Introduction aux méthodes statistiques

26 Février 2008

Christian CAPARIN
christian.caparin@wanadoo.fr

BUTS DE LA PRESENTATION

- Faire une présentation non théorique des méthodes statistiques utilisées en pratique
- Partir des problèmes et voir les exemples de solutions apportées

Champs d'application

- Finance : analyse de risque, prévision de séries financières
- Marketing: fidélité client, détection de fraude, validation de campagnes
- Bioinformatique

Les caractéristiques du problème des statisticiens

- Un nombre important de variables
  - Parfois des centaines
  - Pas nécessairement de même nature
- Un nombre important d'individus
  - Notamment à cause de l'apparition de l'outil informatique

Objectif du statisticien

- Réduire le nombre de variables du problème

OU/ET

- Réduire le nombre d'individus

Deux solutions

- Modèles non supervisés
  - Aucune variable est privilégiée l'une par rapport à l'autre
  - Volonté de résumer toute l'information
- Modèles supervisés
  - Explication d'une (ou plusieurs) variable(s) à l'aide des autres

Objectif

- Étude d'un méthode non supervisée

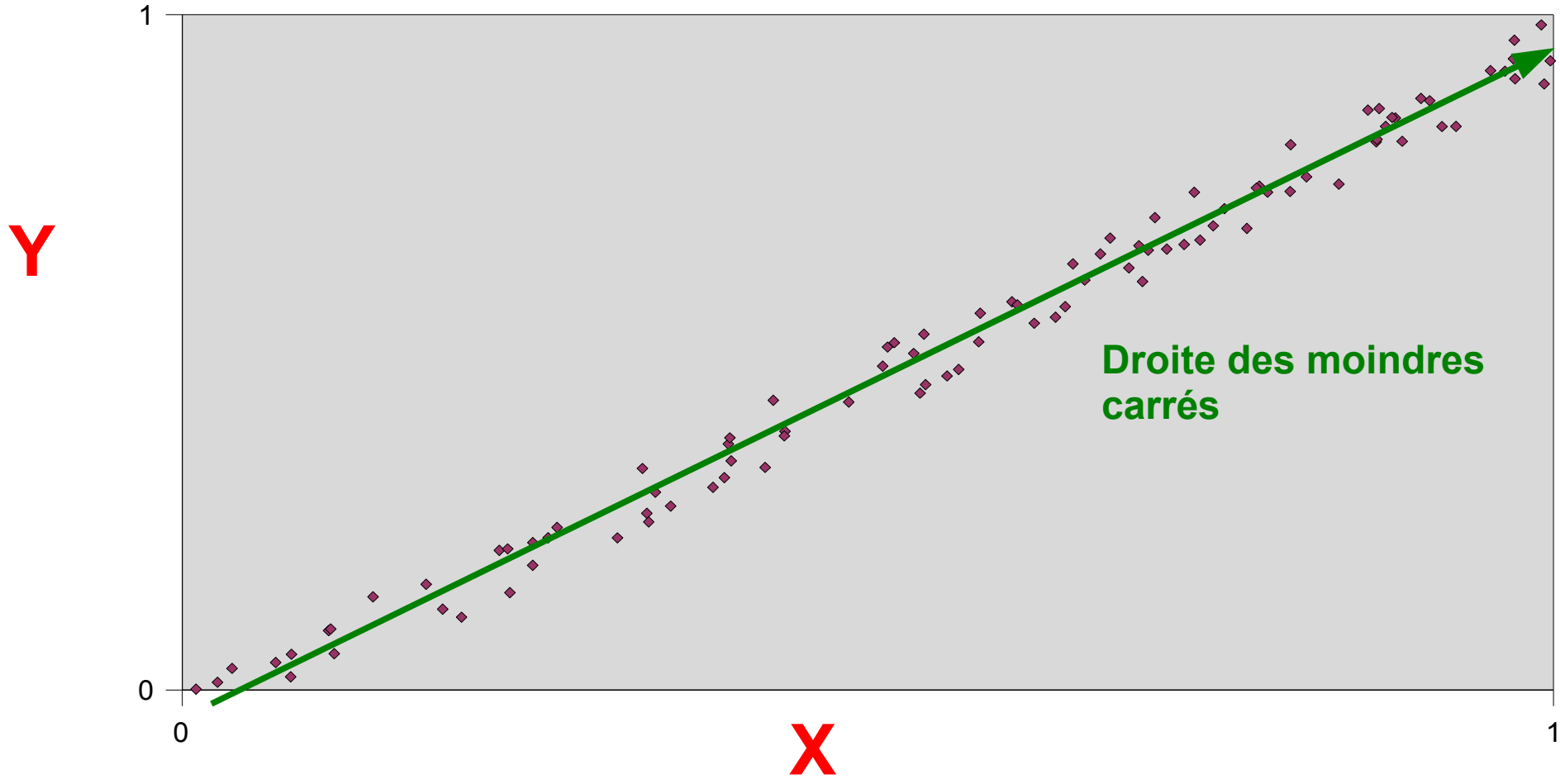
A.C.P.

Analyse en composantes principales

Les enjeux

- Constat: **difficulté** pour l'être humain de voir dans des espaces de dimension importante
  - Facilité uniquement pour les espaces de dimension 1 et 2 voire trois
- Objectif: **« résumer »** des données de dimension importante dans des espaces de dimension acceptable

Un cas simple



Problématique

- Algorithme proposé
  - Utiliser la droite des moindres carrés pour résumer les données
- Inconvénients
  - Quelle régression choisir? Celle de Y sur X ou celle de X sur Y?

