L'analyse en Composantes principales

Stéphane Canu
stephane.canu@litislab.eu

M8 - Principes du traitement de l'information

March 13, 2017

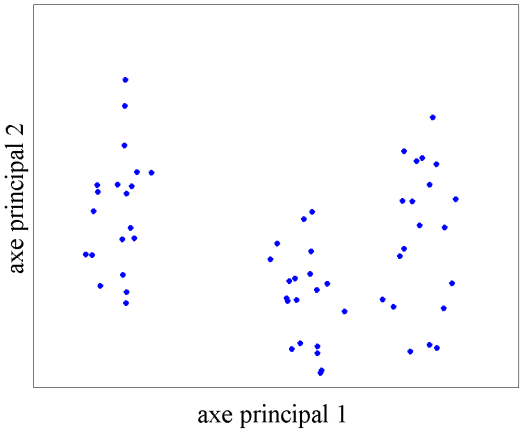
Plan

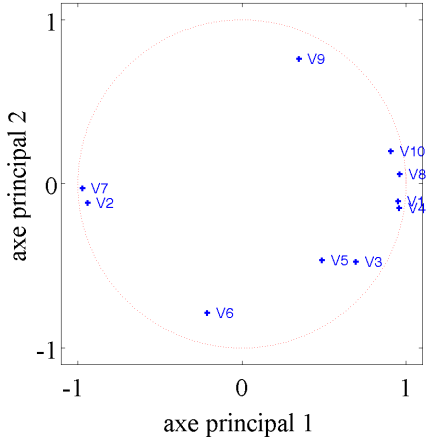
Analyse en composante principale (ACP)

	AGE	SEX	BMI	BP	··· Serum Measurements ···						Response
Patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

Table 1. Diabetes study. 442 diabetes patients were measured on 10 baseline variables. . prediction model was desired for the response variable, a measure of disease progression one yea after baseline.

$$X \Rightarrow \boxed{\text{A.C.P.}} \Rightarrow (U, V, \lambda)$$

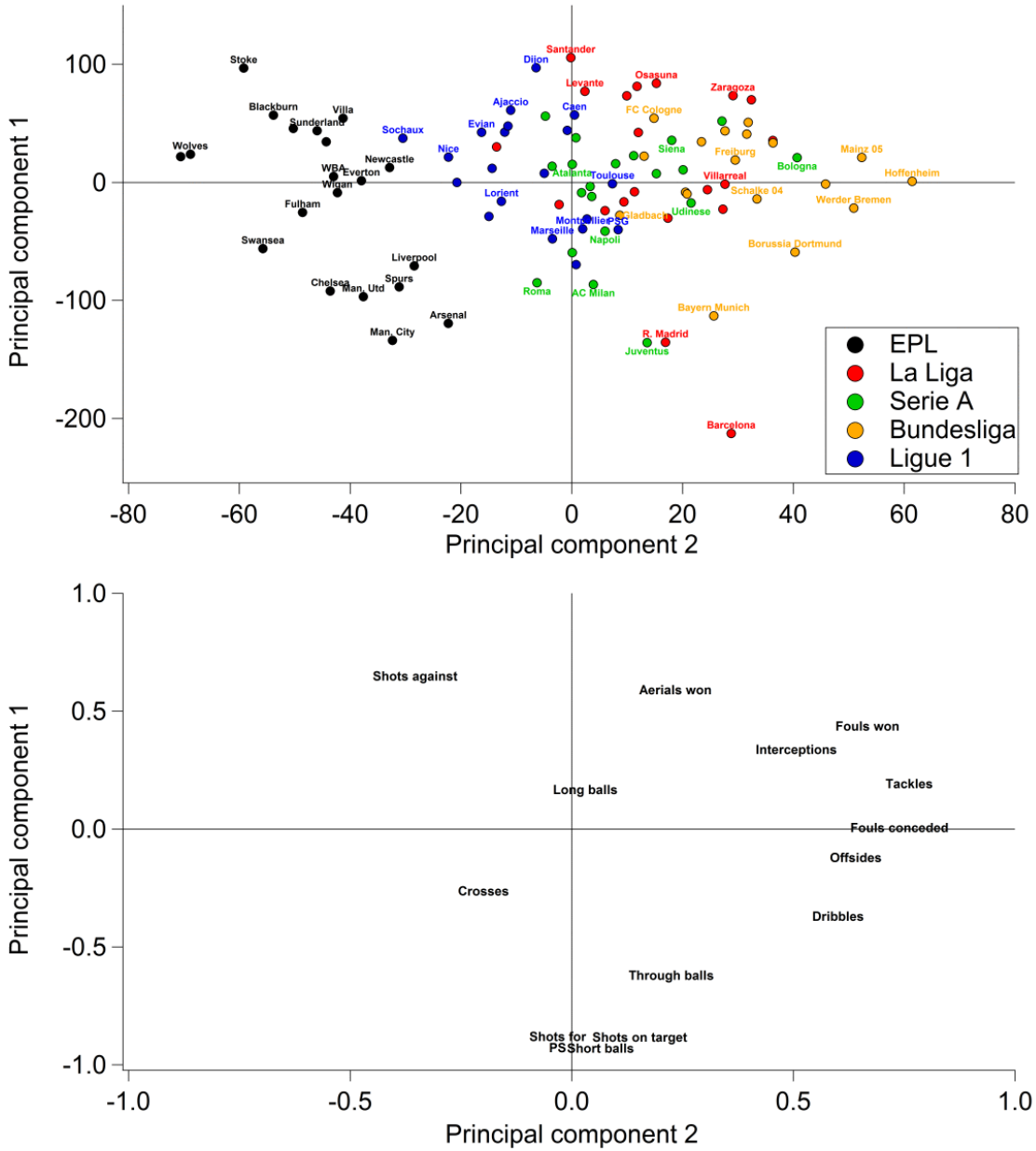




Les 3 étapes de l'ACP

1. centrer et réduire les données `Xn = (X - un*mX)./(un*sX)`
2. calculer les valeurs et vecteurs propres `[V,λ] = eig(Xn'*Xn)`
3. projeter les observations `U = Xn*V`

