Analyse en composantes principales

Christine Decaestecker & Marco Saerens

ULB & UCL

A.C.P.: Analyse en Composantes Principales

- **Analyse de la structure de la matrice variance-covariance**
  c-à-d de la variabilité, dispersion des données.

  Excepté si l'une des variables peut s'exprimer comme une fonction d'autres, on a besoin des p variables pour prendre en compte toute la variabilité du système

  Objectif de l'ACP: ***décrire*** à l'aide de $q < p$ composantes ***un maximum*** de cette variabilité.

  - Ce qui permet :
    - une réduction des données à q nouveaux descripteurs
    - une visualisation des données à 2 ou 3 dimensions (si $q = 2$ ou 3)
    - une interprétation des données : liaisons inter-variables

  - Etape intermédiaire souvent utilisée avant d'autres analyses !

- **Recherche des composantes principales**

Composantes : C_1 , C_2 ,…, C_k ,…, C_q

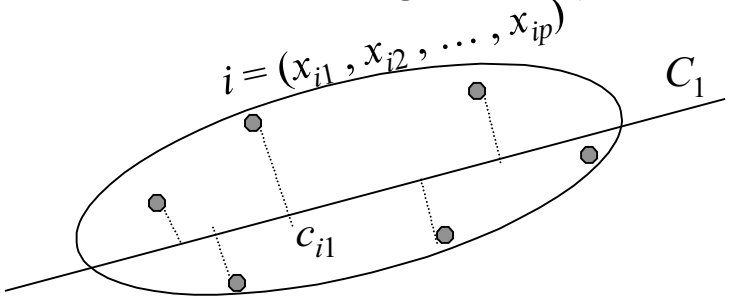
C_k = nouvelle variable = combinaison linéaire des variables d'origine X_1,…, X_p:

$C_k = a_{1k} X_1 + a_{2k} X_2 + \ldots\ldots + a_{pk} X_p$ ⟶ coefficients a_{jk} à déterminer

telle que les C_k soient: - 2 à 2 non corrélées,
- de variance maximale,
- d'importance décroissante.

C_1 = 1ère composante principale doit être de variance maximale

Géométriquement : C_1 détermine une nouvelle direction dans le nuage de points qui suit l'axe d'allongement (étirement) maximal du nuage.



c_{i1} = coordonnée du point i sur l'axe C_1

projection de $\mathbf{x}_i$ sur C_1

$$c_{i1} = \sum_{j=1}^{p} a_{1j} x_{ij}$$

C_1 de variance maximale ⟹ les projections c_{i1} sont les plus dispersées possible.

Pour fixer la droite, on impose qu'elle passe par **g** (centre de gravité) !

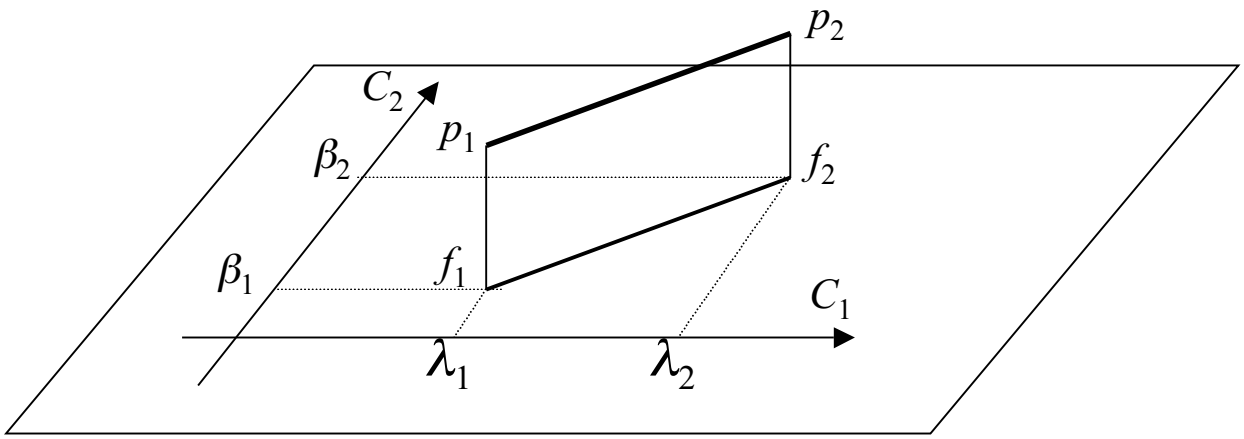
(sinon l'ensemble des droites parallèles conviennent)