Rappel: Les moindres carrés

- Hypothèse

$$Y = a\,X + b + \epsilon$$

- Estimation de a et b

$$min_{a,b} \sum_{i=1}^{N} (y_i - a\,x_i - b)^2$$

Problématique

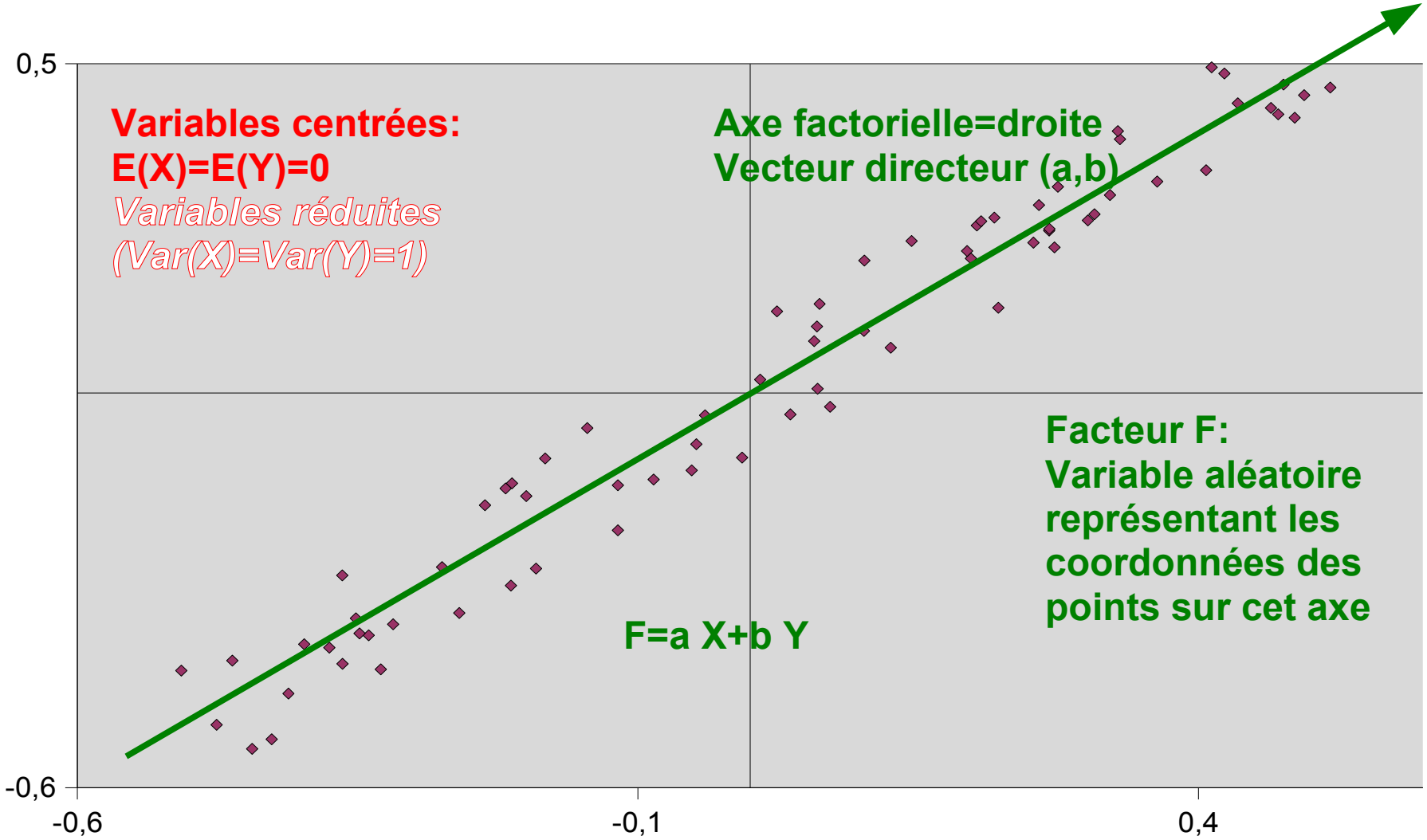
- Décomposition de la variance

$$VAR(Y) = VAR\underbrace{(aX+b)}_{\text{variance expliquée}} + VAR\underbrace{(Y-aX-b)}_{\text{variance résiduelle}}$$

- Non corrélation entre les résidus et le signal

$$cov(Y-aX-b, aX+b) = 0$$

Vocabulaire



Mathématisation

- Une relation fondamentale

$$VAR((X,Y)) = \underbrace{VAR(F\vec{u})}_{\text{variance expliquée par l'axe}} + \underbrace{VAR((X,Y) - F\vec{u})}_{\text{variance résiduelle}}$$

$$COV(F\vec{u}, (X,Y) - F\vec{u}) = 0$$

Idée d'un nouvel algorithme?

Mathématisation

- Algorithme pour la **dimension 2**
  - Je choisis l'axe qui donne la plus grande variance expliquée
  - L'autre axe factoriel est l'axe orthogonal au précédent (et qui passe par O)

A-t-on fini?

Mathématisation

- Problème
  - Si on est en dimension supérieure ou égale à 3, quel sera le deuxième axe factoriel qu'on va choisir?
- Solution
  - Prendre des axes factoriels qui
    - Expliquent le maximum de variance
    - Sont non corrélés aux précédents

Algorithme

- Données:
  - Une population sur laquelle sont renseignées les variables X_1,....., X_p quantitatives
- Initialisation:
  - Je choisis comme premier axe l'axe qui explique la plus grande variance possible
- Phase itérative
  - Je prends comme axe suivant, l'axe qui explique la plus grande variance possible mais qui en'est pas corrélé avec les axes déjà trouvés.