Analyse en composante principale du Foot

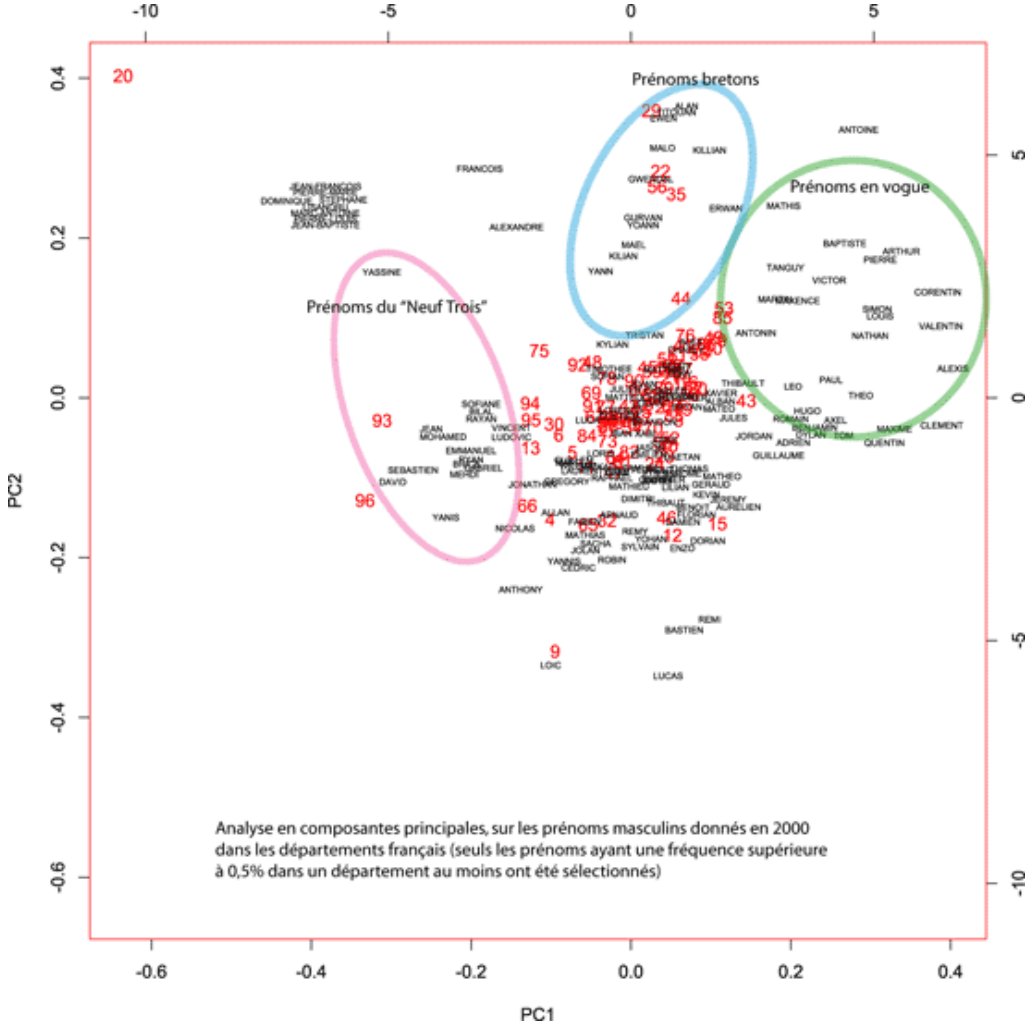


http://2plus2equals11.com/category/principal-component-analysis/

Analyse en composante principale des prénoms

n = 99 départements

p = fréquence des prénoms

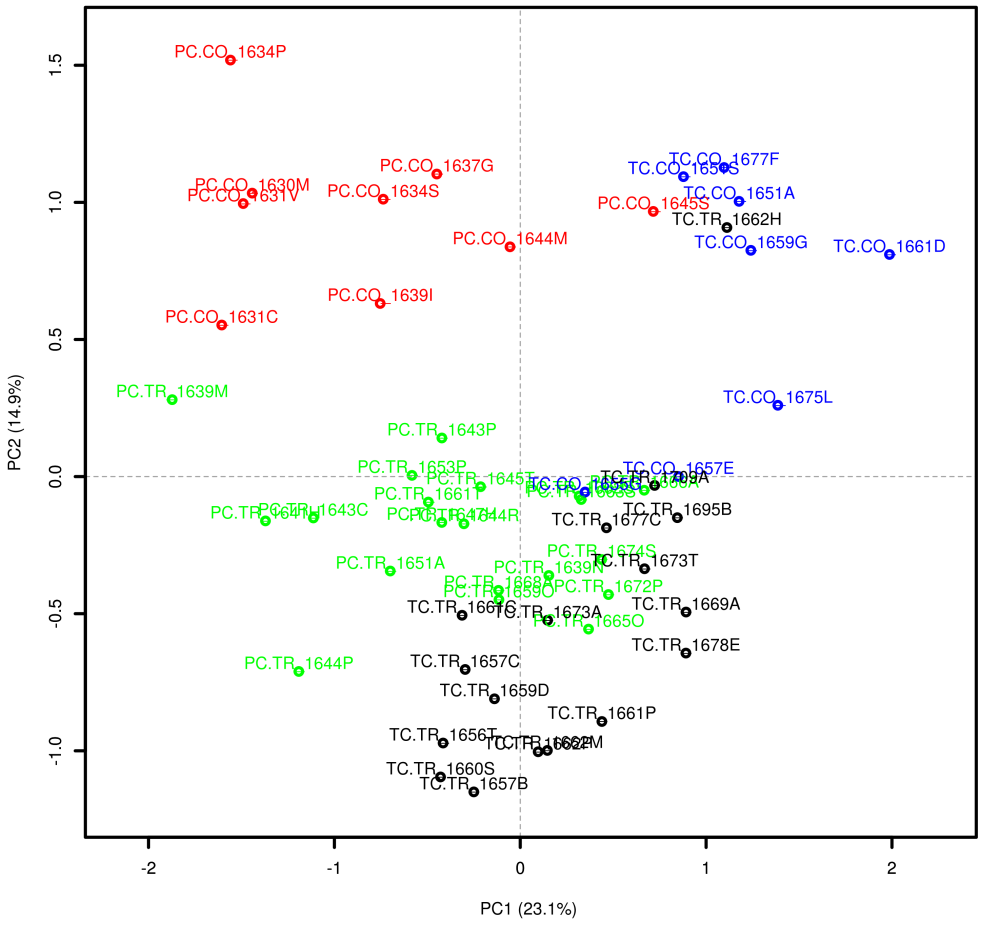


http://coulmont.com/blog/2010/07/27/un-joli-dessin/

Analyse en composante principale des frères Corneille

n = 32 pièces écrites par Thomas et Pierre Corneille

p = mots (frère, mort, ces. . .)