C_1 = droite passant par le centre de gravité réalisant le meilleur ajustement possible du nuage c-à-d : qui conserve au mieux la distance entre les points (après projection)
=> droite de projection assurant une distorsion minimale.

C_2 = 2ème composante, orthogonale à C_1 et de variance maximale.

Géométriquement : C_2 détermine une droite perpendiculaire à C_1 (au point g), suivant un axe (perpendiculaire au 1er) d'allongement maximum.

=> C_1 et C_2 déterminent le plan principal : le meilleur plan de projection (de distorsion minimum).



C_1 est telle que la moyenne des $d^2(\lambda_i, \lambda_{i'})$ max.

C_2 est $\perp$ à C_1 et telle que la moyenne des $d^2(\beta_i, \beta_{i'})$ max.

=> C_1 et C_2 déterminent le plan tel que $d^2(\mathbf{f}_i, \mathbf{f}_{i'})$ soit maximum.

=> C_3 est la droite $\perp$ à C_1 et C_2 (par g) telle que la variance des coord. soit maximum ...

Décomposition de la variance

- La variance se décompose de la manière suivante

$$
\begin{aligned}
\sigma^2 &= \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} (\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} (\mathbf{x}_i - \mathbf{x}_j) \\
&= \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} ((\mathbf{x}_i - \mathbf{g}) + (\mathbf{g} - \mathbf{x}_j))^{\mathrm{T}} ((\mathbf{x}_i - \mathbf{g}) + (\mathbf{g} - \mathbf{x}_j)) \\
&= \frac{1}{2n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \Big((\mathbf{x}_i - \mathbf{g})^{\mathrm{T}} (\mathbf{x}_i - \mathbf{g}) + (\mathbf{g} - \mathbf{x}_j)^{\mathrm{T}} (\mathbf{g} - \mathbf{x}_j) + 2(\mathbf{x}_i - \mathbf{g})^{\mathrm{T}} (\mathbf{g} - \mathbf{x}_j) \Big) \\
&= \frac{1}{(n-1)} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{g})^{\mathrm{T}} (\mathbf{x}_i - \mathbf{g})
\end{aligned}
$$

Projection sur une droite

- L'opérateur de projection orthogonale, $\boldsymbol{\pi}$, sur une droite de vecteur directeur unitaire $\mathbf{v}$ s'écrit

$$\boldsymbol{\pi} = \mathbf{v}\mathbf{v}^{\mathrm{T}}$$

- Avec

$$\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1$$

- La variance des observations projetées s'écrit alors

$$\sigma^2_{\mathbf{v}} = \frac{1}{(n-1)} \sum_{i=1}^{n} (\boldsymbol{\pi}(\mathbf{x}_i - \mathbf{g}))^{\mathrm{T}} (\boldsymbol{\pi}(\mathbf{x}_i - \mathbf{g}))$$

Recherche de la projection de variance maximale

- Nous avons donc

$$
\begin{aligned}
\sigma^2_{\mathbf{v}} &= \frac{1}{(n-1)} \sum_{i=1}^{n} (\boldsymbol{\pi}(\mathbf{x}_i - \mathbf{g}))^{\mathrm{T}} (\boldsymbol{\pi}(\mathbf{x}_i - \mathbf{g})) \\
&= \frac{1}{(n-1)} \sum_{i=1}^{n} (\mathbf{v}\mathbf{v}^{\mathrm{T}}(\mathbf{x}_i - \mathbf{g}))^{\mathrm{T}} (\mathbf{v}\mathbf{v}^{\mathrm{T}}(\mathbf{x}_i - \ \mathbf{g})) \\
&= \frac{1}{(n-1)} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{g})^{\mathrm{T}} (\mathbf{v}\mathbf{v}^{\mathrm{T}}\mathbf{v}\mathbf{v}^{\mathrm{T}}) (\mathbf{x}_i - \mathbf{g}) \\
&= \frac{1}{(n-1)} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{g})^{\mathrm{T}} \mathbf{v}\mathbf{v}^{\mathrm{T}} (\mathbf{x}_i - \mathbf{g})
\end{aligned}
$$