Algorithme: Phase d'initialisation

- Je cherche le premier facteur F_1 comme combinaison linéaire des variables initiales
  - $F_1=\alpha_1 X_1+.....+\alpha_p X_p$
  - $(\alpha_1 ,..,\alpha_p)$ vecteur unitaire
- Je choisis les $\alpha_1,.....,\alpha_p$ de façon à ce qu'ils maximisent var F_1.

Algorithme: Phase itérative

- Données:
  - J'ai déjà trouvé les facteurs F_1, F_{k-1}
- Je cherche le facteur F_k comme combinaison linéaire des variables initiales
  - $F_k=\beta_1 X_1+.....+\beta_p X_p$
- Je choisis les $\beta_{1,}.....,\beta_p$ de façon à ce qu'ils maximisent var F_k sous la contrainte:

$$cov(F_k,F_n)=0 \text{ pour } 1\leq n<k$$

Quelques résultats(I)

- Les vecteurs solutions sont les vecteurs propres de la matrice de covariance des variables.

$$\Sigma = \left(cov\left(X_i, X_j \right) \right)_{1 \le i \le p, 1 \le j \le p}$$

Quelques résultats(II)

- La somme des variances des F_i est égale à la variance totale de l'échantillon
  - Souvent, on ne parle qu'en terme de pourcentage de la variance totale

Quelques résultats(III)

- On peut exprimer X_i une variable en fonction des facteurs
  - $X_i=\alpha_1 F_1+.....+\alpha_p F_p$
- On peut exprimer les α_k facilement
  - $\alpha_k =cov(X_i,F_k)/var(F_k)$

Un exemple

Iris Virginica

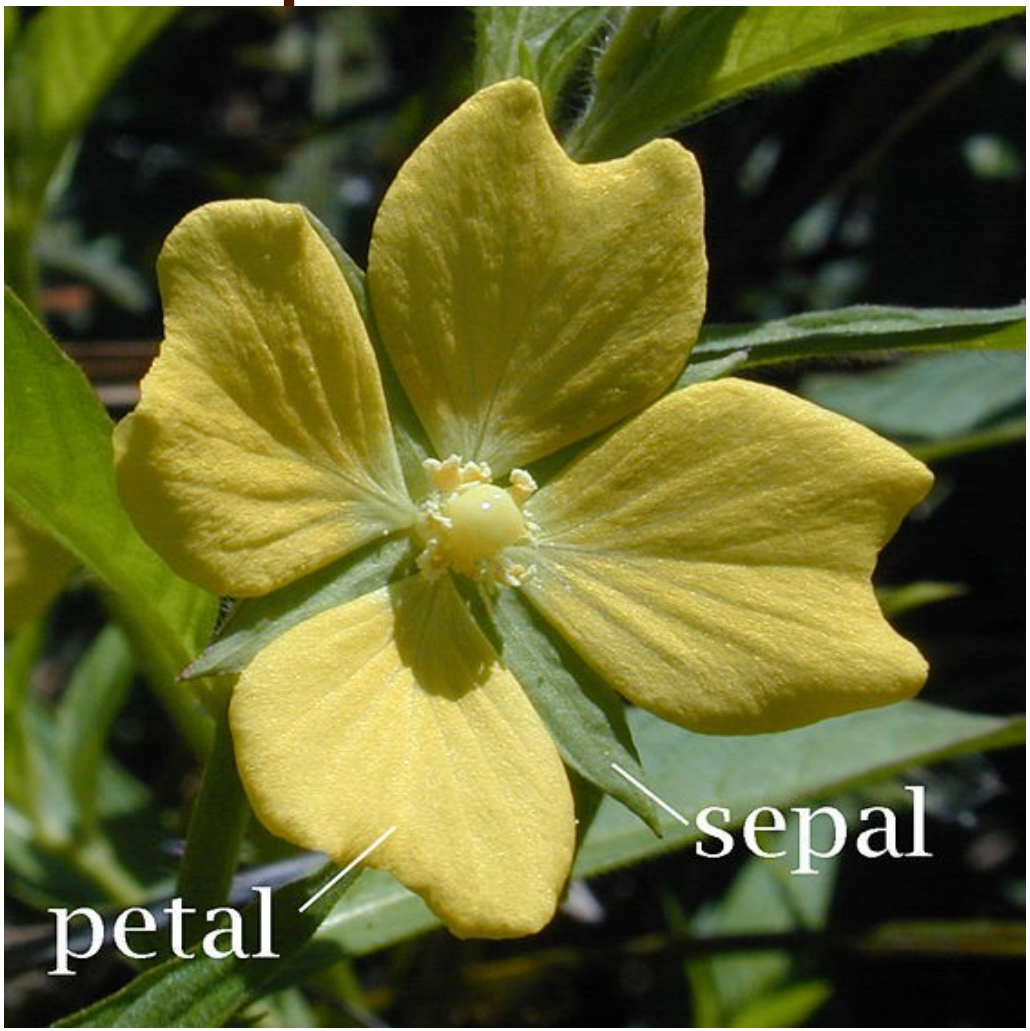
Iris Versicolor

Iris Setosa

Un exemple

- Échantillon de 150 fleurs de ce type (50 de chaque type)
- Description de chaque fleur avec quatre variables
  - Longueur et largeur du sépale
  - Longueur et largeur de la pétale

