Analyse en composantes principales

ACP

AUTEUR·RICE

Benjamin Charlier, Adrien Mazoyer

1 Objectifs

1.1 Description des données

On travaille sur des tableaux de données où

- les **lignes** correspondent aux **individus**
- les **colonnes** correspondent aux **variables quantitatives** étudiées (i.e. des nombres)

	Variable 1	$\cdots$	**Variable p**
Individu 1	$x_{1,1}$	$\cdots$	$x_{1,p}$
$\vdots$	$\vdots$		$\vdots$
Individu n	$x_{n,1}$	$\cdots$	$x_{n,p}$

$x_{i,j}$: mesure de la variable j pour l'individu i.

▸ **Iris (R)**

```
> head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

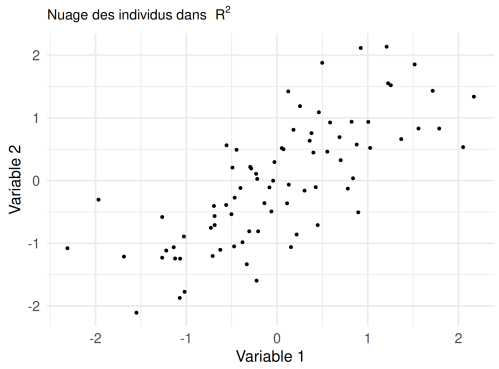
Objectif : trouver une représentation, dans un espace de dimension réduite, permettant de mettre en évidence d'éventuelles structures au sein des données.

1.2 Représentation graphique

On est rapidement limité

- En dimension 1 ou 2 on peut les représenter graphiquement
- Si 3 voir 4 c'est plus compliqué, mais on on peut bricoler
- Si $p > 4$ ca devient franchement compliqué

▸ **Nuage 2D**



1.3 Le tableau de données

Le tableau (individus, variables) est représenté sous la forme d'une matrice X de taille $n \times p$:

$$X = \begin{pmatrix} X_{:,1} & \cdots & X_{:,p} \end{pmatrix} = \begin{pmatrix} X_{1,:} \\ \vdots \\ X_{n,:} \end{pmatrix} = \begin{pmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \cdots & x_{n,p} \end{pmatrix}.$$

Note

- Colonnes **actives** : variables utilisées pour construire l'ACP (par défaut).
- Dans le cas contraire, on parle de variables **supplémentaires**, utiles pour l'**interprétation** des axes.

▸ **Notation**

- La valeur $x_{i,j}$ est la mesure de la variable j pour l'individu i.
- La ligne i, notée $X_{i,:} \in \mathbb{R}^p$, représente l'individu i.
- La colonne j, notée $X_{:,j} \in \mathbb{R}^n$, contient les valeurs de la variable j.

1.4 Dualité

- Espace E des individus (direct) : Une matrice $n \times p$ code un nuage de n points dans $\mathbb{R}^p$. À chaque individu i on associe le vecteur colonne $X_{i,:} \in \mathbb{R}^p$ ("les valeurs d'une ligne mis en colonne"). Ce sont des éléments d'un espace vectoriel E de dimension p : on peut prendre $\mathbb{R}^p$ muni de la base canonique $\mathcal{E}$ et d'une métrique de matrice M lui conférant une structure d'espace euclidien.

- Espace F des variables (dual) : de la même manière, cette même matrice $n \times p$ code un nuage de p points dans $\mathbb{R}^n$... À la j-ème variable, on associe le vecteur colonne $X_{:,j} \in \mathbb{R}^n$. C'est un élément d'un espace vectoriel noté F de dimension n : prendre $\mathbb{R}^n$ muni de la base canonique $\mathcal{F}$ et d'une métrique de matrice diagonale $D = \mathrm{Diag}(w_1, \cdots, w_n)$ lui conférant une structure d'espace euclidien.

En pratique: le plus souvent on a $M = Id_p$ et $D = Id_n$ (i.e $w_1 = \cdots = w_n = 1$). Mais cette formulation plus générale permet d'inclure nombre de cas particuliers dans le formalisme de l'ACP.

1.5 Objectifs

L'**analyse en composantes principales (ACP)** est une méthode de statistique exploratoire permettant de décrire un grand tableau de données quantitatives de type individus variables.

Question : Lorsqu'on étudie simultanément un nombre important de variables, comment en faire une représentation globale ? On recherche de plus à :

- Individus : identifier des profils similaires ou différents, étudier leur variabilité et les représenter graphiquement dans un espace de faible dimension.
- Variables: étudier leur structure de dépendance (corrélations), identifier d'éventuels groupes de variables liées et de les représenter graphiquement.

Idée : L'ACP est un simple changement de base : passer d'une représentation dans la base canonique des variables initiales à une représentation dans la base des ***facteurs***. Les facteurs sont des combinaisons linéaires des variables (vecteurs propres de la matrice des corrélations).

Interprétation : On ne sélectionne qu'une petite partie des facteurs et cela permet de réaliser des graphiques dans cet espace de faible dimension (dimension égale au nombre de facteurs retenus). Les graphiques peuvent éventuellement permettre de comprendre la structure des données analysées. Cette interprétation est guidée par des indicateurs numériques.