Recherche de la projection de variance maximale (suite)

- Et donc

$$\begin{aligned}
\sigma_{\mathbf{v}}^2 &= \frac{1}{(n-1)} \sum_{i=1}^{n} ((\mathbf{x}_i - \mathbf{g})^{\mathrm{T}} \mathbf{v})(\mathbf{v}^{\mathrm{T}} (\mathbf{x}_i - \mathbf{g})) \\
&= \frac{1}{(n-1)} \sum_{i=1}^{n} \mathbf{v}^{\mathrm{T}} (\mathbf{x}_i - \mathbf{g})(\mathbf{x}_i - \mathbf{g})^{\mathrm{T}} \mathbf{v} \\
&= \frac{1}{(n-1)} \mathbf{v}^{\mathrm{T}} \left[\sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{g})(\mathbf{x}_i - \mathbf{g})^{\mathrm{T}} \right] \mathbf{v} \\
&= \mathbf{v}^{\mathrm{T}} \boldsymbol{\Sigma} \mathbf{v}
\end{aligned}$$

- Observons que la matrice $\boldsymbol{\Sigma}$ est la matrice variance-covariance
- Cette matrice est symétrique définie positive

Recherche de la projection de variance maximale (suite)

- Nous devons donc maximiser cette variance des observations projetées

$$\max_{\mathbf{v}}(\mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v}) \text{ subject to } (\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1)$$

- Il s'agit d'un problème d'optimisation sous contrainte
- Nous formons donc la fonction de Lagrange

$$\pounds = \mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v} + \lambda(1 - \mathbf{v}^{\mathrm{T}}\mathbf{v})$$

- Et nous calculons les conditions nécessaires d'optimalité

$$\partial_{\mathbf{v}}\pounds = 0$$

Recherche de la projection de variance maximale (suite)

- Nous obtenons ainsi l'équation aux valeurs propres

$$\boldsymbol{\Sigma}\mathbf{v} = \lambda\mathbf{v}$$

- Comme la matrice variance-covariance est symétrique définie positive, les valeures propres sont réelles positives
- Les vecteurs propres peuvent être choisis orthonormés

Recherche de la projection de variance maximale (suite)

- La variance des observations projetées s'écrit alors

$$\begin{aligned} \sigma_{\mathbf{v}}^2 &= \mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v} \\ &= \mathbf{v}^{\mathrm{T}}\lambda\mathbf{v} \\ &= \lambda \end{aligned}$$

- Et donc la solution est de projeter les données sur le vecteur propre ayant la valeur propre λ la plus élevée

Recherche des projections de variance maximale orthogonales au premier axe

- Afin de trouver le second axe de variance maximale, nous recherchons

$$\max_{\mathbf{v}}(\mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v}) \text{ subject to } (\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1) \text{ and } (\mathbf{v}^{\mathrm{T}}\mathbf{v}_1 = 0)$$

- Avec $\mathbf{v}_1$ étant le premier vecteur propre à valeur propre maximale
- Comme les vecteurs propres de $\boldsymbol{\Sigma}$ sont naturellement orthonormés, la solution est de choisir le deuxième vecteur propre de $\boldsymbol{\Sigma}$ (à deuxième valeur propre maximale)

Matrice variance-covariance

- Notons que si $\mathbf{X}$ est la matrice de données
- Qui contient les vecteurs $(\mathbf{x}_i - \mathbf{g})^T$ en ligne
- La matrice $\mathbf{\Sigma} = (n-1)^{-1}\, \mathbf{X}^T\mathbf{X}$