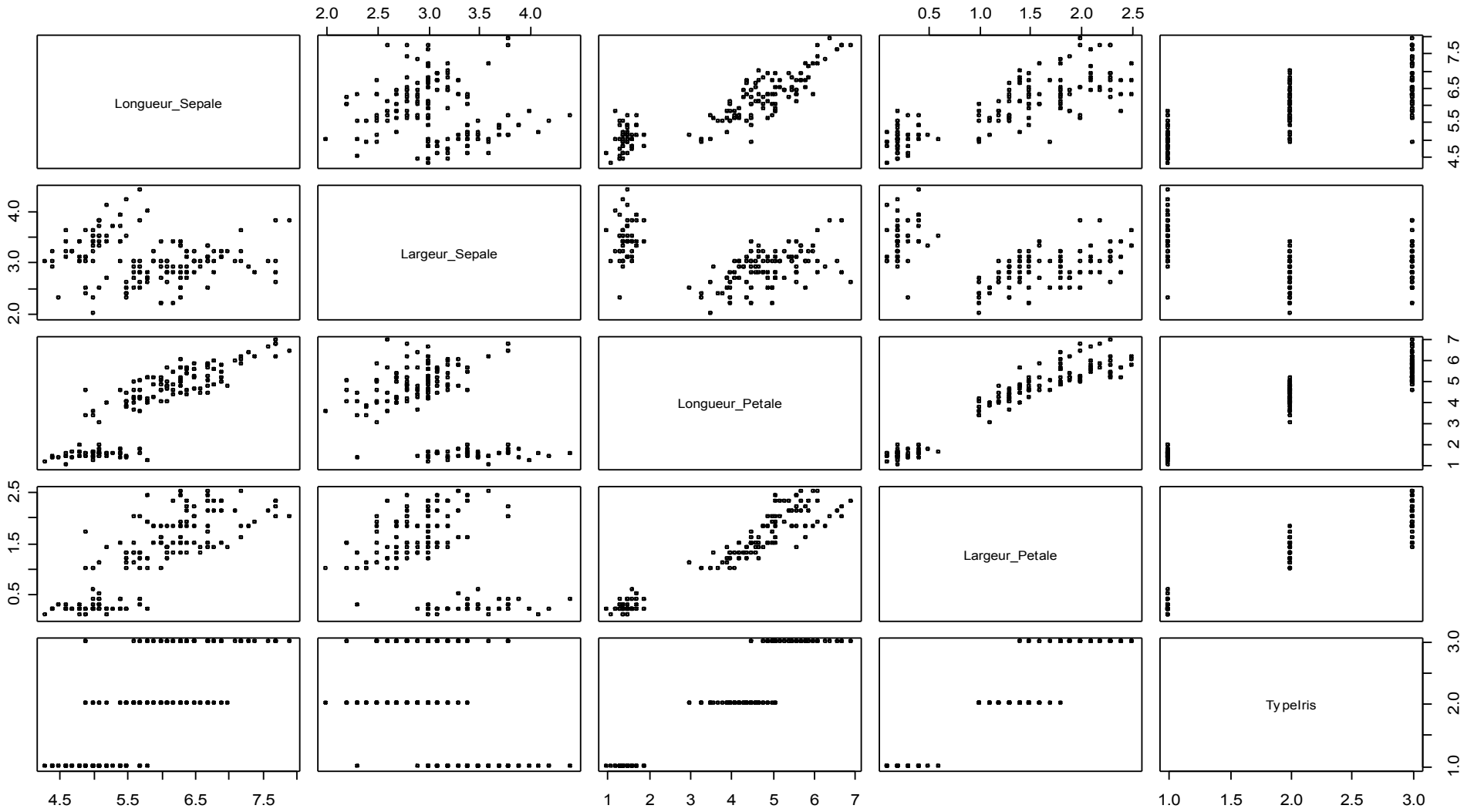
Une petite définition



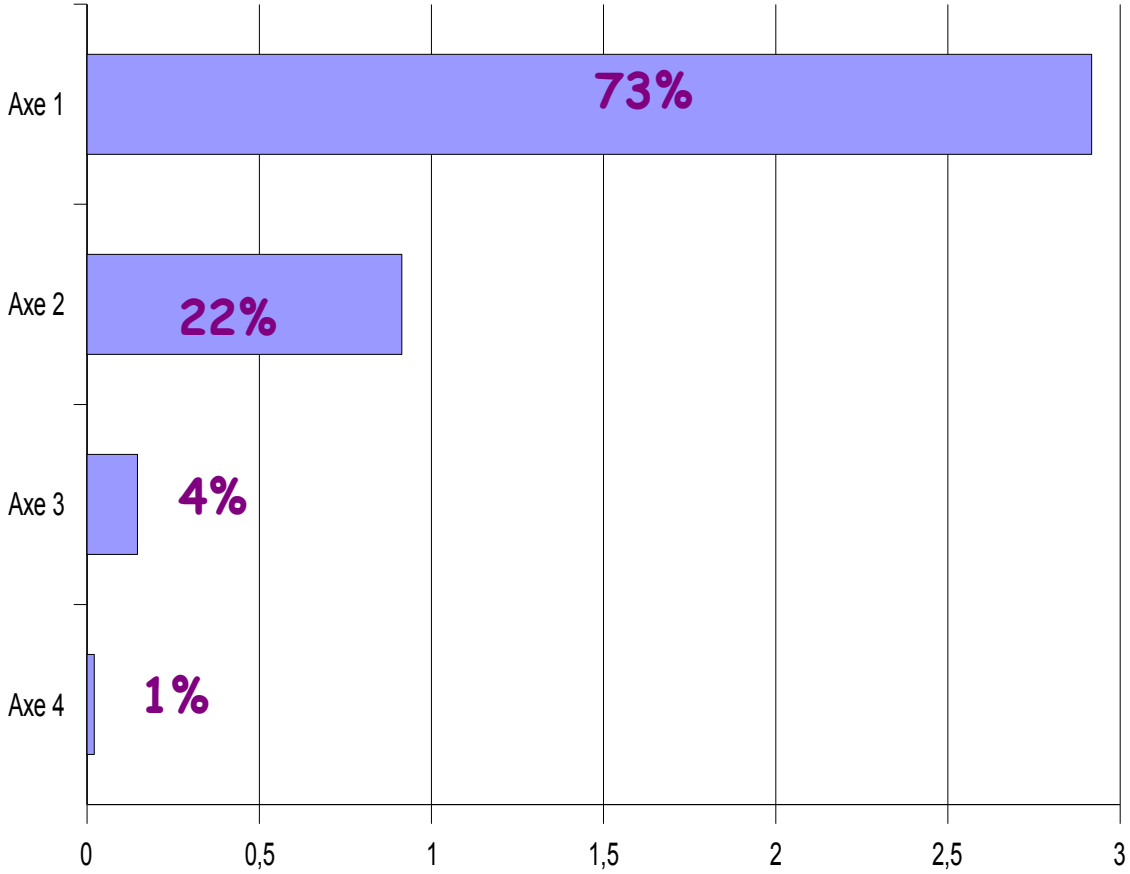
Un exemple

LONGUEUR SEPALE	LARGEUR SEPALE	LONGUEUR PETALE	LARGEUR PETALE	TYPE
5.1	3.5	1.4	0.2	Iris-setosa
4.9	3.0	1.4	0.2	Iris-setosa
4.7	3.2	1.3	0.2	Iris-setosa
4.6	3.1	1.5	0.2	Iris-setosa
5.0	3.6	1.4	0.2	Iris-setosa

