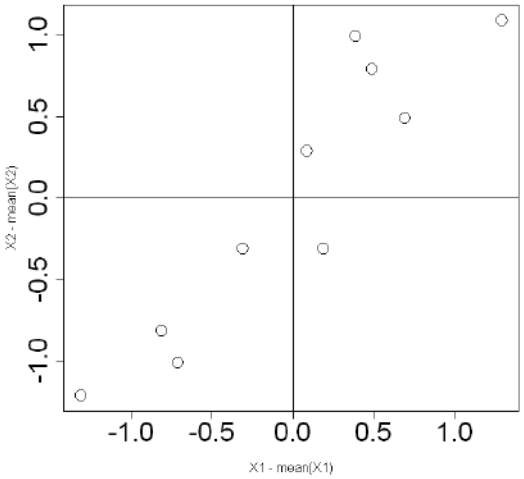
Un cas simple à 2 variables

- **Prenons le cas simple de notre exemple à deux variables X_1 et X_2 :**
  - > X_1 = c(2.5, 0.5, 2.2, 1.9, 3.1, 2.3, 2, 1, 1.5, 1.1)
  - > X_2 = c(2.4, 0.7, 2.9, 2.2, 3, 2.7, 1.6, 1.1, 1.6, 0.9)



- **X_1 et X_2 varient d'un même ordre de grandeur?**
  - var(X_1) = 0.6165556
  - var(X_2) = 0.7165556
  - cov(X_1, X_2) = 0.6154444
  - ACP non-normée = matrice de covariances

$$C = \begin{pmatrix} 0.6165556 & 0.6154444 \\ 0.6154444 & 0.7165556 \end{pmatrix}$$

- **Utilisation de la matrice de covariances pour changer de référentiel :**
  - Calcul du vecteur directeur de CP_1
  - Calcul du vecteur directeur de CP_2

- **Une histoire d'algèbre linéaire et de calculs matriciels :**
  - Détermination des *p* valeurs propres λ_j
  - Détermination des *p* vecteurs propres V_j

- **Soit la matrice de covariances *C* de taille *p* x *p*, elle admet *p* valeurs propres et *p* vecteurs propres associés, tels que :**
  - $CV_j = \lambda_j V_j$

- **Dans notre exemple à deux variables, *C* admet 2 valeurs propres et 2 vecteurs propres tels que soit vérifié les égalités suivantes :**

$$\begin{pmatrix} 0.6165556 & 0.6154444 \\ 0.6154444 & 0.7165556 \end{pmatrix} \begin{pmatrix} v_{1,1} \\ v_{1,2} \end{pmatrix} = \lambda_1 \begin{pmatrix} v_{1,1} \\ v_{1,2} \end{pmatrix}$$

$$\begin{pmatrix} 0.6165556 & 0.6154444 \\ 0.6154444 & 0.7165556 \end{pmatrix} \begin{pmatrix} v_{2,1} \\ v_{2,2} \end{pmatrix} = \lambda_2 \begin{pmatrix} v_{2,1} \\ v_{2,2} \end{pmatrix}$$

- **Détermination des 2 valeurs propres λ_1 et λ_2 :**
  - Calcul du déterminant de $C - \lambda I$
  - Résolution de l'équation $\det(C - \lambda I) = 0$

$$C - \lambda I = \begin{pmatrix} 0.6165556 & 0.6154444 \\ 0.6154444 & 0.7165556 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$C - \lambda I = \begin{pmatrix} 0.6165556 - \lambda & 0.6154444 \\ 0.6154444 & 0.7165556 - \lambda \end{pmatrix}$$

$$\det(C - \lambda I) = (0.6165556 - \lambda)(0.7165556 - \lambda) - 0.6154444^2$$

$$\lambda^2 - 1.333111\,\lambda + 0.06302444 = 0$$

$$\Delta = b^2 - 4ac$$

$$\lambda_1 = \frac{-b + \sqrt{\Delta}}{2a} = 1.284028 \qquad \lambda_2 = \frac{-b - \sqrt{\Delta}}{2a} = 0.04908323$$