Interprétation des valeurs propres

- La somme des valeurs propres correspond à la variance totale

$$\mathrm{tr}(\mathbf{\Sigma}) = \sigma^2 = \sum_{i=1}^{n} \lambda_i$$

- Chaque valeur propre mesure la part de variance expliquée par l'axe factoriel correspondant

Approche alternative de la PCA

- Nous présentons une approche alternative de la PCA
- En utilisant cette fois-ci la notion de vecteur aléatoire
  - Et donc en ne partant pas des données empiriques comme précédemment
- Soit $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ le vecteur aléatoire des n variables aléatoires (caractéristiques) mesurées sur les individus
- Nous définissons une nouvelle variable y qui est une combinaison linéaire des variables aléatoires x_i

$$y = \mathbf{v}^{\mathrm{T}}\mathbf{x}$$

- Nous supposons que $\mathbf{v}$ est normalisé

$$\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1$$

Approche alternative de la PCA

- Nous recherchons la projection du vecteur aléatoire **x** qui maximise la variance projetée:

$$\mathrm{var}(y) = \mathrm{E}[(y - \mathrm{E}[y])^2]$$

- Calculons d'abord la moyenne de *y*

$$\begin{aligned} \mathrm{E}[y] &= \mathrm{E}[\mathbf{v}^{\mathrm{T}}\mathbf{x}] \\ &= \mathbf{v}^{\mathrm{T}}\,\mathrm{E}[\mathbf{x}] \\ &= \mathbf{v}^{\mathrm{T}}\mathbf{g} \end{aligned}$$

Approche alternative de la PCA

- Et ensuite la variance:

$$\begin{aligned} \mathrm{var}(y) &= \mathrm{E}[(y - \mathrm{E}[y])^2] \\ &= \mathrm{E}[(\mathbf{v}^{\mathrm{T}}\mathbf{x} - \mathbf{v}^{\mathrm{T}}\mathbf{g})^2] \\ &= \mathrm{E}[\mathbf{v}^{\mathrm{T}}(\mathbf{x} - \mathbf{g})\mathbf{v}^{\mathrm{T}}(\mathbf{x} - \mathbf{g})] \\ &= \mathbf{v}^{\mathrm{T}}\,\mathrm{E}[(\mathbf{x} - \mathbf{g})(\mathbf{x} - \mathbf{g})^{\mathrm{T}}]\mathbf{v} \end{aligned}$$

- Il faut donc calculer le maximum de cette variance par rapport à **v**, ce qui nous ramène au problème d'optimisation suivant (le même que pour l'approche précédente)

$$\max_{\mathbf{v}}(\mathbf{v}^{\mathrm{T}}\mathbf{S}\mathbf{v}) \text{ subject to } (\mathbf{v}^{\mathrm{T}}\mathbf{v} = 1)$$

Approche alternative de la PCA

- Avec **S** étant la matrice variance-covariance (notons que **Σ** était la matrice variance-covariance empirique):

$$\mathbf{S} = \mathrm{E}[(\mathbf{x} - \mathbf{g})(\mathbf{x} - \mathbf{g})^{\mathrm{T}}]$$

- Et **S** peut être estimé à partir de l'échantillon par

$$\begin{aligned} \mathbf{S} &\simeq \frac{1}{(n-1)} \sum_{i=1}^{n} (\mathbf{x}_i - \mathbf{g})(\mathbf{x}_i - \mathbf{g})^{\mathrm{T}} \\ &\simeq \boldsymbol{\Sigma} \end{aligned}$$

- Nous sommes donc amenés, comme précédemment, à calculer les valeurs/vecteurs propres de **Σ**

Résultats:

L'ACP remplace les p variables de départ (variances ≠, corrélation inter-variable)
en q nouvelles composantes ($q \leq p$) C_k

- orthogonales 2 à 2 c-à-d cov(C_k , $C_{k'}$) = 0 (pour tout $k \neq k'$), et
- de variances maximales

On peut noter que

- $V(C_1) \geq V(C_2) \ \ldots \geq V(C_q)$ ™ d'importance décroissante
- le nombre maximum de composantes principales $q \leq p$

  avec $q < p$ dès que l'une des variables d'origine est une combinaison linéaire d'autres!