http://dragonfly.hypotheses.org/472

Xn =

-0.3393	-0.5775	0.9802	-0.4268	-0.0513	0.9483	0.4674
-1.5740	0.0188	-0.0172	0.8806	-0.7188	1.5195	-0.0684
0.5678	-1.0150	1.9092	1.0828	-1.3888	1.7471	-0.2692
-0.9404	0.5585	1.2807	-0.1368	-0.9082	1.4323	-0.0584
-0.5138	0.4148	1.3217	-0.8962	-1.0307	0.2355	-0.2256
-0.9136	-0.2863	1.0695	0.1055	-0.7645	0.8271	0.7587
-1.2495	-0.9377	1.6086	0.1441	-1.7061	1.6852	0.2501
-0.9139	0.2438	1.1964	-0.4372	-2.2392	1.0304	-0.1304
0.4414	-0.6329	1.1367	0.7691	-1.4089	1.4982	-0.2027
-0.9298	0.7957	1.0126	0.3820	-1.5871	0.5316	-0.1956
-0.6952	-0.4563	1.3876	0.3609	-1.2391	1.9481	-0.3715
-0.6289	0.6420	1.1921	-0.4564	-1.7554	0.4721	-0.2147
-0.6215	-0.3100	1.4032	0.5022	-0.9087	1.1956	0.3328
-1.6193	-0.0890	1.1068	0.9947	-1.2511	2.1292	-0.3633
-0.6763	-0.1060	0.3007	-0.7081	-1.2344	1.6665	0.5152
-0.7855	-0.2062	1.6197	-0.1047	-2.1923	1.0576	-0.2072
-0.5527	0.0951	0.8251	0.1652	-1.4881	0.5919	-0.0448
0.1821	-0.4625	1.1065	-0.0788	-1.1373	1.6941	0.2177
-1.0216	0.3328	1.6178	0.7603	-0.6325	1.4086	-0.0444
-0.7464	-0.4383	1.0400	-0.8900	-0.6159	0.7964	-0.5719
-0.8502	-1.6189	-0.3538	0.8044	-0.0720	-0.1088	-0.8360
-0.7510	-0.9706	0.3582	1.0655	0.4216	-0.3015	-1.5726
-0.4739	-0.6319	0.6422	2.0299	-0.2603	-0.1102	-0.5091
-0.7330	-0.3950	-0.2020	0.5043	0.0663	-0.8902	-0.2179
-0.4330	-1.4785	0.0731	1.9445	0.0572	-0.3216	-1.3364
-0.2844	-0.8040	-0.5510	0.8029	0.4260	-1.1629	-1.4486
-0.6800	-0.4118	-0.5706	1.5380	0.4093	-0.0960	-0.9428
-0.6017	-1.1102	-0.3159	0.8827	1.2581	-0.9915	-0.3513
-0.4944	-1.5812	-0.8136	0.8441	1.5874	-0.0908	-0.8191
-0.4706	-1.4967	0.1321	0.6719	0.2776	-0.0711	-1.7956
-1.2173	-1.6905	-0.1295	0.9166	0.3423	-0.5926	-1.7682
0.0359	-0.8376	-0.7469	0.7333	0.8639	-0.6880	-1.1263
0.1118	-1.2214	-1.3014	1.6383	0.6306	-0.9105	-1.0789
0.4007	-1.4843	-1.1421	0.6611	0.8225	-0.7386	0.0907
-0.2838	-0.9521	-0.9017	0.7185	0.3125	-0.0816	-1.5057
0.3965	-1.1379	-0.5380	0.1953	1.0424	-0.7673	-1.0867
0.1059	0.0361	0.2630	1.2519	-0.0240	-0.7852	-1.4863
-0.6911	-0.6407	-0.2283	0.3998	0.7485	-0.0239	-1.0408
-2.2776	-1.1729	-0.2019	0.8760	0.5644	-0.1469	-0.6024
-0.1502	-0.5223	-1.1485	1.0422	0.7335	-1.5077	-1.1760
2.0272	1.4750	-0.9801	-1.9614	1.2820	-0.7261	1.3569
0.1989	0.8576	-1.0152	-0.6927	0.4431	-0.7846	1.7169
2.2220	1.0396	-0.5927	-0.4109	0.3165	-1.6929	0.9554
1.1895	1.6842	-1.1238	-0.8307	0.3544	-0.9759	1.5552
1.0515	1.6428	-0.7463	-1.5116	0.5674	-0.6314	1.3239
1.8664	1.0513	-0.7221	-0.8922	1.1087	-0.2142	1.2243
1.5677	1.5059	-0.3269	-1.3851	-0.1779	-0.3927	0.4779
1.3562	1.8998	-0.2993	-1.1442	0.3210	0.3170	1.7762
0.2630	0.5303	-0.6793	-2.1539	1.6170	-0.4378	0.7714
1.1919	1.2938	0.3356	-1.5879	1.1011	-0.3599	1.5303

La matrice des corrélations

Quand on considère
la matrice des données
centrées réduites, l'ACP revient à
la recherche des vecteurs propres
de la matrice des corrélations.

$$\bar{X} = ones(n,1) * mean(X)$$
$$S = ones(n,1) * std(X)$$
$$Xn = (X - \bar{X})./S$$
$$rho = \frac{1}{n}\, Xn' * Xn$$

1.						
0.63	1.					
-0.49	-0.17	1.				
-0.58	-0.77	0.21	1.			
0.45	0.10	-0.78	-0.19	1.		
-0.50	-0.17	0.81	0.16	-0.76	1.	
0.59	0.80	-0.14	-0.73	0.12	-0.05	1.

LA première composante principale

	AGE	SEX	BMI	BP	··· Serum Measurements ···						Response
Patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

Table 1. Diabetes study. 442 diabetes patients were measured on 10 baseline variables. prediction model was desired for the response variable, a measure of disease progression one yea after baseline.

$$X \Rightarrow \boxed{\text{A.C.P.}} \Rightarrow (u \in \mathbb{R}^n, v \in \mathbb{R}^p, \lambda \in \mathbb{R}^+)$$

Les 3 étapes de l'ACP

1. centrer et réduire les données — `Xn = (X - un*mX)./(un*sX)`
2. calculer la plus grande valeur propre — `[v,λ] = eig(Xn'*Xn)`
3. projeter les observations — `u = Xn*v`

