Analyse de Composante principale (ACP)

Fouille de données avancées (2016-2017)

UFR MIME

Université Lille 3

7 décembre 2016

Sommaire

Motivation

Comment retrouver un système des *coordonnées* qui

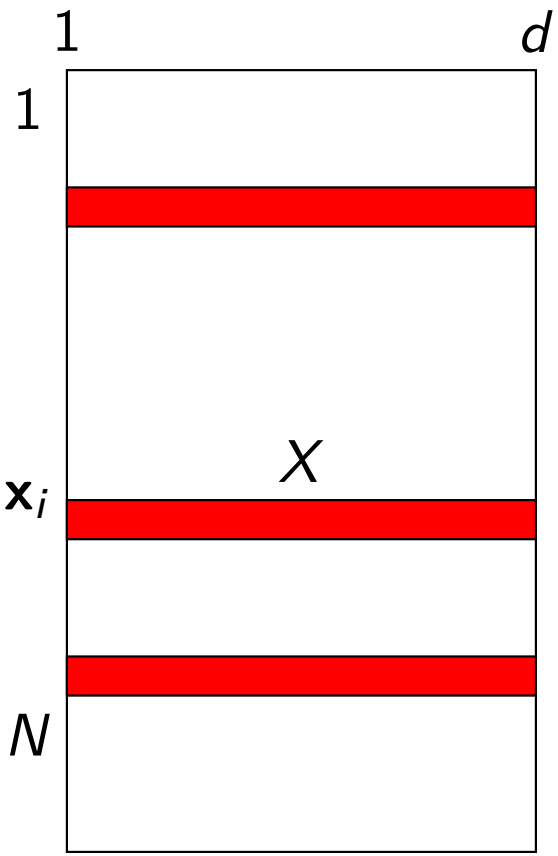
- Qui explique les sources principales de variations
  - En séparant dans des bases orthogonales (pour pouvoir analyser les composantes individuellement)
  - Qui ordonnes des bases par rapport leur contributions pour expliquer les variations

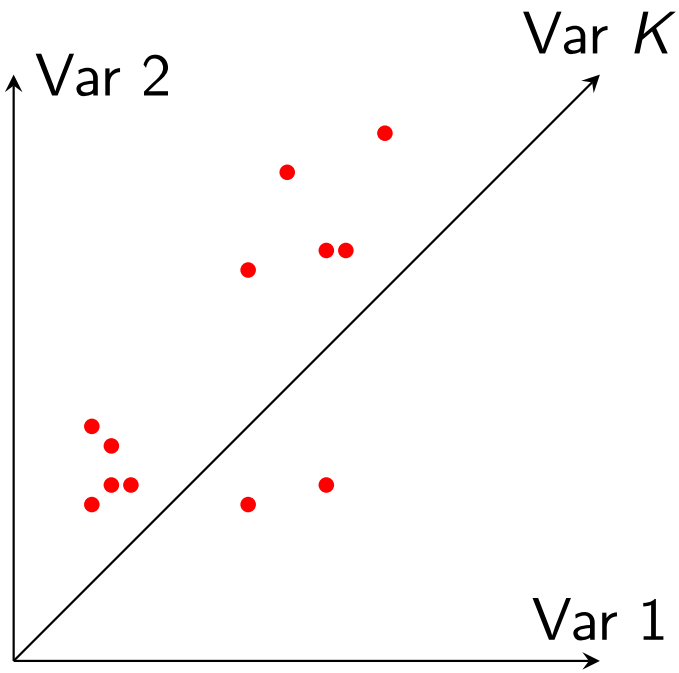
  pour un jeu de données multi-variées.

