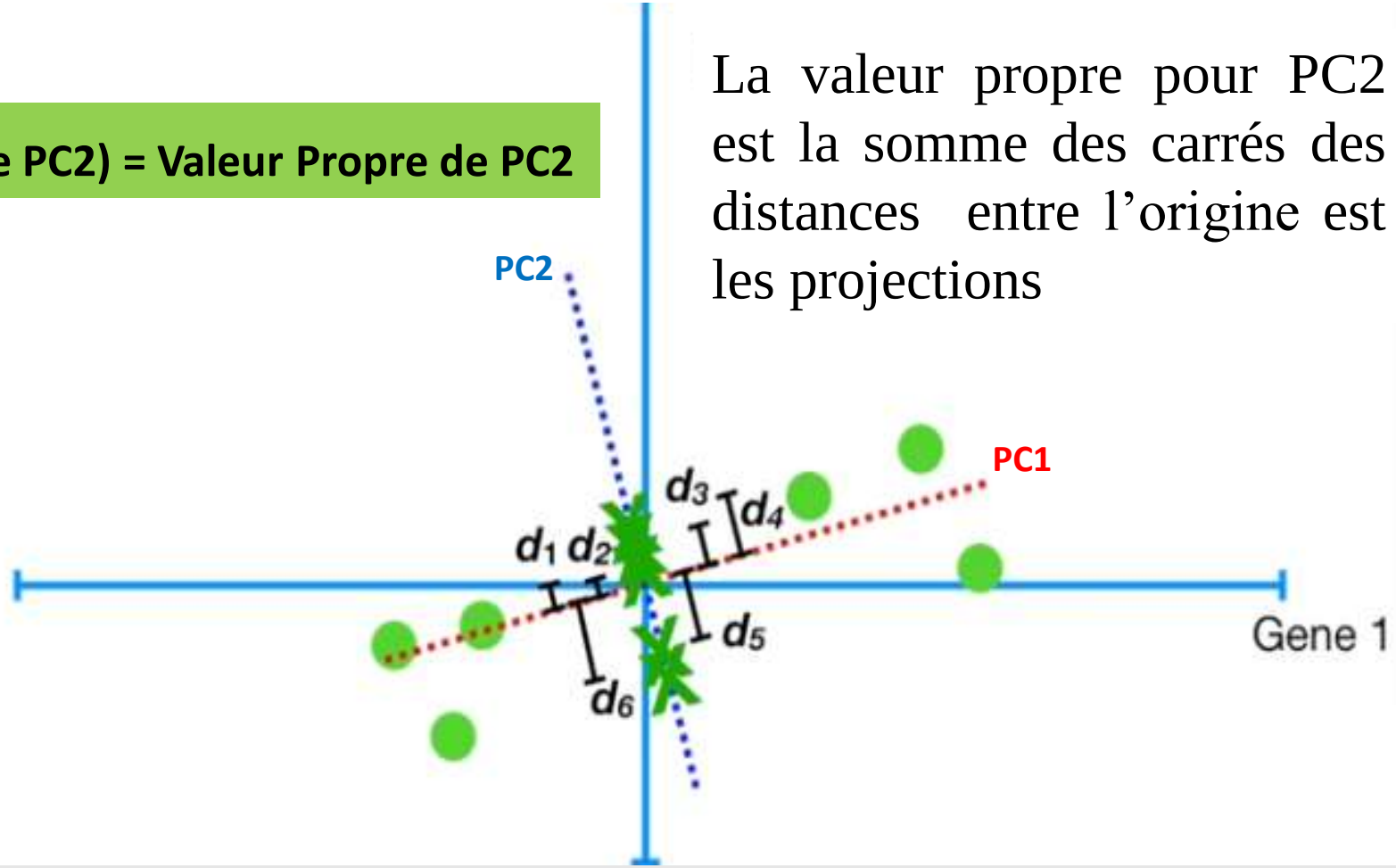


Illustration graphique de l'ACP

Pour avoir le graphique final de l'ACP ➔ une rotation complète du système afin d'avoir PC1 horizontal

