

Pourquoi l'algèbre linéaire ?

Cours 1 — Algèbre linéaire et analyse de données

Christophe Ambroise

L3 GBI — Université d'Évry

Un problème concret

Génétique des populations

Novembre *et al.*, **Genes mirror geography within Europe**, *Nature*, 2008.

3 000 individus européens

500 000 SNP par individu

Un tableau de 3 000 lignes et 500 000 colonnes.

Génétique des populations

Novembre *et al.*, **Genes mirror geography within Europe**, *Nature*, 2008.

3 000 individus européens

500 000 SNP par individu

Un tableau de 3 000 lignes et 500 000 colonnes.

À l'échelle En écrivant petit : **1 mètre** de haut, **166 mètres** de large.

Un SNP

Une position du génome, variable d'un individu à l'autre.
Deux allèles possibles : **C** ou **T**.

Un SNP

Une position du génome, variable d'un individu à l'autre.
Deux allèles possibles : **C** ou **T**.
Nombre d'allèles **T** hérités :

Mère \ Père	C	T
C	0	1
T	1	2

Le tableau

	SNP 1	SNP 2	SNP 3	SNP 4	...	SNP 500 000
Individu 1	0	2	1	0	...	1
Individu 2	1	2	0	2	...	0
Individu 3	0	1	1	0	...	2
⋮	⋮	⋮	⋮	⋮		⋮
Individu 3 000	2	0	1	1	...	1

Le tableau

	SNP 1	SNP 2	SNP 3	SNP 4	...	SNP 500 000
Individu 1	0	2	1	0	...	1
Individu 2	1	2	0	2	...	0
Individu 3	0	1	1	0	...	2
⋮	⋮	⋮	⋮	⋮		⋮
Individu 3 000	2	0	1	1	...	1

Des nombres. Rien d'autre.

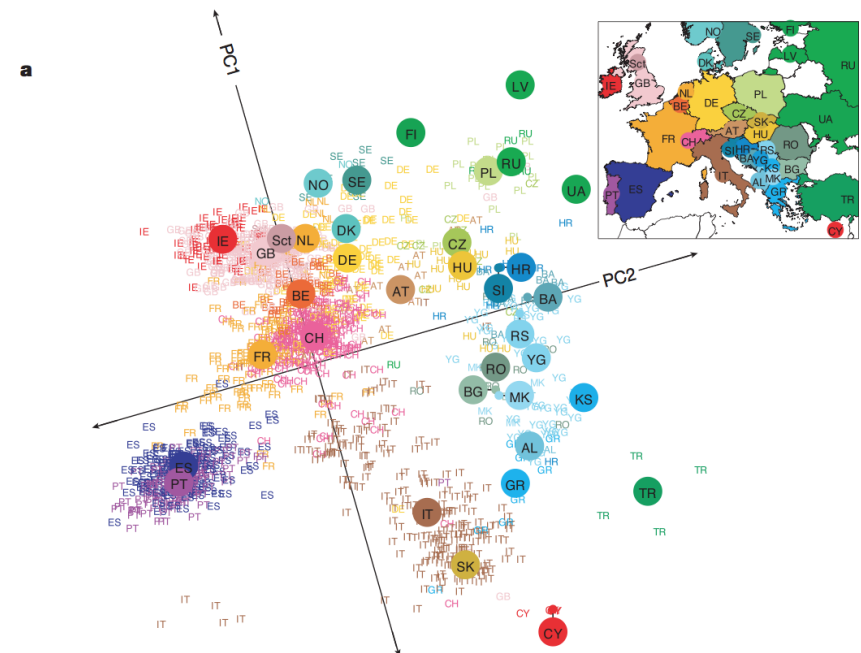
Autrement dit

Un individu = un point dans un espace à **500 000 dimensions**.

La question Comment **résumer** ce tableau sans le trahir ?

La réponse de l'article

Indessinable.



500 000 colonnes résumées en 2. Novembre et al/ 2008

Ce que vous venez de voir

Aucune information géographique fournie à la méthode.

De l'ADN, uniquement.

Trois problématiques

Ce que vous venez de voir

Aucune information géographique fournie à la méthode.

De l'ADN, uniquement.

Notre point d'arrivée Cette figure : une **analyse en composantes principales**. À votre portée en **séance 18**.

Trois problématiques

Résoudre — $Ax = b$

Les inconnues, puis *toutes* les solutions. Séances 2 à 9

Trois problématiques

Résoudre — $Ax = b$

Les inconnues, puis *toutes* les solutions. Séances 2 à 9

Mesurer — comparer deux profils

Une distance, un angle, une corrélation. Séances 10 à 13

Des vecteurs

Trois problématiques

Résoudre — $Ax = b$

Les inconnues, puis *toutes* les solutions. Séances 2 à 9

Mesurer — comparer deux profils

Une distance, un angle, une corrélation. Séances 10 à 13

Synthétiser — extraire l'information

500 000 colonnes vers 2, sans perte de l'essentiel. Séances 14 à 18

L'objet de base

Un individu décrit par n mesures :

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$$

L'objet de base

Un individu décrit par n mesures :

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$$

Un **gène** aussi : ses niveaux d'expression dans n échantillons.

L'objet de base

Un individu décrit par n mesures :

$$\mathbf{x} = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$$

Un **gène** aussi : ses niveaux d'expression dans n échantillons.

Le même objet Lignes et colonnes : les mêmes vecteurs, les mêmes outils.