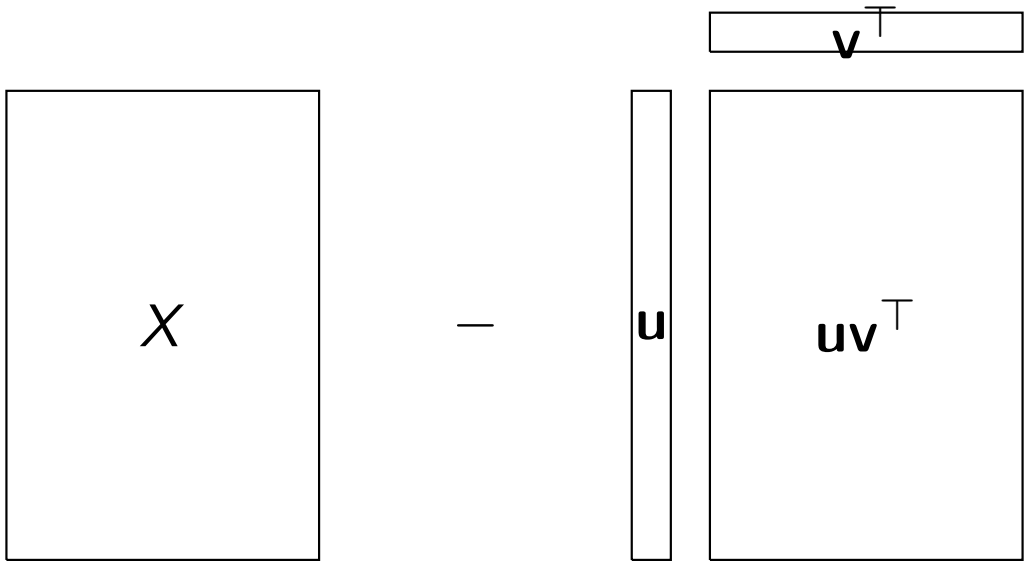
Comment interpréter $u = X_n v$? $\Rightarrow$ u est une nouvelle variable

Quelle est la dimension de uv^t ?

ACP : pourquoi rechercher des valeurs propres ?

La meilleur représentation linéaire du nuage de points est donnée par le couple de vecteurs $\mathbf{u} \in \mathbb{R}^n$ et $\mathbf{v} \in \mathbb{R}^p$ permettant au mieux de reconstruire la matrice X. Il sont alors solution du problème de minimisation suivant :

$$\min_{\mathbf{u},\mathbf{v}} J(\mathbf{u},\mathbf{v}) \qquad \text{avec} \qquad J(\mathbf{u},\mathbf{v}) = \sum_{i}^{n}\sum_{j}^{p}(x_{ij} - u_i v_j)^2$$



aussi noté : $\qquad J(\mathbf{u},\mathbf{v}) = \|X - \mathbf{u}\mathbf{v}^\top\|_F^2$

Résumé d'un tableau de données : résultat principal

Théorème (Eckart & Young, 1936)

La solution unique du problème d'optimisation suivant

$$\min_{\mathbf{u},\mathbf{v}} J(\mathbf{u},\mathbf{v}) \qquad \text{avec} \quad J(\mathbf{u},\mathbf{v}) = \|X - \mathbf{u}\mathbf{v}^\top\|_F^2$$

est donnée par

$$\mathbf{v}^\star \qquad \text{et} \qquad \mathbf{u}^\star = X\,\mathbf{v}^\star,$$

où $\mathbf{v}^\star$ est le vecteur propre normé associé à λ la plus grande valeur propre de la matrice $X^\top X$. De plus on a $\|\mathbf{v}^\star\| = 1$ et $\|\mathbf{u}^\star\| = \sqrt{\lambda}$.

Démonstration : A l'optimum on a

$$\begin{cases} \nabla_{\mathbf{u}} J(\mathbf{u},\mathbf{v}) = & -2X\mathbf{v} + 2\|\mathbf{v}\|^2\mathbf{u} = 0 \\ \nabla_{\mathbf{v}} J(\mathbf{u},\mathbf{v}) = -2X^\top\mathbf{u} + 2\|\mathbf{u}\|^2\mathbf{v} = 0 \end{cases}$$

Résumé d'un tableau de données : calculs

La fonction cout[1] peut se réécrire :

$$\begin{aligned}
\sum_{i}^{n}\sum_{j}^{p}(x_{ij}-u_i v_j)^2 &= \sum_{i}^{n}\sum_{j}^{p}x_{ij}^2 - 2\sum_{i}^{n}\sum_{j}^{p}x_{ij}u_i v_j + \sum_{i}^{n}\sum_{j}^{p}(u_i v_j)^2 \\
&= \sum_{i}^{n}\sum_{j}^{p}x_{ij}^2 - 2\sum_{i}^{n}\left(\sum_{j}^{p}x_{ij}v_j\right)u_i + \sum_{i}^{n}u_i^2\ \sum_{j}^{p}v_j^2 \\
&= \sum_{i}^{n}\sum_{j}^{p}x_{ij}^2 - 2(X\mathbf{v})^\top\mathbf{u} + \|\mathbf{u}\|^2\ \|\mathbf{v}\|^2
\end{aligned}$$

et donc

$$\min_{\mathbf{u},\mathbf{v}} \underbrace{\|X-\mathbf{u}\mathbf{v}^\top\|_F^2}_{J(\mathbf{u},\mathbf{v})} \quad \text{est analogue à} \quad \min_{\mathbf{u},\mathbf{v}} \underbrace{-2(X\mathbf{v})^\top\mathbf{u} + \|\mathbf{u}\|^2\ \|\mathbf{v}\|^2}_{\mathcal{J}(\mathbf{u},\mathbf{v})}$$

[1] qui est la norme de Frobenius de la matrice des différences $J(\mathbf{u},\mathbf{v}) = \|X-\mathbf{u}\mathbf{v}^\top\|_F^2$

Minimisation d'une fonction de plusieurs variables

$$\min_{\mathbf{u},\mathbf{v}} J(\mathbf{u},\mathbf{v}) \;=\; \|X - \mathbf{u}\mathbf{v}^\top\|_F^2$$

Definition (Gradient)

soit F une fonction de plusieurs (d) variables :

$$\begin{array}{rccc} F: & \mathbb{R}^d & \longmapsto & \mathbb{R} \\ & \mathbf{x} & \longrightarrow & F(\mathbf{x}) \end{array}$$

on appelle gradient de F au point $\mathbf{x}$ la fonction des dérivées partielles

$$\begin{array}{rccc} \nabla_{\mathbf{x}} F: & \mathbb{R}^d & \longmapsto & \mathbb{R}^d \\ & \mathbf{x} & \longrightarrow & \nabla_{\mathbf{x}} F(\mathbf{x}) = \left(\frac{\partial F}{\partial x_1}(\mathbf{x}), \ldots, \frac{\partial F}{\partial x_i}(\mathbf{x}), \ldots, \frac{\partial F}{\partial x_d}(\mathbf{x})\right)^\top \end{array}$$

Condition d'optimalité

$(u^\star, v^\star)$ est solution du problème de minimisation ssi le gradient de la fonction J s'annule en ce point

Minimisation d'une fonction de plusieurs variables

Exemple

$$\min_{x,y} J(x,y) = 2(x-a)^2 + (y-b)^2$$

Méthode de résolution du problème :