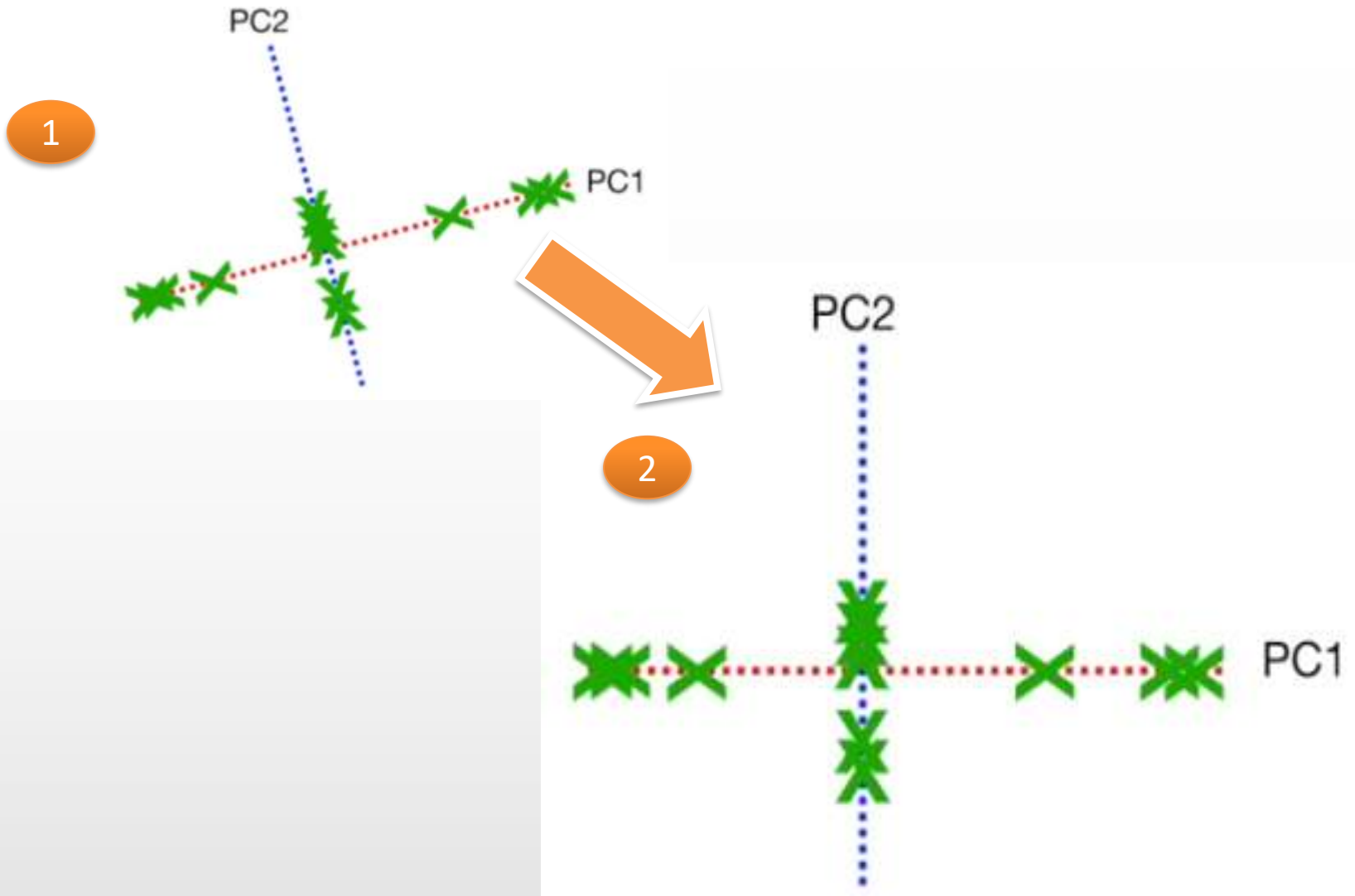


Illustration graphique de l'ACP

Les points projetées sur PC1 et PC2 sont utilisées pour avoir la position des données selon l'ACP