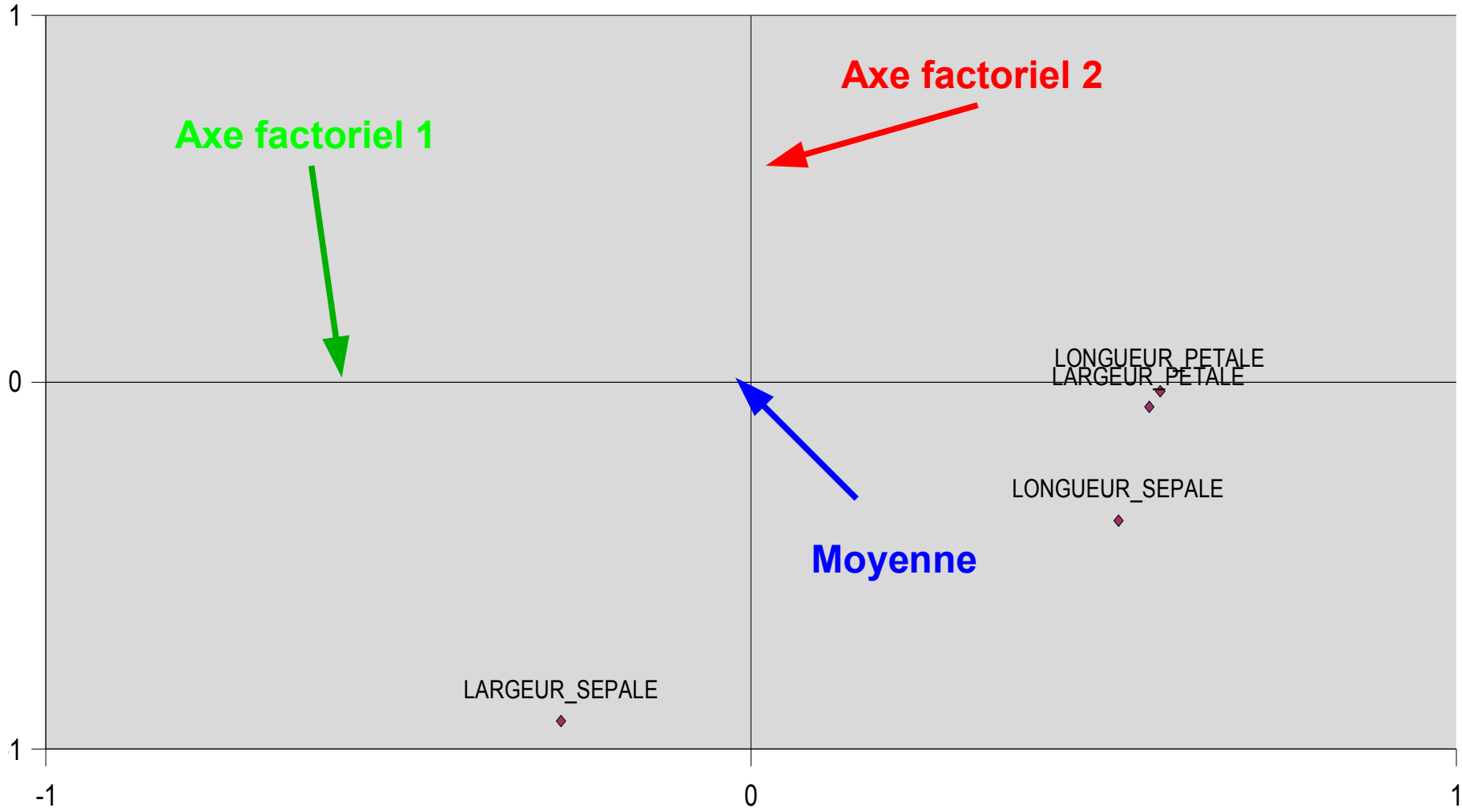
Les graphiques usuels



Variance de chaque axe

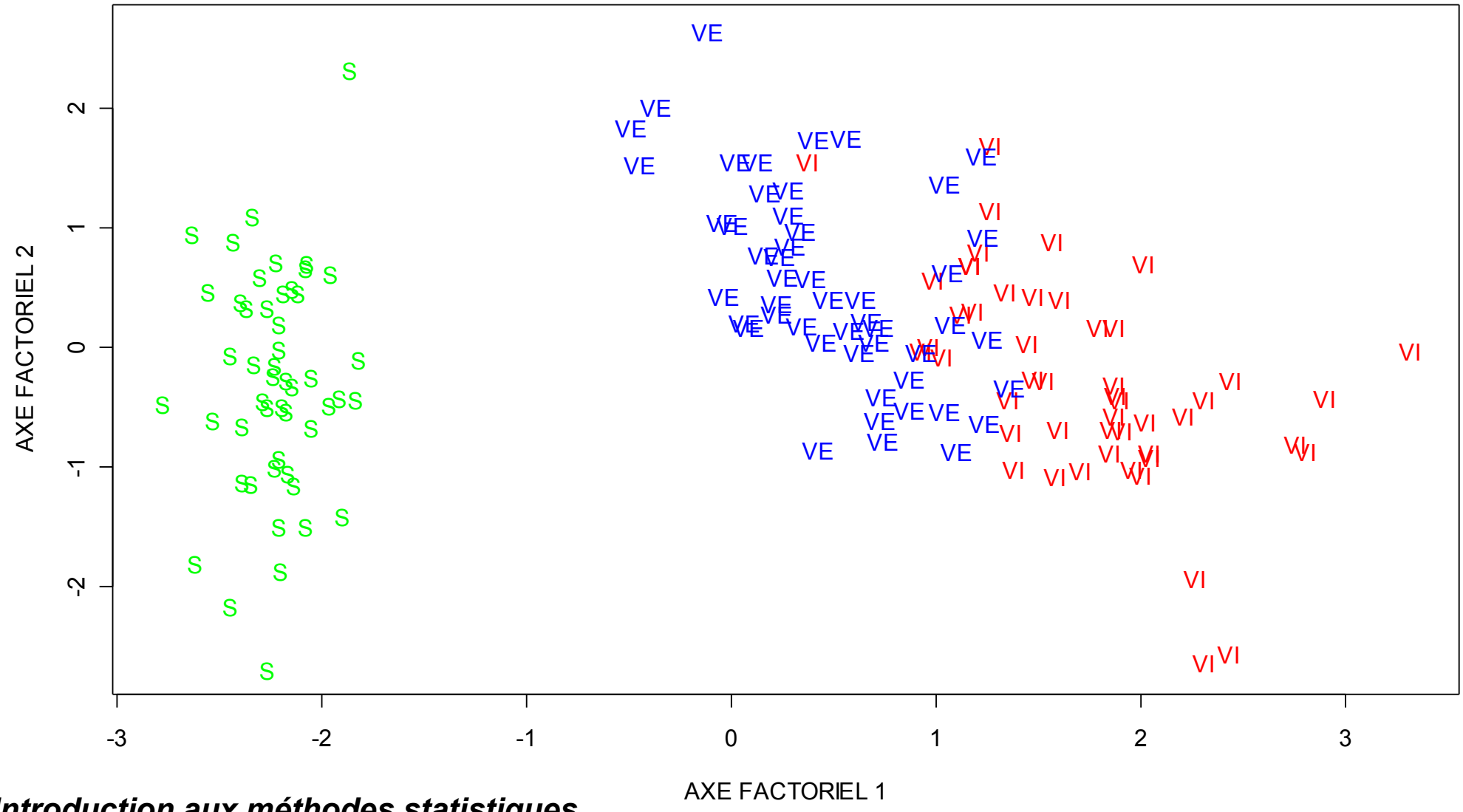


Projection des variables