Addition

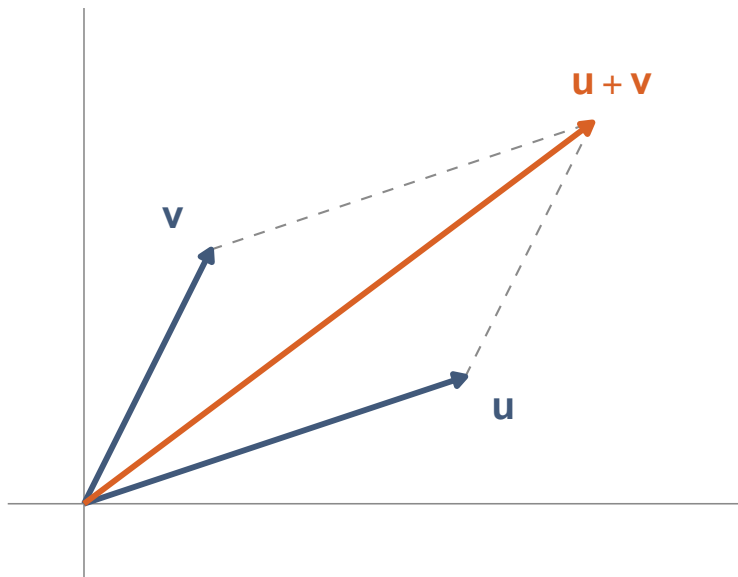


Figure 1

Coordonnée par coordonnée

Multiplication par un scalaire

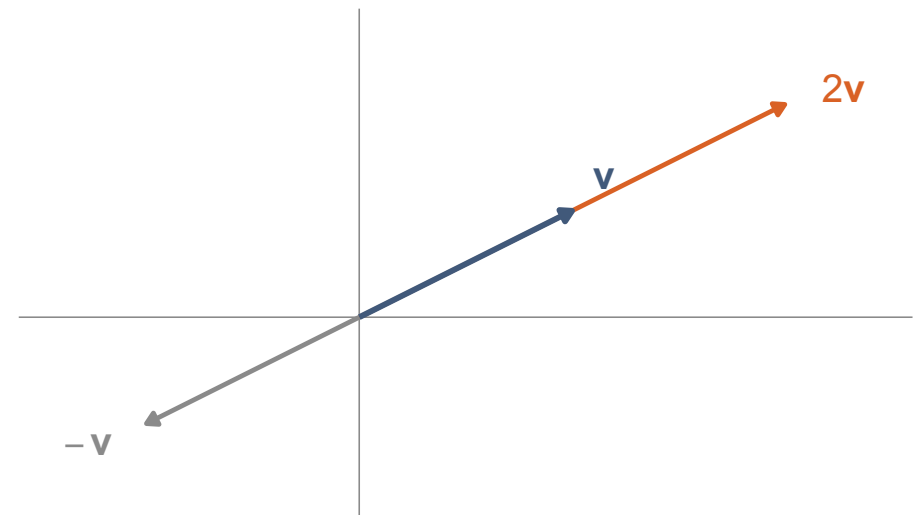


Figure 2

Étirer, compression, retournement. Direction conservée

Combinaison linéaire

$$\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \cdots + \alpha_p \mathbf{v}_p$$

Pourquoi c'est central

L'opération du semestre Résoudre, mesurer, synthétiser : les trois, ramenés à des **combinaisons linéaires**.

Combinaison linéaire

$$\alpha_1 \mathbf{v}_1 + \alpha_2 \mathbf{v}_2 + \cdots + \alpha_p \mathbf{v}_p$$

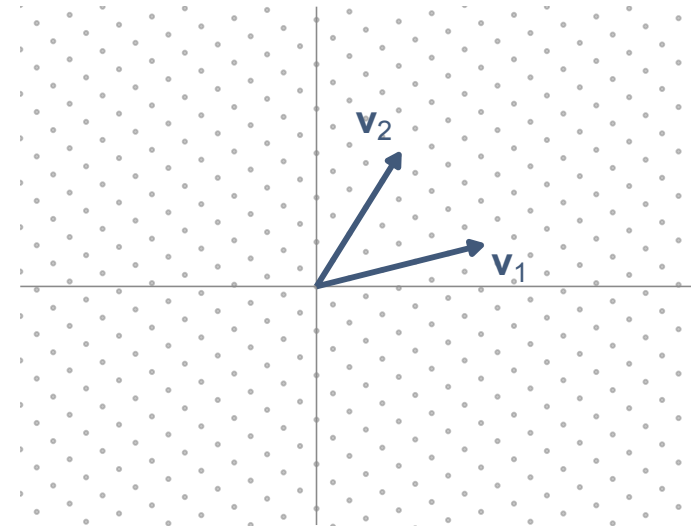


Figure 3

Pourquoi c'est central

L'opération du semestre Résoudre, mesurer, synthétiser : les trois, ramenés à des **combinaisons linéaires**.

Une ACP : deux colonnes, combinaisons linéaires des 500 000 autres.

Organisation

Séances	Contenu
1 – 5	Systèmes linéaires, matrices
6 – 9	Espaces vectoriels, base, dimension, rang
10 – 13	Produit scalaire, orthogonalité, projection
14 – 16	Déterminant, valeurs propres, théorème spectral
17 – 18	ACP sur machine , en R

Évaluation

2 examens — milieu et fin de semestre.

4 QCM de 10 minutes — en début de TD, séances 5, 9, 13 et 15.

Évaluation

2 examens — milieu et fin de semestre.

4 QCM de 10 minutes — en début de TD, séances 5, 9, 13 et 15.

Les TD Sur papier, sauf les deux dernières séances, en **R**.