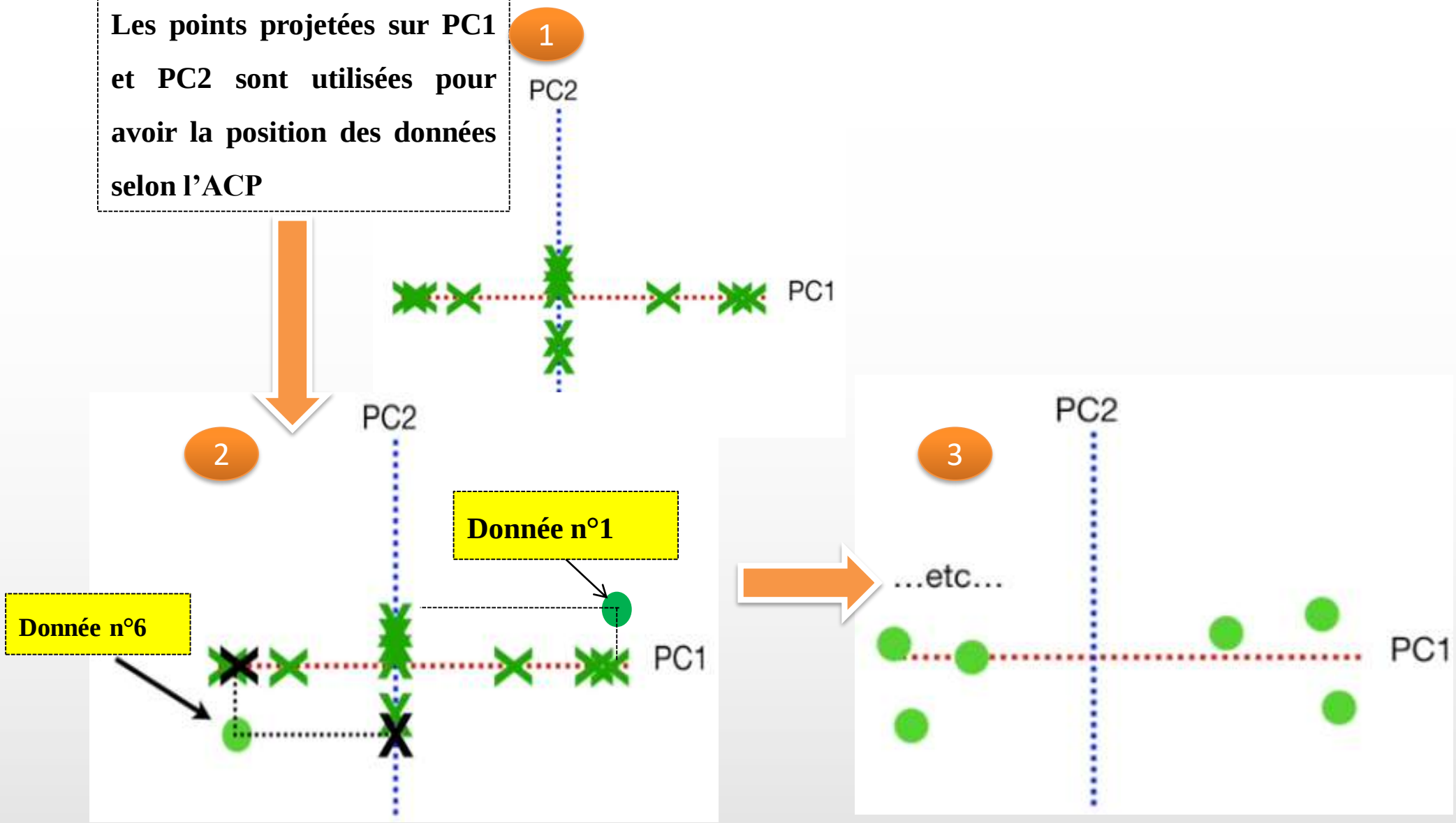
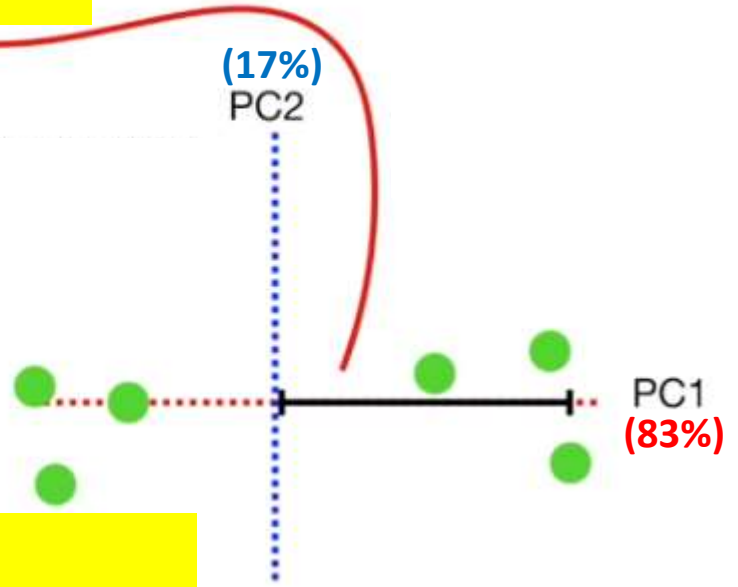