1. calcul du gradient $\nabla_{\mathbf{x}} F(x,y)$
2. résolution des équations $\nabla_{\mathbf{x}} F(x^\star, y^\star) = 0$ (2 équations à 2 inconnues)

Calcul du gradient

Minimiser le cout c'est trouver $\mathbf{u}$ et $\mathbf{v}$ qui annulent le gradient :

$$\min_{\mathbf{u},\mathbf{v}} J(\mathbf{u},\mathbf{v}) \;=\; \|X - \mathbf{u}\mathbf{v}^\top\|_F^2 \qquad \Leftrightarrow \qquad \begin{cases} \nabla_{\mathbf{u}} \mathcal{J}(\mathbf{u}) = 0 \\ \nabla_{\mathbf{v}} \mathcal{J}(\mathbf{v}) = 0 \end{cases}$$

pour le cout :

$$J(\mathbf{u},\mathbf{v}) \;=\; \sum_i^n \sum_j^p x_{ij}^2 - 2 \sum_i^n \left(\sum_j^p x_{ij} v_j \right) u_i + \sum_i^n u_i^2 \sum_j^p v_j^2$$

le gradient s'écrit

$$\begin{cases} \dfrac{\partial J(\mathbf{u},\mathbf{v})}{\partial u_i} &= -2 \displaystyle\sum_j^p x_{ij} v_j + 2 u_i \sum_j^p v_j^2 \\ \dfrac{\partial J(\mathbf{u},\mathbf{v})}{\partial v_j} &= -2 \displaystyle\sum_i^n x_{ij} u_i + 2 v_j \sum_i^n u_i^2 \end{cases}$$

Ecriture matricielle du gradient

$$\begin{cases} \nabla_{\mathbf{u}} J(\mathbf{u}) = -2X\mathbf{v} + 2\|\mathbf{v}\|^2 \mathbf{u} \\ \nabla_{\mathbf{v}} J(\mathbf{v}) = -2X^\top \mathbf{u} + 2\|\mathbf{u}\|^2 \mathbf{v} \end{cases}$$

$\min_{\mathbf{u},\mathbf{v}} \|X - \mathbf{u}\mathbf{v}^\top\|_F^2$

Les conditions d'optimalité s'écrivent :

$$\begin{cases} \nabla_{\mathbf{u}} \mathcal{J}(\mathbf{u}) = 0 & \Leftrightarrow \quad -X\mathbf{v} + \|\mathbf{v}\|^2 \mathbf{u} = 0 \\ \nabla_{\mathbf{v}} \mathcal{J}(\mathbf{v}) = 0 & \Leftrightarrow \quad -X^\top \mathbf{u} + \|\mathbf{u}\|^2 \mathbf{v} = 0 \end{cases}$$

On en déduit :

$$\begin{cases} X\mathbf{v} = \|\mathbf{v}\|^2 \mathbf{u} \\ X^\top \mathbf{u} = \|\mathbf{u}\|^2 \mathbf{v} \end{cases} \Rightarrow X^\top X \mathbf{v} = \|\mathbf{v}\|^2 X^\top \mathbf{u} = \underbrace{\|\mathbf{v}\|^2 \|\mathbf{u}\|^2}_{\lambda} \mathbf{v}$$

La solution est donc donnée par un vecteur propre $\mathbf{v}$ de la matrice $X^\top X$.

lequel ?

Quel vecteur propre choisir ?

A l'otimum

$$\mathcal{J}(\mathbf{u},\mathbf{v}) = -2(X\mathbf{v})^\top\mathbf{u} + \|\mathbf{u}\|^2\,\|\mathbf{v}\|^2 \qquad \text{et} \qquad X\mathbf{v} = \|\mathbf{v}\|^2\mathbf{u}$$

soit

$$\begin{aligned}\mathcal{J}(\mathbf{u},\mathbf{v}) &= -2\|\mathbf{v}\|^2\mathbf{u}^\top\mathbf{u} + \|\mathbf{u}\|^2\,\|\mathbf{v}\|^2 \\ &= -2\|\mathbf{v}\|^2\|\mathbf{u}\|^2 + \|\mathbf{u}\|^2\,\|\mathbf{v}\|^2 \\ &= -\|\mathbf{u}\|^2\,\|\mathbf{v}\|^2 \qquad\qquad = -\lambda\end{aligned}$$

le cout c'est la valeur propre

La solution du problème est donc donnée par le vecteur propre associé à la plus grande valeur propre de la matrice $X^\top X$ car :

$$\nabla_J(\mathbf{v}) = 0 \quad \Leftrightarrow \quad X^\top X\mathbf{v} - \lambda\mathbf{v} = 0 \quad \text{avec} \quad J(\mathbf{u},\mathbf{v}) = \|X\|^2 - \lambda$$

et l'on sait que toutes les valeurs propres de la matrice symétrique définie positive $X^\top X$ sont positives.

et la suite.. On itere le processus

Construisons la matrice des résidus $R = X - \mathbf{u}\mathbf{v}^\top$

$R^\top R$ admet les mêmes valeurs propres que $X^\top X$.

...à l'exception de la plus grande qui devient zéro.

en posant $X^\top X\mathbf{v} = \lambda\mathbf{v}$ et $X^\top X\mathbf{v}_2 = \lambda_2\mathbf{v}_2$ et $\lambda_2 < \lambda$

$$\begin{aligned} \text{on a :} \quad R^\top R\mathbf{v} &= (X - \mathbf{u}\mathbf{v}^\top)^\top(X - \mathbf{u}\mathbf{v}^\top)\mathbf{v} \\ &= X^\top X\mathbf{v} - 2\mathbf{v}\mathbf{u}^\top X\mathbf{v} + \mathbf{v}\mathbf{u}^\top\mathbf{u}\mathbf{v}^\top\mathbf{v} \\ &= \lambda\mathbf{v} - 2\|\mathbf{u}\|^2\mathbf{v} + \|\mathbf{u}\|^2\|\mathbf{v}\|^2\mathbf{v} \\ &= \lambda\mathbf{v} - \|\mathbf{u}\|^2\mathbf{v} \qquad = 0 \end{aligned}$$

car $\mathbf{u} = X\mathbf{v}$, $\|\mathbf{v}\|^2 = 1$ et $\|\mathbf{u}\|^2 = \|X\mathbf{v}\|^2 = \mathbf{v}^\top X^\top X\mathbf{v} = \lambda$

et

$$\begin{aligned} R^\top R\mathbf{v}_2 &= (X - \mathbf{u}\mathbf{v}^\top)^\top(X - \mathbf{u}\mathbf{v}^\top)\mathbf{v}_2 \\ &= X^\top X\mathbf{v}_2 - 2X^\top\mathbf{u}\mathbf{v}^\top\mathbf{v}_2 + \mathbf{v}\mathbf{u}^\top\mathbf{u}\mathbf{v}^\top\mathbf{v}_2 \\ &= \lambda_2\mathbf{v}_2 \end{aligned}$$

car $\mathbf{v}^\top\mathbf{v}_2 = 0$ puisque les vecteurs propres sont orthogonaux entre eux.