- **>>Un exemple**

Projection des individus



Le premier axe

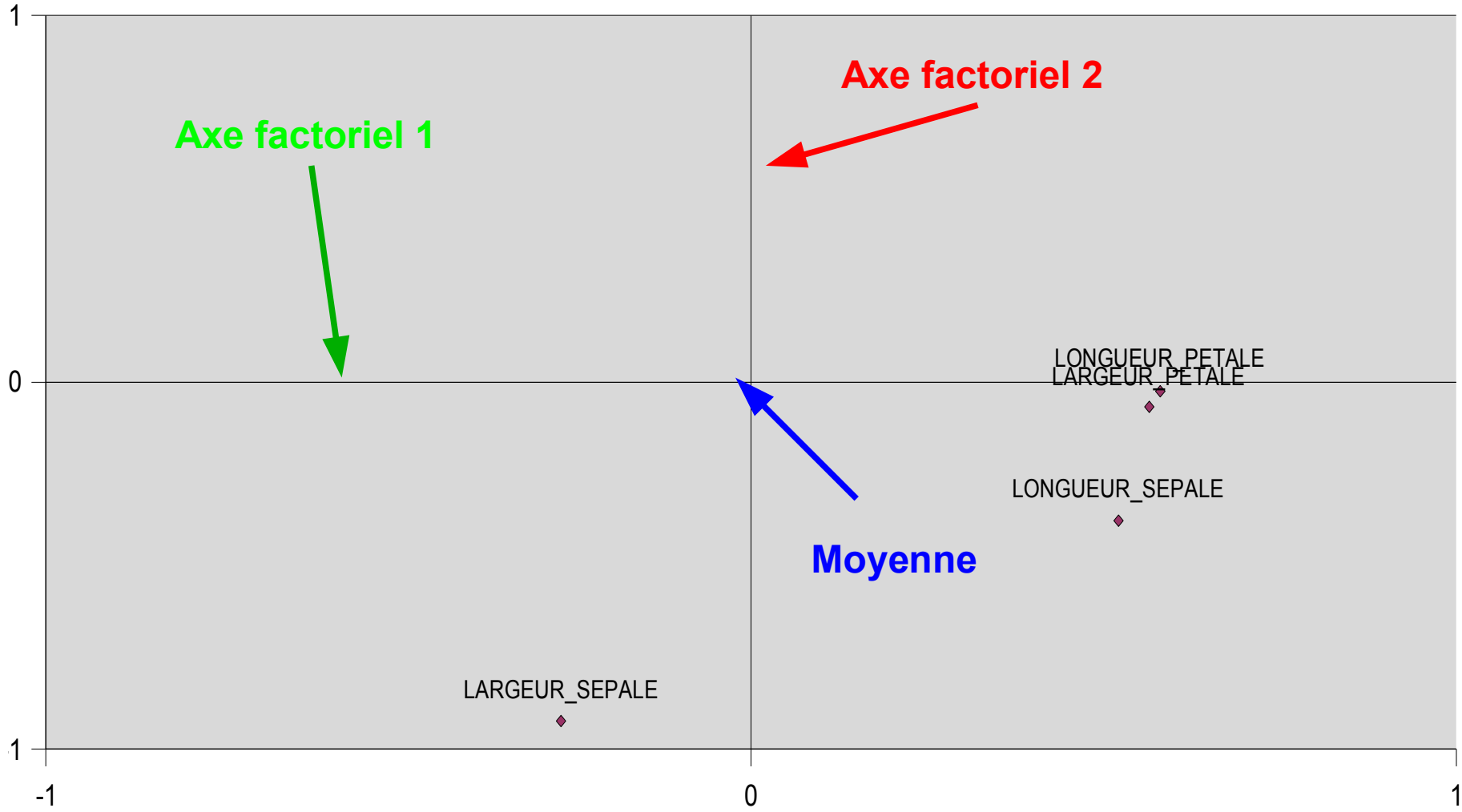
- **Faibles largeur , longueur de pétales et longueur de sépale**