Les enjeux :

- Réduire des redondances et la dimensionnalité dans les variables avec la perte d'information minimale !
- Explorer les liaisons entre les variables

Données et analyse





- Chaque rang d'une matrice X est un vecteur $\mathbf{x}_i$ et un point dans l'espace $\mathbb{R}^d$

Applications réelles

- Analyse croisée : d-attributs et n-échantillions
  (Météorologie, biologie)
- Économie : d-valeur d'indicateur et n-périodes temporelles (mois/années etc)
- Génétique : d-gènes et n-humains ($d >> n$)
- Réduction de la dimensionnalité : d-pixel et n-images
- Finance : Retrouver les portfolios qui changent ensemble (trends)

Le problème

- Comment retrouver les vecteurs $\mathbf{v}_k \in \mathbb{R}^d$, sur lesquels les projections de tous les points (rangs) de X "maximise la variance" ? ou un système de coordonnées qui garde toute information ?
- Quel est le sens de la première composante principale (CP) ?

On peut définir la première CP comme une solution d'un problème d'optimisation :

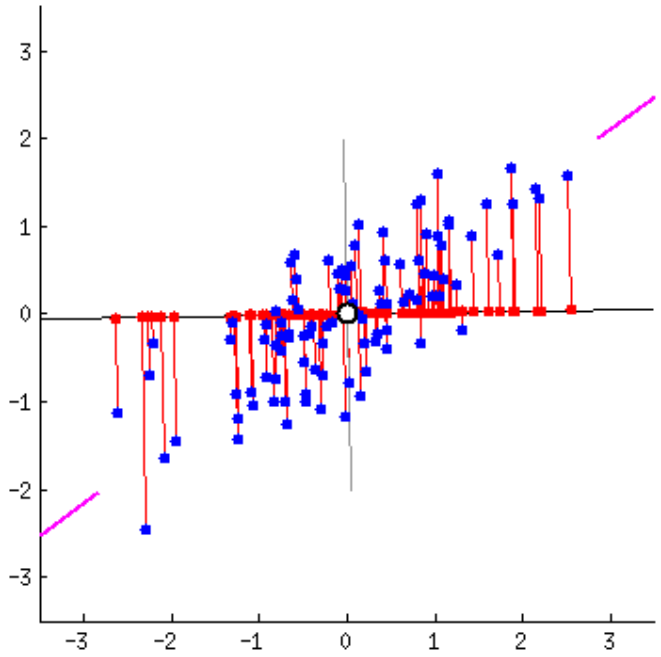
$$\max_{\|\mathbf{v}\|=1} \text{var}(\text{Proj}_{\mathbf{v}} X) = \max_{\|\mathbf{v}\|=1} \text{var}(X^T \mathbf{v}) \tag{1}$$

Maintenant, si on veut calculer une famille de vecteurs sur laquelle on doit maximiser les variances et en même temps capter les variances d'une façon "complementaire" :

$$\operatorname*{argmax}_{VV^T=I_d} \text{var}(\text{Proj}_{\mathbf{v}\in V} X) = \operatorname*{argmax}_{VV^T=I_d} \text{var}(XV) \tag{2}$$

Le critère $VV^T = I_d$ exige que les bases $\mathbf{v}_i$ soient orthogonales.

Exemple visuelle



- les points en rouge sont des projections des points de départ (en bleu)
- Les lignes rouges sont les erreurs de reconstruction

ACP minimises les distances orthogonaux aux points pour retrouver la *meilleure* ligne.

Projection et Variance

- La projection d'un vecteur $\mathbf{x}$ sur $\mathbf{v}$ est un vecteur $c \cdot \mathbf{v}$ dans la même direction que $\mathbf{v}$ celui qui est orthogonale à la résidus/différence $(\mathbf{x} - c \cdot \mathbf{v})$.
- deux vecteurs $\mathbf{a}, \mathbf{b}$ sont orthogonales si $\mathbf{a}^T\mathbf{b} = 0$.