- **Chaque valeur propre représente la variance des données autour d'un nouvel axe *CP* ou « composante principale » qui est une combinaison linéaire des variables de départ**

$$\lambda_1 + \lambda_2 = \mathrm{var}(X_1) + \mathrm{var}(X_2)$$

$$1.284028 + 0.04908323 = 0.6165556 + 0.7165556 = 1.333111$$

- **La première « composante principale » ou CP_1 associée à λ_1 porte 96% de la variance totale**

- **La deuxième « composante principale » ou CP_2 associée à λ_2 porte 4% seulement de la variance totale**

- **A partir d'une seule dimension (CP_1), il est possible ici de résumer 96% de l'information de départ contenue dans deux dimensions (X_1, X_2)**

- **Détermination des 2 vecteurs propres V_1 et V_2 :**
  - Résolution des 2 systèmes d'équations $CV_j - \lambda_j V_j = 0$

λ_1

$$\begin{pmatrix} 0.6165556 & 0.6154444 \\ 0.6154444 & 0.7165556 \end{pmatrix} \begin{pmatrix} v_{1,1} \\ v_{1,2} \end{pmatrix} = 1.284028 \times \begin{pmatrix} v_{1,1} \\ v_{1,2} \end{pmatrix}$$

$$\begin{cases} 0.6165556 \times v_{1,1} + 0.6154444 \times v_{1,2} - 1.284028 \times v_{1,1} = 0 \\ 0.6154444 \times v_{1,1} + 0.7165556 \times v_{1,2} - 1.284028 \times v_{1,2} = 0 \end{cases}$$

λ_2

$$\begin{pmatrix} 0.6165556 & 0.6154444 \\ 0.6154444 & 0.7165556 \end{pmatrix} \begin{pmatrix} v_{2,1} \\ v_{2,2} \end{pmatrix} = 0.04908323 \times \begin{pmatrix} v_{2,1} \\ v_{2,2} \end{pmatrix}$$

$$\begin{cases} 0.6165556 \times v_{2,1} + 0.6154444 \times v_{2,2} - 0.04908323 \times v_{2,1} = 0 \\ 0.6154444 \times v_{2,1} + 0.7165556 \times v_{2,2} - 0.04908323 \times v_{2,2} = 0 \end{cases}$$

- **Une solution possible (cf. sous la contrainte que V_1 et V_2 soient tout 2 des vecteurs unitaires de taille 1)**

$$V_1 = \begin{pmatrix} 0.6778736 \\ 0.7351785 \end{pmatrix} \qquad NB : v_{1,1}^2 + v_{1,2}^2 = 1$$

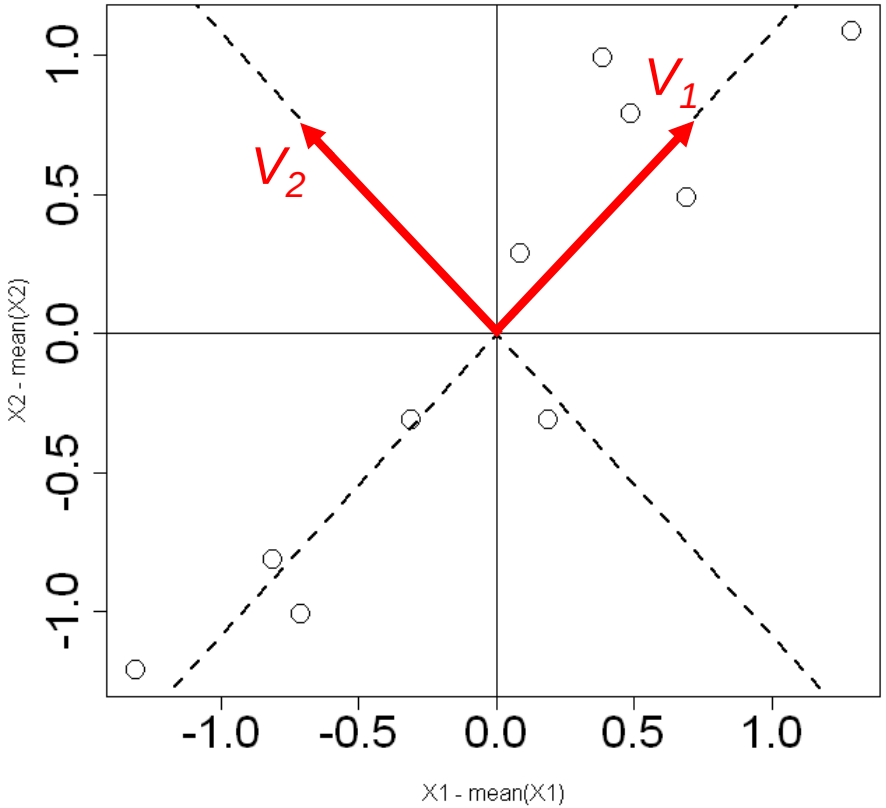
$$V_2 = \begin{pmatrix} -0.7351785 \\ 0.6778736 \end{pmatrix} \quad NB : v_{2,1}^2 + v_{2,2}^2 = 1$$