Les valeurs propres et vecteurs propres en 5 lignes

1. soit X la matrice des observations
2. Xr est la matrice des données centrées réduites $Xr(:,i) = \frac{X(:,i)-\mu_i}{\sigma_i}$
3. on calcule les valeurs et vecteurs propres de la matrice de covariance $Xr^\top Xr$
4. on visualise les observations en les projetant sur les vecteurs propres associés aux plus grandes valeurs propres (typiquement les 2 plus grandes)
5. on visualise aussi les variables

L'analyse en composantes principales (ACP)

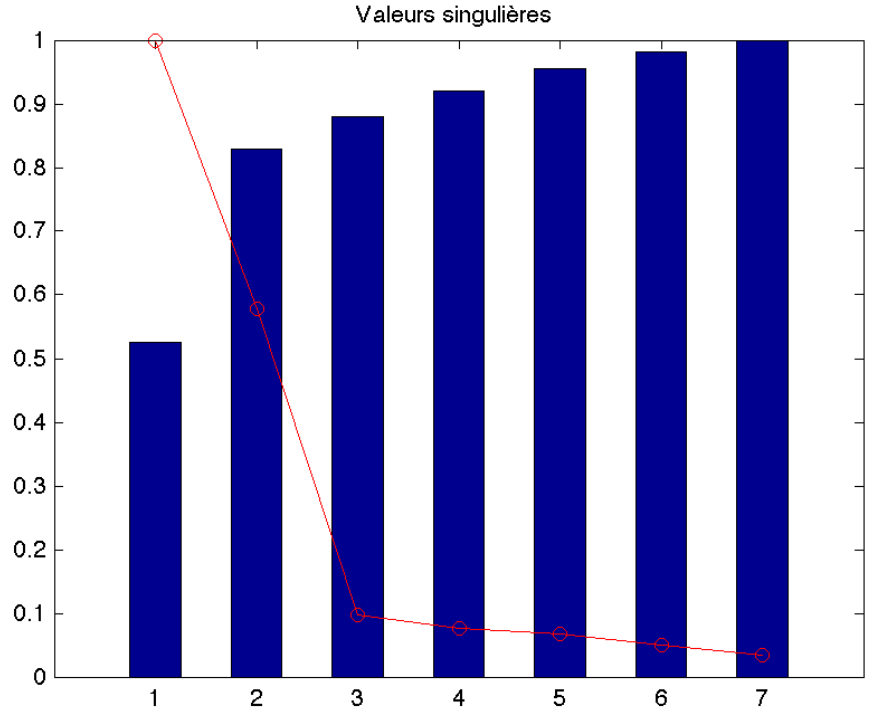
Donc en résumé, si l'on cherche le sous espace V de rang k qui « ressemble » le plus aux données X il faut calculer les k vecteurs proposes correspondant aux k plus grandes valeurs propres.

```
[n,p]=size(X);
Xn = (X - ones(n,1)*mean(X))./(ones(n,1)*std(X));
[v,d] = eig(Xn'*Xn);             % Calcul des valeurs propres

vp = sort(diag(d),'descend');    % on ordonne les valeurs propres
[vp 100*cumsum(vp)/sum(vp)],     % affichage des valeurs propres

  216.9681    52.53
  125.3473    82.89
   21.1190    88.00
   16.5800    92.01
   14.5963    95.55
   10.8213    98.17
    7.5682   100.00

v =
    0.06    0.14   -0.41   -0.16   -0.75    0.05    0.44
    0.53   -0.38    0.44    0.24   -0.07    0.38    0.38
   -0.28   -0.68   -0.33   -0.01   -0.15    0.42   -0.36
    0.13   -0.19    0.49   -0.48   -0.44   -0.34   -0.38
    0.24   -0.46   -0.33   -0.40    0.38   -0.42    0.34
    0.57    0.31   -0.24   -0.40    0.16    0.44   -0.35
   -0.47    0.07    0.31   -0.58    0.17    0.41    0.36
```

L'analyse en composantes principales (ACP)

Donc en résumé, si l'on cherche le sous espace V de rang k qui « ressemble » le plus aux données X il faut calculer les k vecteurs proposes correspondant aux k plus grandes valeurs propres.

```
[n,p]=size(X);
Xn = (X - ones(n,1)*mean(X))./(ones(n,1)*std(X));
[v,d] = eig(Xn'*Xn);              % Calcul des valeurs pro

vp = sort(diag(d),'descend');     % on ordonne les valeurs
[vp 100*cumsum(vp)/sum(vp)],      % affichage des valeurs

  216.9681    52.53
  125.3473    82.89
   21.1190    88.00
   16.5800    92.01
   14.5963    95.55
   10.8213    98.17
    7.5682   100.00

v =
    0.06    0.14   -0.41   -0.16   -0.75    0.05    0.44
    0.53   -0.38    0.44    0.24   -0.07    0.38    0.38
   -0.28   -0.68   -0.33   -0.01   -0.15    0.42   -0.36
    0.13   -0.19    0.49   -0.48   -0.44   -0.34   -0.38
    0.24   -0.46   -0.33   -0.40    0.38   -0.42    0.34
    0.57    0.31   -0.24   -0.40    0.16    0.44   -0.35
   -0.47    0.07    0.31   -0.58    0.17    0.41    0.36
```



La représentation des individus et des variables par l'ACP

Facteurs : les axes factoriels $\mathbf{v}, \lambda$

$$X_n^\top X_n \mathbf{v} = \lambda \mathbf{v} \qquad \text{et } \|\mathbf{v}\| = 1$$