On peut du coup calculer la projection en sachant le scalaire c.

$$(\mathbf{x} - c \cdot \mathbf{v})^T\mathbf{v} = 0 \implies c = \frac{\mathbf{x}^T\mathbf{v}}{\mathbf{v}^T\mathbf{v}} \tag{3}$$

Variance d'un vecteur x :

- Variance est une mesure de déviation d'une suite des valeurs autour de leur moyenne.
- $\text{var}(x) = \frac{1}{N}\sum_{i=1}^{N}(\bar{x} - x_i)^2$ ou $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- si on soustrait la moyenne, on peut réécrire : $\frac{1}{N}\sum_{i=1}^{N}(x_i^2)$
- $\text{var}(x) = X^TX = [x_1x_2...x_n] * [x_1x_2...x_n]^T$

Bases

Les bases standard :

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = c_1 \cdot \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + c_2 \cdot \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + c_3 \cdot \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Dans la notation matricielle :

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

ou

$$\mathbf{x} = B * \mathbf{c} \tag{4}$$

ou $\mathbf{c}$ sont des coordonnées.

Bases

Pour le problème de l'ACP, comment retrouver des bases $\mathbf{v}_i$ dans les colonnes d'une matrice $V \in \mathbb{R}^{d,r}$ avec leurs nouvelles coordonnées $\mathbf{c}$:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} | & | & | \\ \mathbf{v_1} & \mathbf{v_2} & \mathbf{v_3} \\ | & | & | \end{bmatrix} * \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix}$$

ou

$$\mathbf{x} = V * \mathbf{c} \tag{5}$$

On est en train de chercher un changement des bases depuis l'espace du départ $\mathbb{R}^d$ vers un nouvel espace avec :

- $VV^T = I_d$ les vecteurs soient orthogonales,
- Et la variance des projections soient maximale

Maximisation de la variance de projection

Le principe de l'ACP est de trouver un axe, issu d'une combinaison linéaire des vecteurs $\mathbf{x}_i \in X$, tel que la variance du nuage autour de cet axe soit maximale.
Pour la première composante principale :

$$\max_{\|\mathbf{v}\|=1} \text{var}(\text{proj}_{\mathbf{v}} X)$$

on sait que $\text{Proj}_{\mathbf{b}}\mathbf{a} = \frac{\mathbf{a}^T\mathbf{b}}{\mathbf{b}^T\mathbf{b}}$,

$$\max_{\|\mathbf{v}\|=1} \text{var}(X^T\mathbf{v})$$

et on sait que $\text{var}(x) = X^T X$,

$$\max_{\|\mathbf{v}\|=1} (X^T\mathbf{v})^T(X^T\mathbf{v}) = \max_{\|\mathbf{v}\|=1} \mathbf{v}^T XX^T\mathbf{v} = \max_{\|\mathbf{v}\|=1} \mathbf{v}^T A\mathbf{v}$$

ou $A = A^T$ est la *matrice de covariance*.

Maximisation de variance de projection (inertie)

Quant on calcule la projection sur un ensemble des vecteurs orthogonales dans les colonnes de matrice V, on peut réécrire le problème d'optimisation comme :

$$\underset{V \in \mathbb{R}^{d,r} : V^T V = I_d}{\operatorname{argmax}} \quad \operatorname{trace}(V^T A V) \tag{6}$$

ou la fonction trace() est la somme des entrées diagonales d'une matrice. Le problème d'optimisation est résolu **analytiquement** en cherchent la maximum de cette fonction. La solution est donner par une décomposition en vecteurs/valeurs propres de A, car A est une matrice **symétrique**. Il implique aussi que les meilleurs bases seront les premier $r \leq d$ vecteurs propres.

$$XX^T = A = VDV^T$$

ou V a des colonnes orthogonales et D est une matrice diagonale avec les entrées ordonnées.

Étapes de calcul

- Centrer les données X par le moyennes de chaque variable (colonnes).
- Calculer la matrice de covariance $A = \frac{1}{N} X^T X \in \mathbb{R}^{d,d}$
- Décomposer en vecteurs/valeurs propres de $A = VDV^T$
- Sélectionner les composantes principales en faisant des projections $XV \in \mathbb{R}^{n,r}$ et leur reconstructions $XVV^T \in \mathbb{R}^{n,d}$

Transformation des données (Pre-traitement)

- Centrer les données (par soustraction de la valeur moyenne de variable). C'est une rotation des données dans l'espace $\mathbb{R}^d$.
- Réduire les données si les unités de mesure sont différentes d'une variable à l'autre

$$T(\mathbf{x}_i) = \frac{x_{ik} - \bar{x}_k}{s_k} \tag{7}$$

- si on ne réduit pas le nuage : une variable à forte variance va « tirer » tout l'effet de l'ACP à elle
- si on réduit le nuage : une variable qui n'est qu'un bruit va se retrouver avec une variance apparente égale à une variable informative.