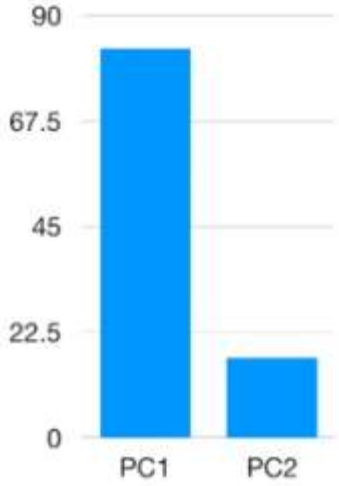
Illustration graphique de l'ACP

Rappelant les valeurs propres

SS(distances de PC1) = Valeur Propre de PC1

SS(distances de PC2) = Valeur Propre de PC2





Calculer la variance par rapport à l'origine (0,0)

$$\frac{SS(\text{distances for PC1})}{n-1} = \text{Variation for PC1}$$

$$\frac{SS(\text{distances for PC2})}{n-1} = \text{Variation for PC2}$$

- **Var PC1 = 15 et Var PC2=3 ➔ Total Var = 18**
- **PC1 compte 15/18=0.83=83% de la variation totale par rapport PCs,**
- **PC2 compte 3/18=0.17=17% de la variation totale**

Illustration graphique de l'ACP

- L'ACP avec 3 variables ?