⇒ mise en évidence de relations linéaires dans les données

⇒ les données occupent, en réalité, un sous-espace de dimensions réduites ($q < p$)

Le nombre maximum de composantes principales = dimension intrinsèque des données

- **Choix des *r* premières composantes principales**

  $r << p$ ⟶ réduction de la dimension

  objectif : garder un maximum d'information des données initiales.

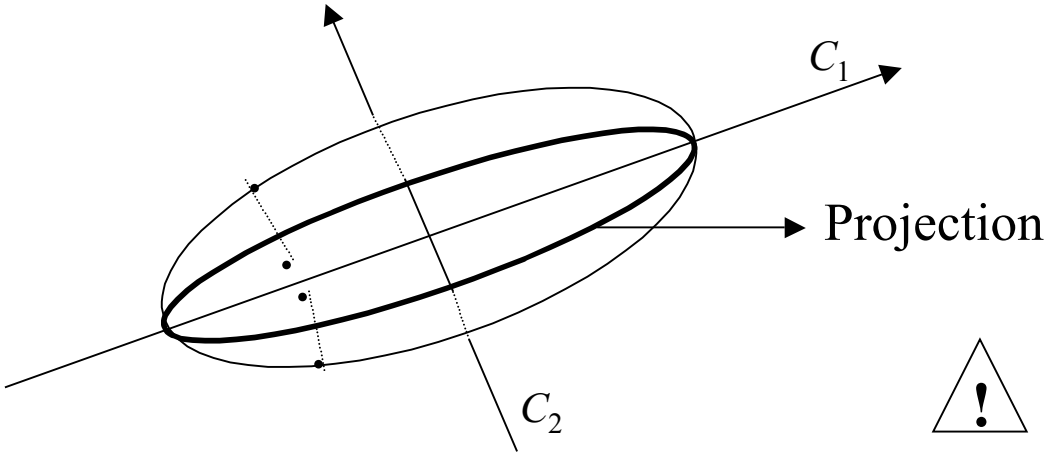
  Mesure de cette information : le % de variance expliquée

$$= \frac{\sum_{k=1}^{r} V(C_k)}{Inertie \quad totale}$$

  Si les variables originales sont fortement corrélées entre elles, un nombre réduit de composantes permet d'expliquer 80% à 90% de variance !

  Géométriquement : revient à projeter les données dans un sous-espace de dimension *r*, centré sur *g*, reprenant les *r* premiers axes principaux d'allongement du nuage ! ⇒ les projections c_{ij} sont les plus dispersées possibles !!

Exemple : données initiales à 3 dimensions distribuées dans un « ballon de rugby »



proximité sur le plan C_1 , C_2 $\not\Rightarrow$
proximité dans l'espace initial

Plus le nuage est aplati sur C_1 , C_2 $\Rightarrow$ moins de variance sur la 3iè dimension.
$\Rightarrow$ % de variance expliquée par C_1 , C_2

<u>En général</u> :

- Le % de variance expliquée par C_1 , C_2 , … , C_r = mesure d'aplatissement du nuage sur le sous-espace des composantes (à r dim.). Plus ce % est grand, meilleure est la représentation des données dans le sous-espace !

- Les composantes principales sont entièrement déterminées par la matrice **V** variance-covariance (vecteurs propres).

=> toute modification de **V** $\longrightarrow$ modification des composantes !!

Remarques :

- Si certaines variables initiales sont très dispersées (σ_j^2), elles vont prendre le pas sur les autres.

  => les composantes principales tenteront essentiellement d'expliquer la variance due à ces variables !

  => on peut travailler en données réduites (variables normalisées par s_j)

  => toutes les variables auront la même importance (il se peut qu'on perde de l'information)

  > données centrées-réduites $z_{ij} = \frac{(x_{ij} - g_j)}{\sigma_j}$

  => matrice variance-covariance = **R** et l'ACP explique la structure de **R** !