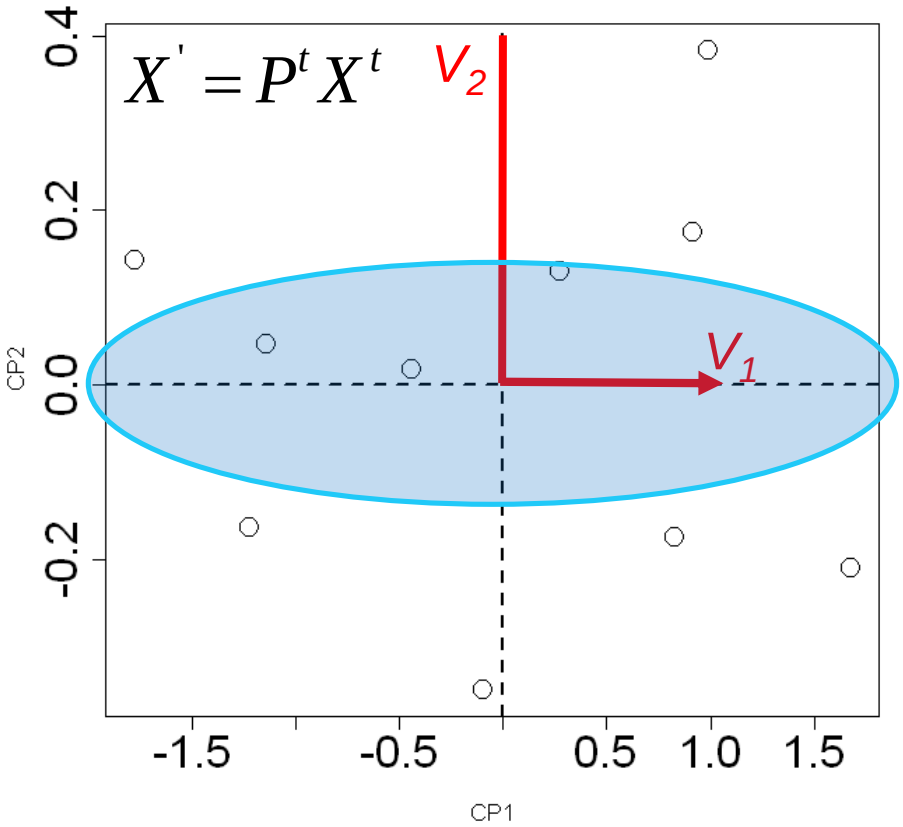
$$P = \begin{pmatrix} 0.6778736 & -0.7351785 \\ 0.7351785 & 0.6778736 \end{pmatrix}$$

$$D = \begin{pmatrix} 1.284028 & 0 \\ 0 & 0.04908323 \end{pmatrix}$$

$$C = PDP^t$$

- **V_1 et V_2 sont les vecteurs directeur de CP_1 et CP_2 :**

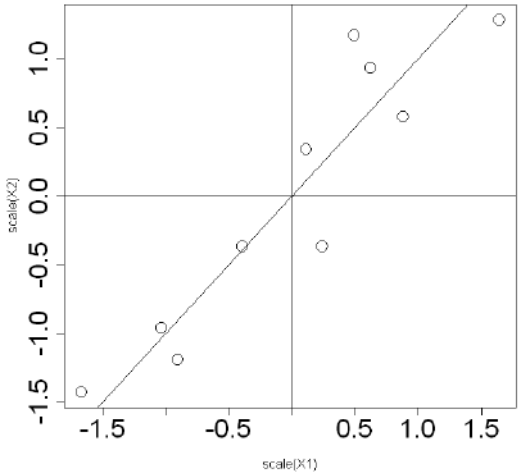




- **CP_1 porte 96% de l'inertie totale du nuage de point**
- **NB : $\rho(X_1, X_2) = 0.93$ mais $\rho(CP_1, CP_2) = 0$**