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2

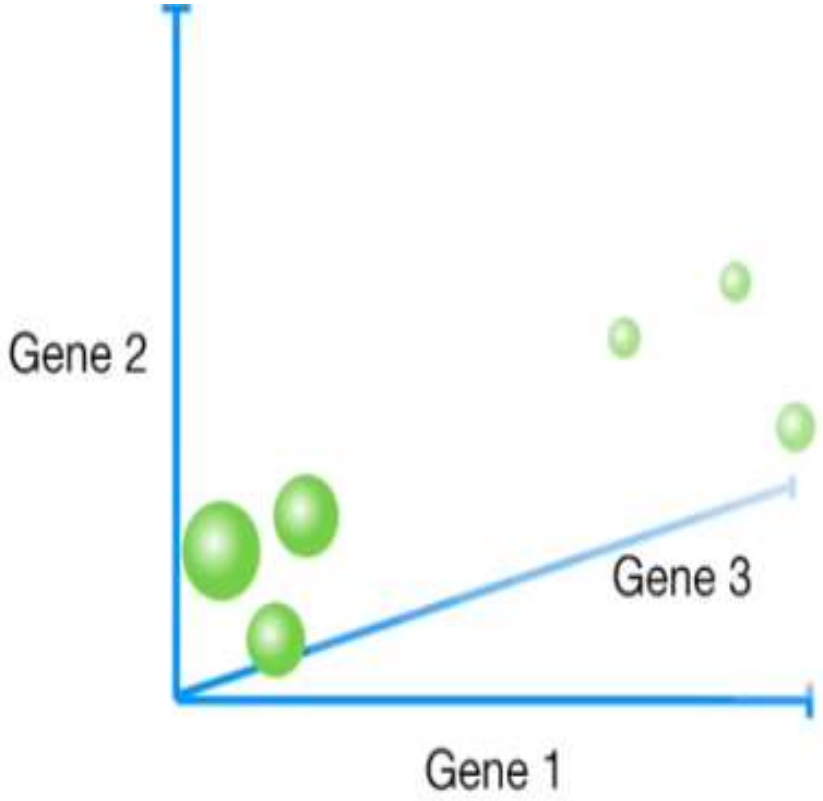


Illustration graphique de l'ACP

Si on a plus de variables, il suffit de trouver plus de composantes PCs. En théorie il y'a une PC par variable…