- Autre possibilité : travailler sur les rangs

  => ACP non-paramétrique

  => plus robuste : - pour des données très hétérogènes
  - aux dissymétries des distributions
  - aux valeurs extrêmes ! (augmente anormalement la variance !)

  => permet d'intégrer des variables qualitatives ordinales !

- **Contributions des variables aux composantes**

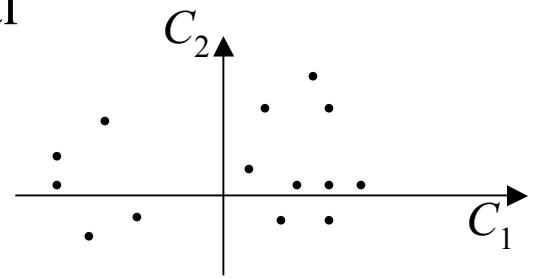
  Composante = combinaison linéaire des variables : $C_k = a_{1k}X_1 + a_{2k}X_2 + \ldots. + a_{pk}X_p$

  coeff. a_{jk} = contribution de la variable X_j à la composante C_k

- **Interprétation des résultats**

  1/ Représentation des individus dans le plan principal

  => peut faire apparaître des groupes d'individus présentant des similitudes.



  ! proximités abusives dues aux projections

  => la représentation n'est valable que si le % de variance expliquée par C_1 et C_2 est suffisamment grand ! (nuage assez aplati sur le plan)

  => vérifier si les proximités se maintiennent dans d'autres plans de projection: C_1 - C_3 , C_2 - C_3 , …

  les individus les mieux représentés: points proches du plan (projection peu importante).

2/ Interprétation des composantes principales

corrélations avec les variables initiales

	C_1	C_2	C_3	…
X_1	r_{11}	r_{12}	r_{13}	…
X_2	r_{21}	r_{22}	$_{23}$	…
⋮	⋮	⋮	⋮	…
X_p	r_{p1}	r_{p2}	r_{p3}	…

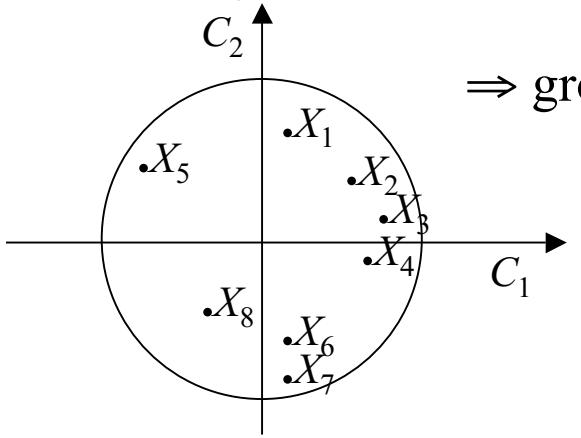
repérer les variables très corrélées

($r \approx 1$ ou $r \approx -1$)

Interprétation des 2 premières composantes C_1 , C_2 : <u>cercle des corrélations</u> :

C_1 et C_2 étant non-corrélées, on a $r^2 (c_1, x_j) + r^2 (c_2, x_j) \leq 1$

=> chaque variable représentée par les coordonnées : ($r (c_1 , x_j)$, $r (c_2 , x_j)$) est dans un cercle de rayon 1

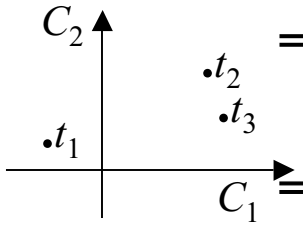


⇒ groupes de variables liées ou opposées

⚠ si proches de la circonférence, bien représentées par les 2 composantes !

3/ Projection de points supplémentaires sur le plan principal après le calcul des composantes

- <u>individus typiques</u> de groupes d'individus : exemple t_1 , t_2 , t_3 pour 3 groupes ≠



⇒ permet de voir si les composantes C_1 et C_2 sont à même de les distinguer

⇒ idée de ce qui est distinguable (t_1 des autres) ou non (t_2, t_3)!