Matrice de covariance

Pour une matrice d'entrée X sa matrice de covariance est une matrice qui contient des covariances pour chaque paire des variables i, j.

$$\Sigma_{ij} = \text{cov}(X_i, X_j) = \mathbf{E}[(X_i - \mu_i)(X_j - \mu_j)] \tag{8}$$

ou, $\mu_i = \mathbf{E}(X_i)$ est la moyenne à travers des échantillions pour la variable i. La diagonale de cette matrice contient des variances de chaque variable.

Centrage

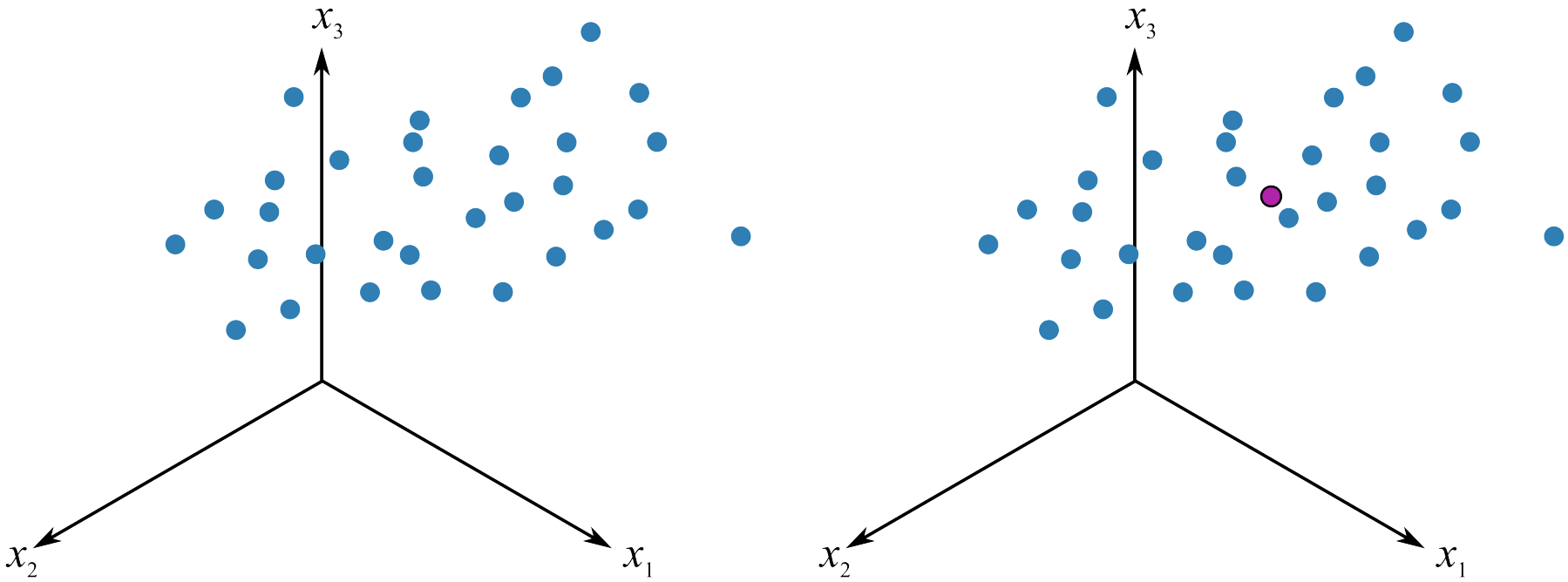


FIGURE – Centrage par les moyennes.