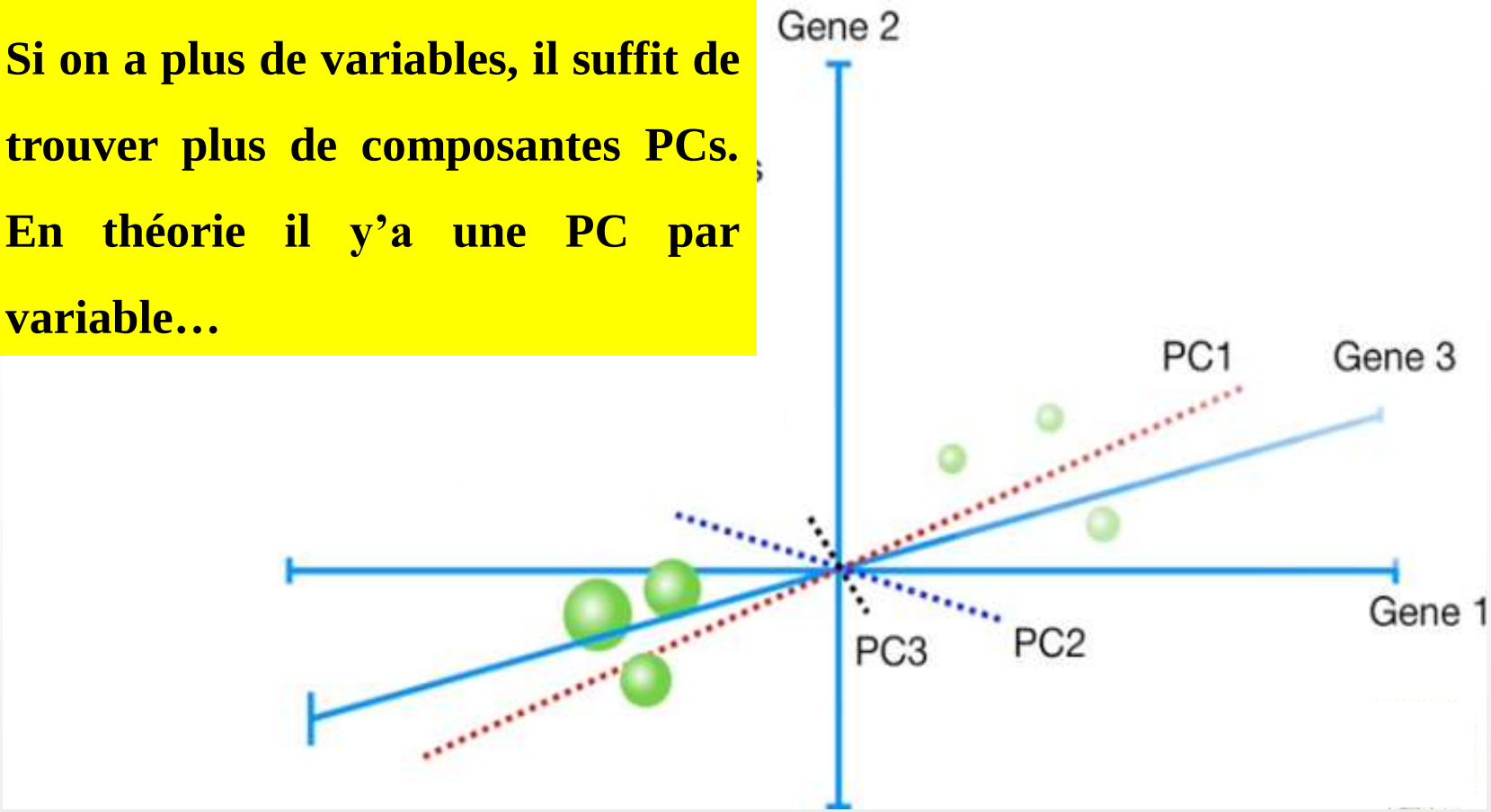


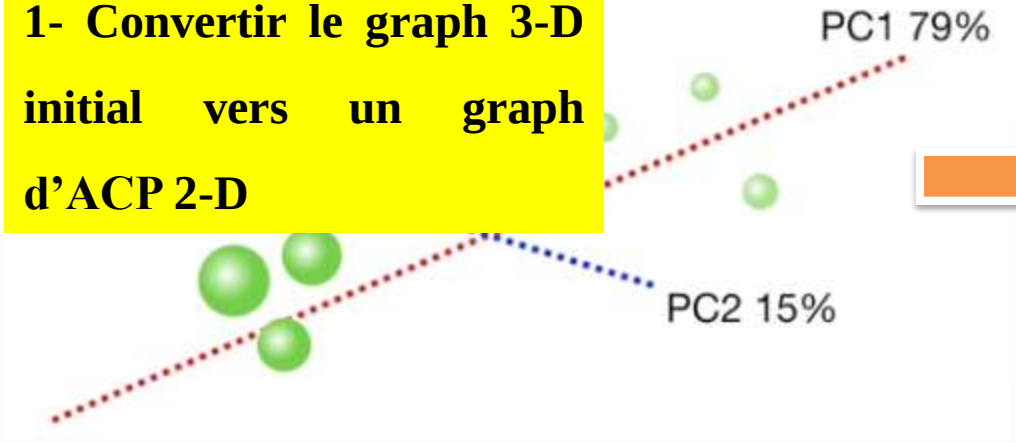
Illustration graphique de l'ACP

PC1 et PC2 compte 94% = la majorité des variations des données

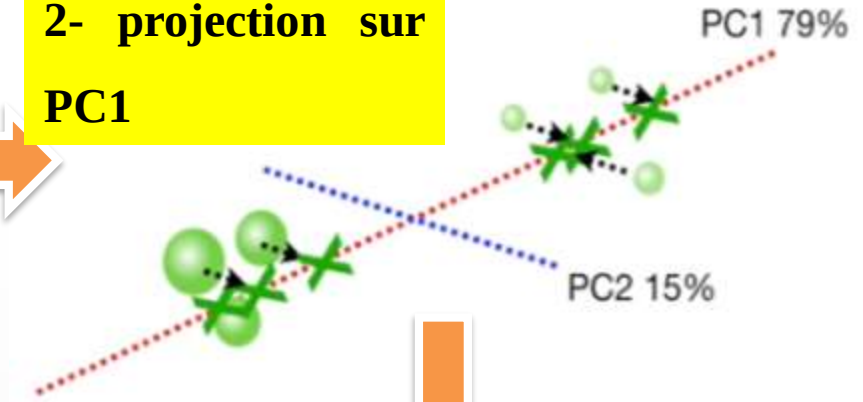
➔ Donc un graph 2-D utilisant juste PC1 et PC2 représente une meilleure approximation de l'espace 3-D initial