Individus : les composante principale qui sont les projection des individus

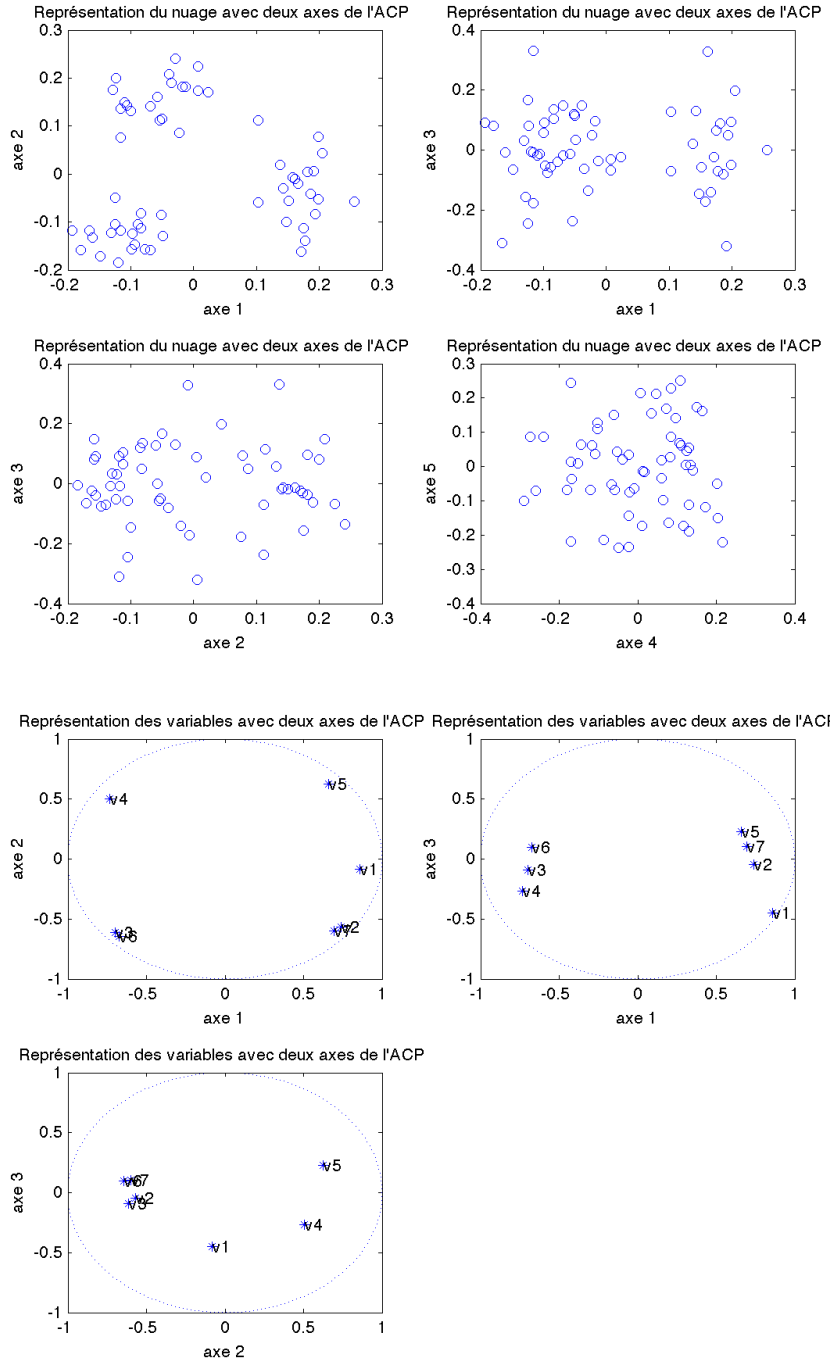
$$\mathbf{u} = X_n \mathbf{v}$$

Variables : calcul des corrélations entre variables observées normalisées X_n (centrées réduites) et les composantes principales $\mathbf{u}$.

$$\text{cor}(X_n, \mathbf{u}) = \frac{X_n^\top \mathbf{u}}{\sqrt{n}\,\|\mathbf{u}\|} = \frac{X_n^\top X_n \mathbf{v}}{\sqrt{n}\,\|\mathbf{u}\|} = \frac{\lambda}{\sqrt{n}\sqrt{\lambda}}\mathbf{v} = \frac{\sqrt{\lambda}}{\sqrt{n}}\mathbf{v}$$

car pour toutes les variables $\|X_{\bullet j}\|^2 = n$ et $\|\mathbf{u}\|^2 = \lambda$

```
[V,D] = eig(Xn'*Xn);
U = Xn*V;

% les nouvelles variables

plot(U(:,1),U(:,2),'o');
plot(U(:,1),U(:,3),'o');
plot(U(:,2),U(:,3),'o');
plot(U(:,4),U(:,5),'o');

% le role des variables

Vn = (V*sqrt(D(1:p,1:p))/sqrt(n));

plot(Vn(:,1),Vn(:,2),'*');
plot(Vn(:,1),Vn(:,3),'*');
plot(Vn(:,2),Vn(:,3),'*');
```



L'ACP en matlab

On centre et on réduit les données .

Fonction $V_n, U, \lambda \leftarrow \texttt{ACP}(X, k)$

$X = (X - un * \texttt{mean}(X))./(un * \texttt{std}(X))$
$(V, \lambda) = \texttt{eig(X'*X,k)}$ ou $(U, V, \mu) = \texttt{svd(X,k)}$
$U = X * V$
$V_n = V * \sqrt{\lambda}/\sqrt{n}$ ou $V_n = V * \mu/\sqrt{n}$

on projette les données sur le sous espace engendré par les k vecteurs propres associés aux k plus grandes valeurs propres
et on calcule les corrélations entre les observation et ces nouvelles variables.

L'ACP reconstruit la matrice des données

Meilleur approximation de rang k

Soit X une matrice. La meilleure approximation de rang k de X, la matrice A_k minimisant le critère suivant :

$$\min_{A_k} \|X - A_k\|_F^2 \qquad \text{avec} \qquad \text{rang}(A_k) = k$$

s'obtient à l'aide des k vecteurs propres $\mathbf{v}_i,\ i = 1, k$ associées aux k plus grandes valeurs propres de la matrice $X^\top X$ de la manière suivante :

$$A_k = U_k V_k^\top$$

où V_k est la matrice des vecteurs propres $\mathbf{v}_i,\ i = 1, k$ et U_k est la matrice des vecteurs $\mathbf{u}_i = X\mathbf{v}_i,\ i = 1, k$ et donc $A_k = XV_k V_k^\top$