Première composante

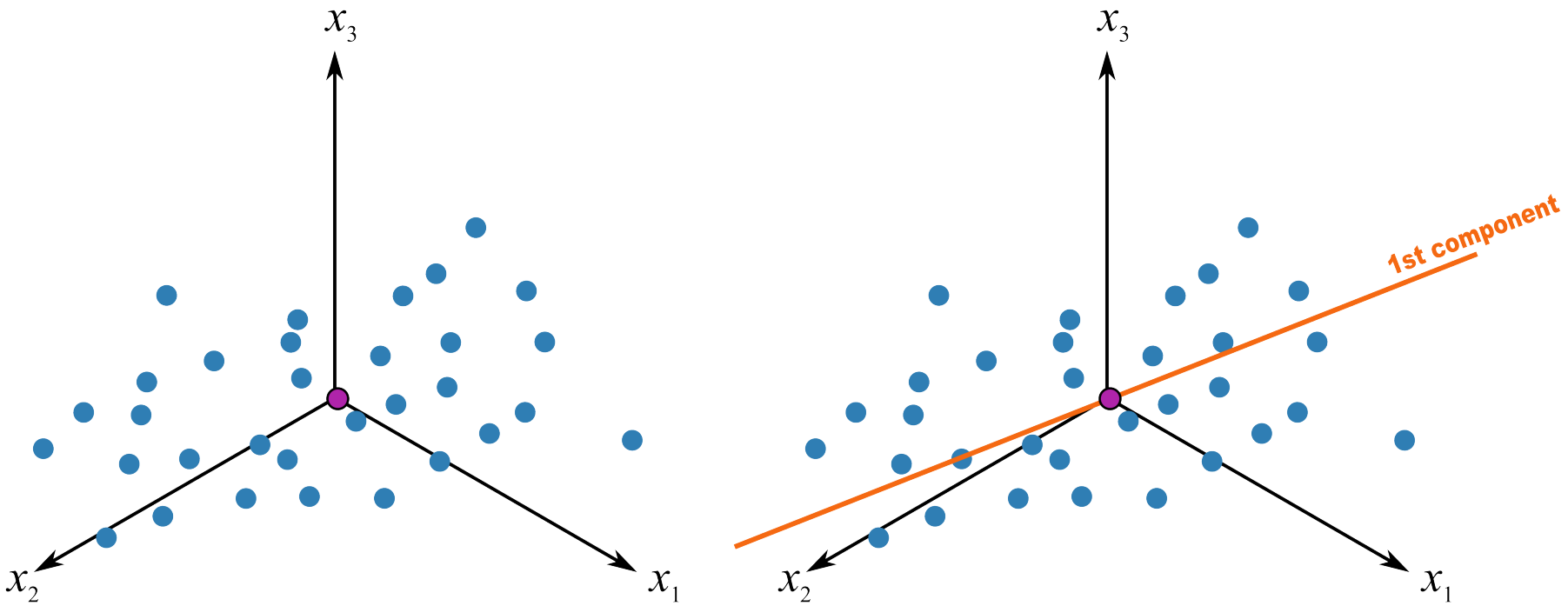


FIGURE – Meilleure ligne (Première direction principale) avec la moindre erreur de reconstruction.