SETOSA

- **Fortes largeur , longueur de pétales et forte longueur de pétale**

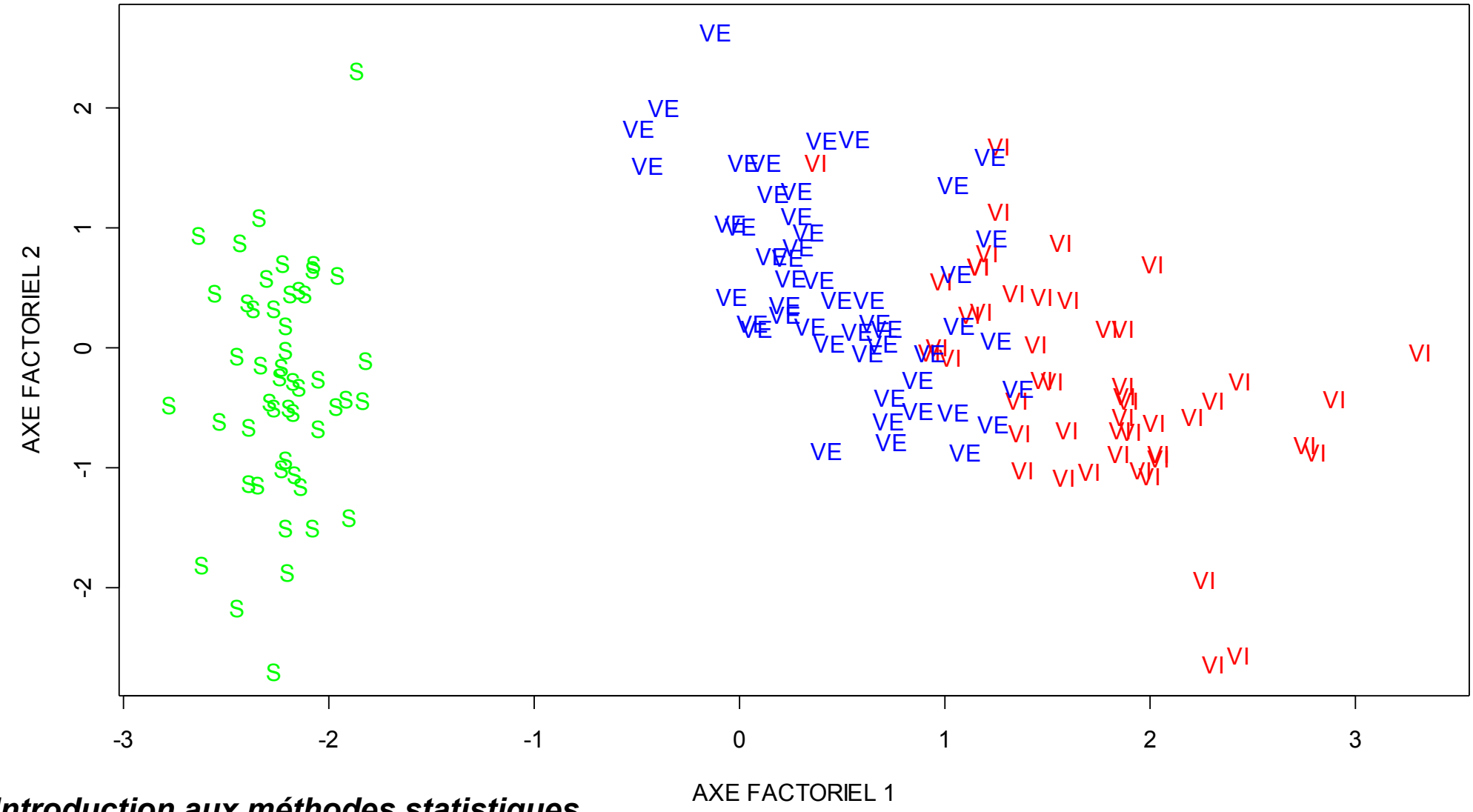
VIRGINICA

Projection des variables



- **>>Un exemple**

Projection des individus



Le second axe

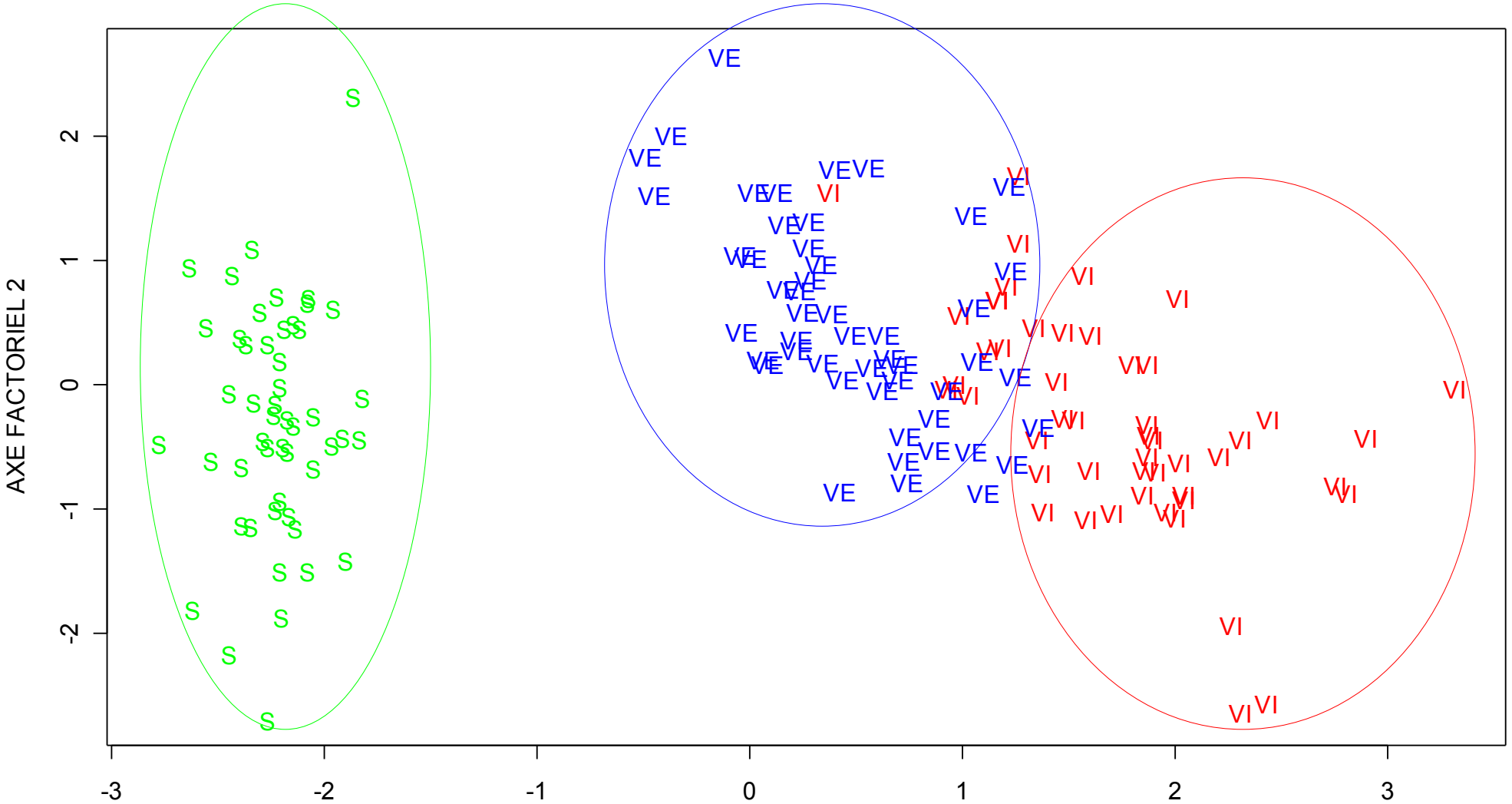
- **Forte largeur de sépale**

SETOSA

- **Faible largeur de sépale**

VERSICOLOR

Le rapport



Petit bilan

- Possibilité de réduire de façon intéressante le nombre de variables
- Possibilité de faire des typologies uniquement à partir des graphiques

L'ACP est une technique utilisée dans beaucoup de domaines.....

Réduction de variables

- Objectif: économiser du temps machine
  - Exemple: réduction de variables pour un logiciel de simulation de risque
- Objectif: Améliorer les précisions du modèle
  - Cas de la régression

Résumé/Classification

- Résumé d'un échantillon
  - Des axes d'opposition dans une population
- Classification des individus dans une population
  - Trouver des groupes dans une population
  - Expliquer des groupes prédéterminés par une ACP généralisée.