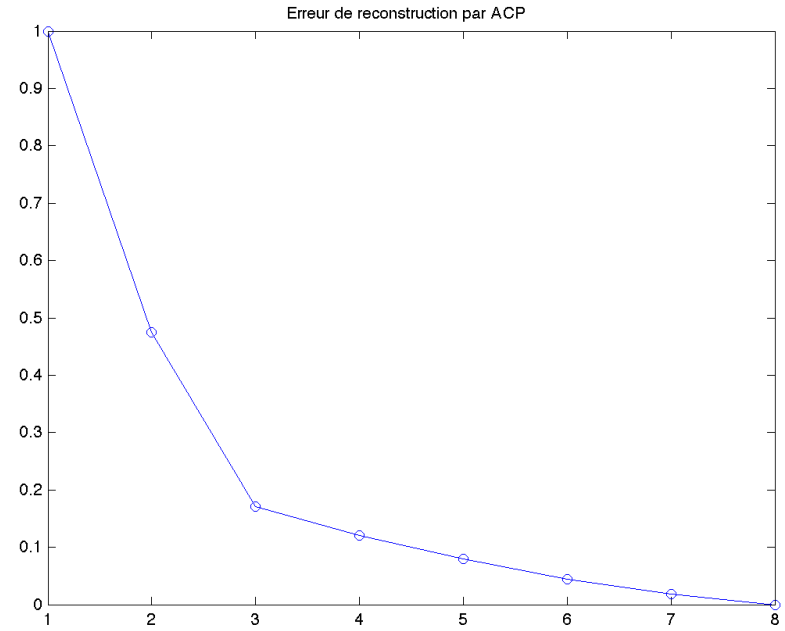
l'erreur d'approximation est donnée par la somme des valeurs propres restantes :

$$\|X - A_k\|_F^2 = \sum_{i=k+1}^{n} \lambda_i$$

L'ACP reconstruit la matrice des données

$$J_k = \|X - U_k V_k^\top\|_F^2 \qquad k = 1, p$$

```
J(1) = (sum(sum(abs(Xn).^2)));

for k=1:p
    J(k+1) = norm(Xn - (U(:,1:k)*D(1:k,1:k)*(V(:,1:k)'))
end

figure(5)
plot(J/J(1));
hold on
title('Erreur de reconstruction par ACP')
plot(J/J(1),'o');
hold off
```



Avec les deux plus grandes valeur propres, on reconstruit plus de 80 % de la matrice des données (ébouli des valeurs propres selon Baccini et Besse)

observations $X =$ Information $A_k +$ Bruit

Plus loin sur le même principe...

- D'autres applications possible de l'ACP :
  - compression d'image
- D'autres analyses à base d'analyse spectrale
  - sur des tableaux de distance
  - Tables de contingences : l'analyse factorielle des correspondances (AFC)
  - L'analyse canonique (AC) : comparaison de deux groupes de variables
  - L'analyse des correspondances multiples, l'analyse discriminates
- D'autres développements :
  - ACP et modèle probabiliste
  - ACP non linéaire,
  - ACP généralisée
  - ACP parcimonieuse

Conclusion

- Pourquoi faire une analyse en composantes principale (ACP)
  - permet de représenter les individus
  - permet de représenter les variables
  - permet de compresser les données et d'éliminer du bruit

- Comment faire une L'analyse en composantes principale (ACP)
  - réduire et centrer les données Xr
  - calculer la matrice des correlations $Xr^\top Xr$
  - en extraire les valeurs et les vecteurs propres V
  - reconstruire les nouvelles variables $U = XrV$
  - choisir k le nombre de facteurs significatifs k

- Méthode exploratrice
  - complexe dans sa formulation
  - facile dans sa mise en œuvre (5 lignes de code)

Repéres bibliographiques

- Statistique descriptive multidimensionnelle : Techniques factorielles de base Baccini, Alain et Besse, Philippe (UPS. Université Paul Sabatier, Toulouse 3. LSP. Laboratoire de statistique et probabilités. France)

  `http://www.math.univ-toulouse.fr/~besse/pub/sdm2.pdf`
- le cours de l'UTC sur l'ACP (G. Govaert)

  http://www.hds.utc.fr/sy09/doku.php?id=fr:transparents
- un tutorial de CMU sur l'ACP

  `http://www.cs.cmu.edu/~elaw/papers/pca.pdf`
- sur wikipedia

  `http://en.wikipedia.org/wiki/Principal_component_analysis`

Rappel sur les valeurs propres

soit Σ une matrice carrée symétrique définie positive ($\Sigma = X^\top X$) de dimension p. Il existe :

- p réels positifs $\lambda_i, \quad i = 1, p$ et
- p vecteurs de $\mathbb{R}^p$, $\mathbf{v}_i, \quad i = 1, p$

tels que

$$\Sigma \mathbf{v}_i = \lambda_i \mathbf{v}_i$$

quelques faits à propos des valeurs et des vecteurs propres

- par convention on ordonne les valeurs propres $\lambda_i \leq \lambda_{i-1}, \ i = 2, p$
- par convention les vecteurs propres sont normés $\mathbf{v}_i^\top \mathbf{v}_i = 1$
- les vecteurs propres sont orthogonaux entre eux : $\mathbf{v}_i^\top \mathbf{v}_j = 0 \ \ i \neq j$
- la matrice des vecteurs propres V forme une base de $\mathbb{R}^p$
- décomposition spectrale : $\Sigma = \sum_{i=1}^{p} \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$
- on les calcule en utilisant un programme (`[V,D] = eig(Σ)`)

```
X = randn(10,5)

X =

   -0.6416   -0.3934   -0.7313    0.9820    0.5524
    0.8415    2.1730    0.3442    1.6411    0.7336
    0.6521    0.5006    3.4307   -0.5842   -1.6001
    0.1848   -0.2535    0.3812   -0.8544   -0.3328
    0.3635    0.5542    1.9298    0.6564    0.6291
    0.5890   -0.2903    1.5685   -1.6603   -0.5648
    0.5153   -0.4606    0.2263    1.6934   -1.0449
    0.9598    1.2378   -2.0753    0.5395   -0.4543
    1.9332    2.2013   -0.2789    0.2030   -1.4092
   -0.0956   -2.0578   -0.9055   -2.3387    1.0096

S = X'*X

S =

    6.9914    7.7941    2.3642    1.4789   -4.7412
    7.7941   16.4077    1.8471    9.0969   -4.0880
    2.3642    1.8471   24.0085   -2.4957   -5.2547
    1.4789    9.0969   -2.4957   16.5856   -0.3456
   -4.7412   -4.0880   -5.2547   -0.3456    8.5323

[V,D] = eig(S)

V =

   -0.8033    0.0653    0.4722    0.0324   -0.3555
    0.4619    0.4924    0.2197    0.3078   -0.6334
   -0.0481    0.1162   -0.3922   -0.7853   -0.4623
   -0.2109   -0.3453   -0.6437    0.5158   -0.3948
   -0.3074    0.7877   -0.4008    0.1468    0.3206

D =

    1.2257         0         0         0         0
         0    4.9599         0         0         0
         0         0   10.6614         0         0
         0         0         0   25.8086         0
         0         0         0         0   29.8698

diag(D)

D =

    1.2257
    4.9599
   10.6614
   25.8086
   29.8698
```

$$S\mathbf{v} = \lambda\mathbf{v}$$

```
S*V(:,1) - D(1,1)*V(:,1)
                    =  1.0e-13  *
                       0
                      -0.7105
                       0.5684
                       0.5684
                       0

S - V*D*V';
                    =  -5.6843e-14  *
                          ...
```