Deuxième composante

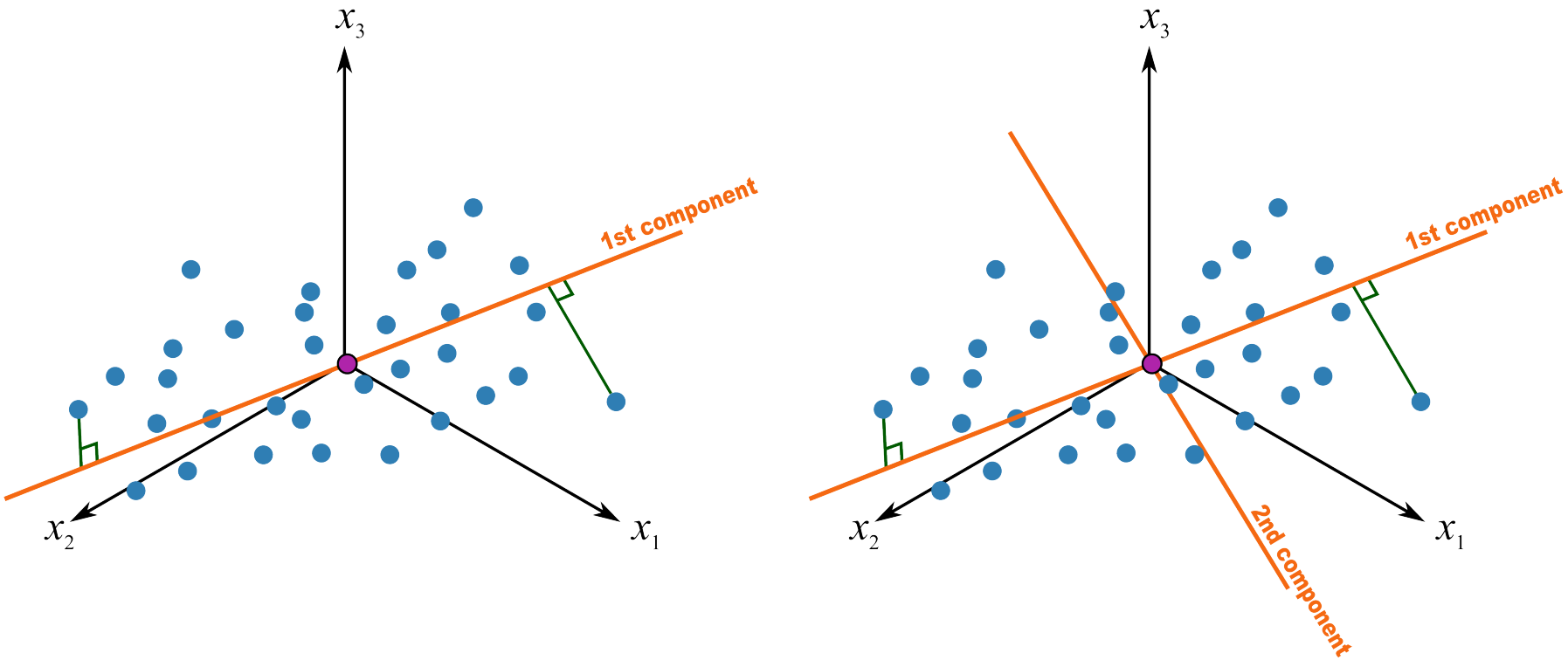


FIGURE – la première composante (projections sur la première direction) et la deuxième composante.