- **L’information (variance) portée par CP1 est tellement importante que l’on peut se passer de CP2 :**
  - Cela revient à compresser l’information originale portée par deux dimensions sur une seule dimension avec une perte ici de 4% de l’information d’origine
  - Par analogie, une fois que l’on a vu le chameau de profil, le voir de face n’apporte pas beaucoup plus d’information...
- **Attention :**
  - Dans le cas de l’ACP non-normée, chacune des variables représente *a priori* un poids égal à sa propre variance
  - L’ACP non-normée est une application rare et en général, on travail avec la matrice des corrélations (ACP normée)

- Cas de l'ACP normée sur le même jeu de donnée :



$$C = \begin{pmatrix} 1 & 0.93 \\ 0.93 & 1 \end{pmatrix}$$

$$C - \lambda I = \begin{pmatrix} 1-\lambda & 0.93 \\ 0.93 & 1-\lambda \end{pmatrix}$$

$$\det(C - \lambda I) = (1-\lambda)(1-\lambda) - 0.93^2$$

$$\lambda^2 - 2\lambda + 0.1351 = 0$$

$$\Delta = b^2 - 4ac$$

$$\lambda_1 = \frac{-b + \sqrt{\Delta}}{2a} = 1.93 \qquad \lambda_2 = \frac{-b - \sqrt{\Delta}}{2a} = 0.07$$

$$\lambda_1 = 1 + \rho \qquad \lambda_2 = 1 - \rho$$

- **Une solution possible (cf. sous la contrainte que V_1 et V_2 soient tout 2 des vecteurs unitaires de taille 1)**

$$V_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \quad V_2 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$$