Illustration graphique de l'ACP

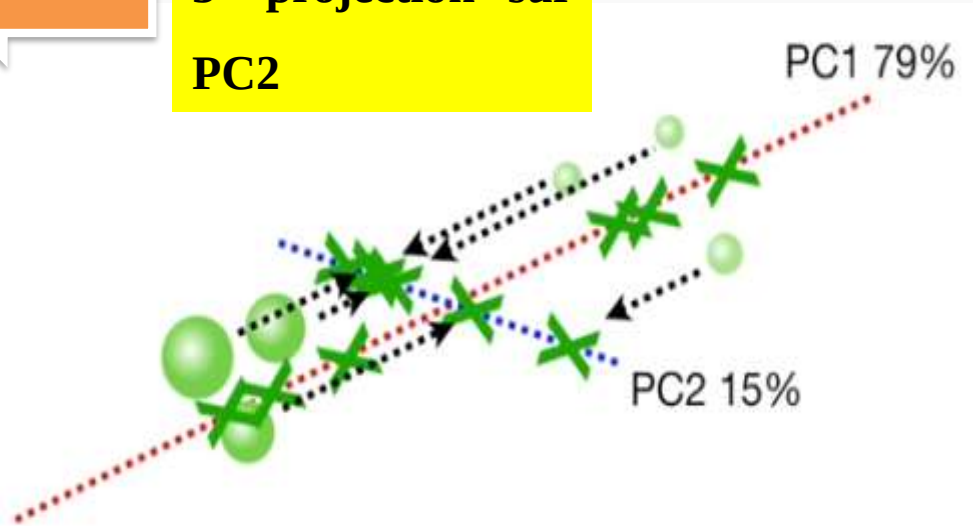
1- Convertir le graph 3-D initial vers un graph d'ACP 2-D



2- projection sur PC1



3- projection sur PC2



4- Rotation de PC1, PC2 ensuite représenter les données

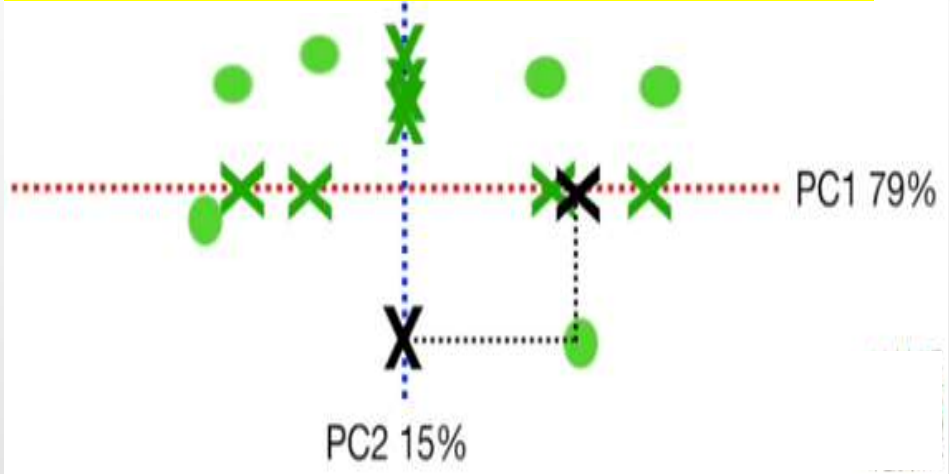


Illustration graphique de l'ACP

- ACP avec 4 variables ?

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	20	6	2	18	19

PC1 et PC2 compte 90% de variation ➔ on peut les utilisées pour le graph d'ACP