Deuxième composante

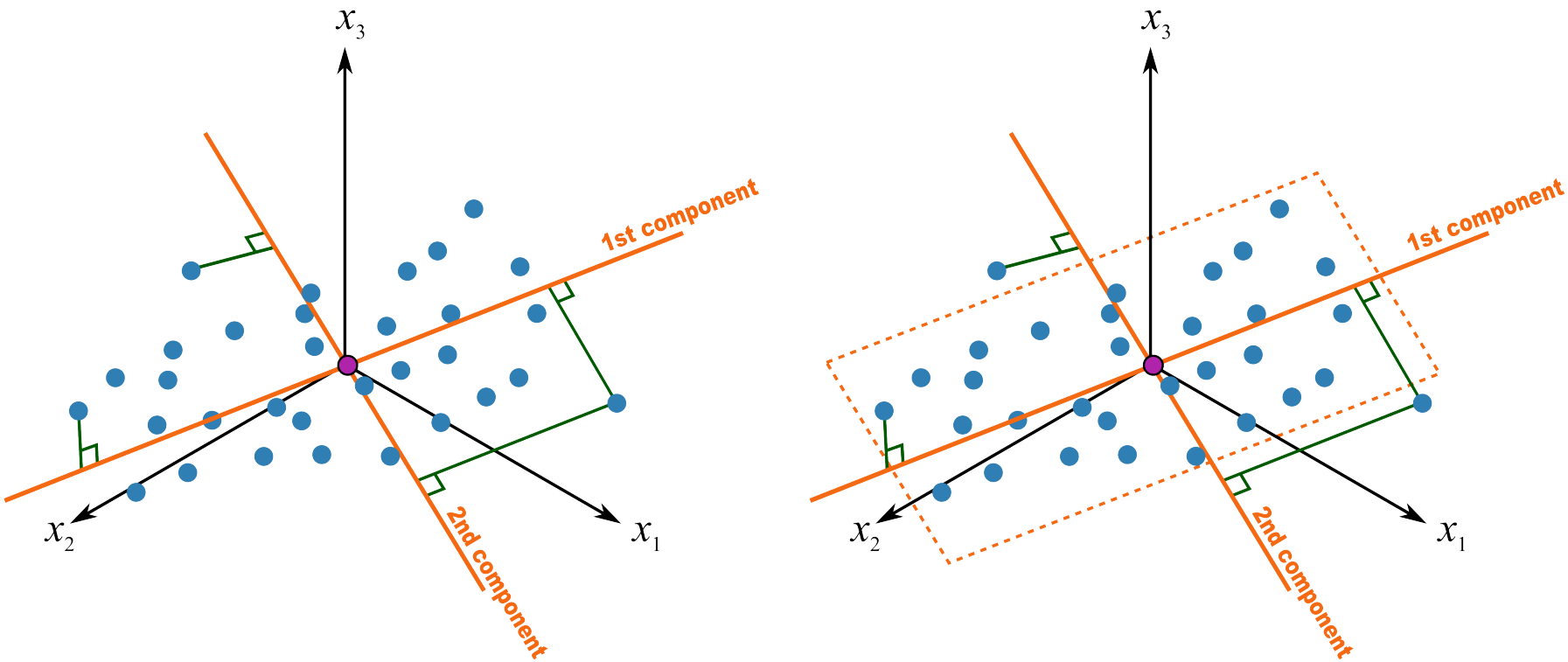


FIGURE – Première et deuxième composantes avec leurs projections.

Instructions : TP

Fonctions utiles :

- var() cov() cor() : pour calculer les variances, covariances et corrélations
- scale(x, center = TRUE, scale = TRUE) : Pour centrer et réduire les données
- prcomp() : Pour calculer les composantes principales en R
- eigen() : Pour calculer la decomposition eigen en R
- Sorties de prcomp() : sdev (les variances pour chaque composante), rotation (le matrice V avec les colonnes qui sont les vecteur propres), x (les projections de X sur V)
- summary() : summary(objetACP) vous rends une résume de la décomposition en ACP effectuer.
- pairs() : Pour plotter les nuages des points entre chaque paires de variables

Instructions : TP

Taches :

- Calculer le matrice de covariance pour une entrée X.
- Calculer le ACP d'une matrice d'entrée par la fonction **prcomp**.
- Maintenant produire les résultats d'ACP en calculant la matrice de covariance et sa décomposition avec **eigen** et vérifier que vous avez produit la même résultats que la fonction **prcomp**.

Conclusions

- Le principe de l'ACP est de trouver un axe, issu d'une combinaison linéaire des $\mathbf{x}_i$, tel que la variance du nuage autour de cet axe soit maximale. Ceci est appliquer itérativement sur la différence entre la projection sur l'axe principale est données de départ. La solution analytique existe, et un des solution efficace est de calculer le SVD de matrice de covariance.
- Finalement, la question de l'ACP se ramène à un problème de diagonalisation de la matrice de covariance ou corrélation.
- La diagonalisation de la matrice de corrélation (ou de covariance si on se place dans un modèle non réduit), nous a permis d'écrire que le vecteur qui explique le plus d'inertie du nuage est le premier vecteur propre. De même le deuxième vecteur qui explique la plus grande part de l'inertie restante est le deuxième vecteur propre, etc.
- Proportion de variance expliquée par le k-ième vecteur propre vaut λ_k