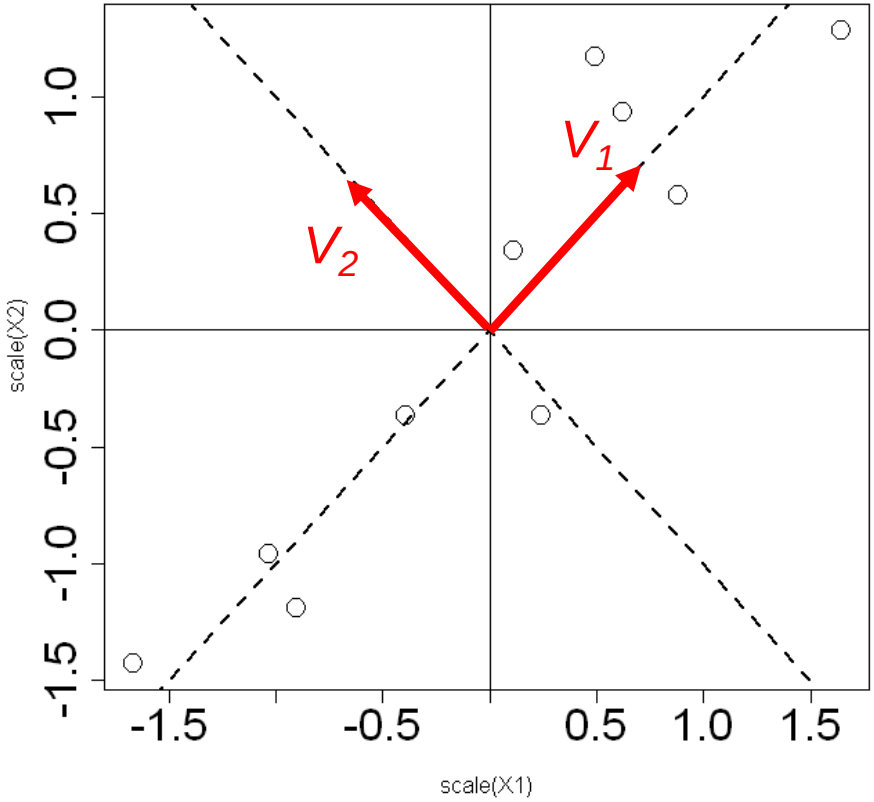
$$NB : v_{1,1}^2 + v_{1,2}^2 = 1$$

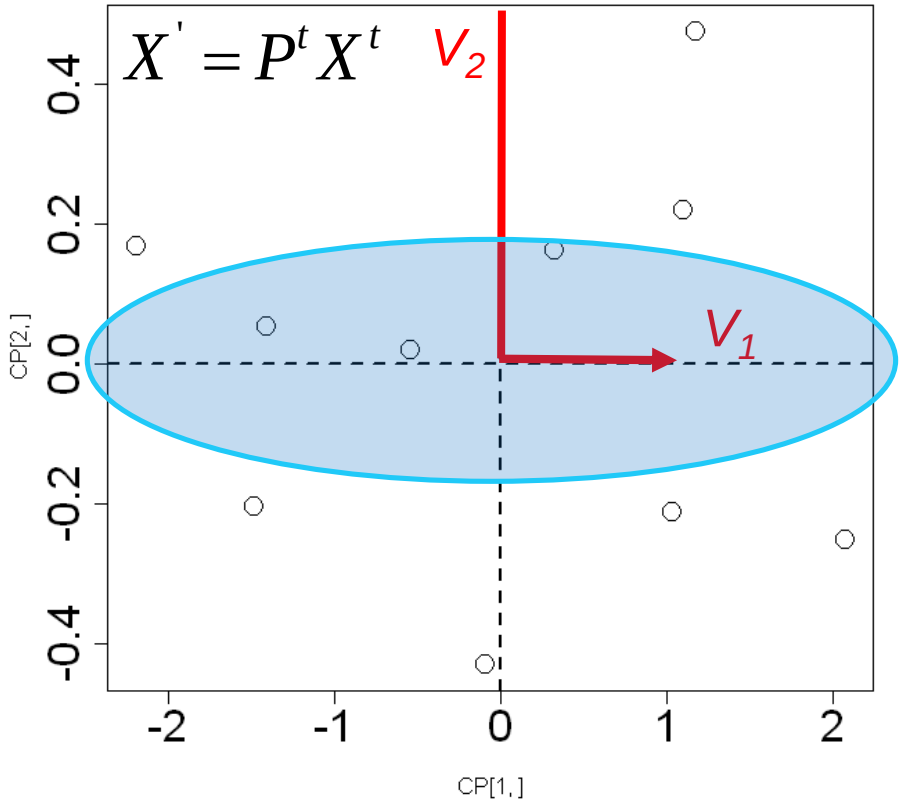
$$NB : v_{2,1}^2 + v_{2,2}^2 = 1$$

$$P = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

$$D = \begin{pmatrix} 1+\rho & 0 \\ 0 & 1-\rho \end{pmatrix}$$

- **V_1 et V_2 sont les vecteurs directeur de CP_1 et CP_2 :**





- **CP_1 porte 96% de l'inertie totale du nuage de point**
- **NB : $\rho(X_1, X_2) = 0.93$ mais $\rho(CP_1, CP_2) = 0$**

- **L'ACP d'un tableau de données à *P* variables et *N* individus admet *P* valeurs propres, *P* vecteurs propres, et *P* composantes principales :**
  - Conservez au moins 50-70% de la variance en cumulé
  - Conservez toute les composantes principale dont $\lambda > 1$ (limite de Kaiser)
  - Utilisez l'histogramme des valeurs propres (scree plot)

- **Attention :**
  - L'ACP sur un tableau de données tel que $P > N$ est impossible

Typologies des individus et des variables

- **Typologie des individus :**
  - La lecture graphique de la position des individus le long des composantes principales permet de dresser une typologie
  - Les indidus proches le long d'une composante principale sont des individus qui partagent les mêmes caractéristiques vis-à-vis des variables quantitatives étudiées
- **Typologie des variables :**
  - Chaque composante principale est une combinaison linéaire des variables de départ auxquelles sont affecté des poids
  - La lecture graphique du cercle des corrélations permet de juger du poids des différentes variables de départ sur chacune des composantes principales