Référence: en ligne (**Bloem 2024**), (**Besse, s. d.**). Des livres (**Escofier et Pagès 2016**), (**Saporta 2006**).

Bloem, Peter. 2024. ***Unraveling Principal Component Analysis***. 1.1.0. **https://peterbloem.nl/publications/unraveling-pca**.

Besse, Philippe. s. d. « Analyse en Composantes Principales (ACP) ». **https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-explo-acp.pdf**.

Escofier, Brigitte, et Jérôme Pagès. 2016. ***Analyses factorielles simples et multiples : cours et études de cas***. Dunod.

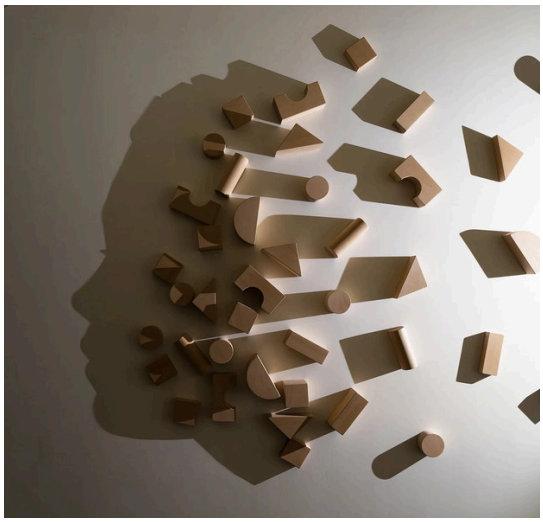
Saporta, Gilbert. 2006. ***Probabilités, analyse des données et statistique / Gilbert Saporta,...*** 2e édition révisée et augmentée. Éditions Technip.

1.6 Les méthodes linéaires

- On observe les nuages de points suivant différents points de vue (**projection**)
- Certains points de vue sont meilleurs que d'autres : ce sont ceux où la forme du nuage est la moins altérée.



Kumi Yamashita



https://kumiyamashita.com/light-shadow

Objectif de l'ACP

Trouver le plan de projection permettant de visualiser une approximation en 2 dimensions du nuage des individus qui vit dans $\mathbb{R}^p$.

2 Principe de l'ACP

2.1 Centrage et réduction

Dans ce cours, (et par défaut dans de nombreuses implémentations de l'ACP), le nuage de points des individus est centré et réduit (i.e. « on centre et on réduit colonne par colonne »).

- **Centrage :** on note $\bar{X} = (\bar{X}_{:,j})_{j=1}^{p} = \left(\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\right)_{j=1}^{p}$ le vecteur ligne des moyennes pour les p variables. La matrice des données centrées est

$$Y = X - \mathbf{1}_n \bar{X}, \qquad \text{où } \mathbf{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \in \mathbb{R}^n.$$

- **Réduction :** pour éviter les problèmes d'échelle, on réduit chaque colonne de X à variance 1[1]. On note $s^2_{X_{:,j}} = \frac{1}{n}\sum_{i=1}^{n}\left(X_{i,j} - \bar{X}_{:,j}\right)^2$ la variance[2] de la colonne j et on pose

[1] cela ne détruit pas la structure de corrélation du nuage de points

[2] Nous utiliserons l'écart-type non corrigé.

$$M_{1/s} = \mathrm{Diag}\left(\frac{1}{s_{X_{:,1}}}, \dots, \frac{1}{s_{X_{:,p}}}\right) \in \mathbb{R}^{p\times p}.$$

La matrice des données centrée-réduite est

$$Z = Y M_{1/s}$$

2.2 Pré-traitement

- Travailler avec la métrique $M = M_{\frac{1}{s^2}}$ (dans l'espace des individus) sur le tableau centré Y revient donc à travailler avec la métrique identité $M = I_p$ sur le tableau centré réduit.

- L'ACP « usuelle » revient à centrer et réduire les observations, puis à utiliser la métrique identité : c'est l'ACP normée.

2.3 Inertie d'un nuage de point

- Le **centre de gravité** G des points $A_i = (a_{i,1}, \dots, a_{i,p}) \in \mathbb{R}^p$ est le point de coordonnées

$$G = (\bar{a}_{:,1}, \dots, \bar{a}_{:,p}) = \left(\frac{1}{n}\sum_{i=1}^{n} a_{i,1}, \dots, \frac{1}{n}\sum_{i=1}^{n} a_{i,p}\right) \in \mathbb{R}^p.$$

- La **matrice d'inertie** est définie par

$$S = \frac{1}{n}\begin{pmatrix} A_1 - G \\ \vdots \\ A_n - G \end{pmatrix}^{\top}\begin{pmatrix} A_1 - G \\ \vdots \\ A_n - G \end{pmatrix} \in \mathbb{R}^{p\times p}.$$

Pour les données centrées, le centre de gravité du nuage des individus est l'origine.

- L'**inertie** du nuage autour de son centre de gravité est la moyenne des distances **au carré** des points A_i à G :

$$\text{Inertie} = \frac{1}{n}\sum_{i=1}^{n}\|A_i - G\|^2 = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}(a_{i,j} - g_j)^2 = \text{tr}(S).$$

2.4 Lien statistique et géométrie

Les n valeurs relevées pour chaque variable peuvent être vues comme n réalisations indépendantes d'une v.a.

- Pour des variables centrées Y et $D = Id_n$, on a $\langle Y_{:,j}, Y_{:,k}\rangle_F = \sum_{i=1}^{n} Y_{i,j}Y_{i,k}$ et à un facteur $\frac{1}{n}$ près

$$\frac{1}{n}\|Y_{:,j}\| \propto \text{écart type} \quad \text{et} \quad \langle Y_{:,j}, Y_{:,k}\rangle_F \propto \text{covariance}$$

- La **matrice de variance :** c'est la matrice $p \times p$ notée V dont l'entrée j,k contient $\text{cov}(X_{:,j}, X_{:,k}) = \frac{1}{n}\sum_{i=1}^{n}(X_{i,j} - \bar{X}_{:,j})(X_{i,k} - \bar{X}_{:,k})$. On a, pour les données centrée Y,

$$V = \frac{1}{n}Y^\top Y$$

  C'est donc aussi la matrice d'inertie de Y

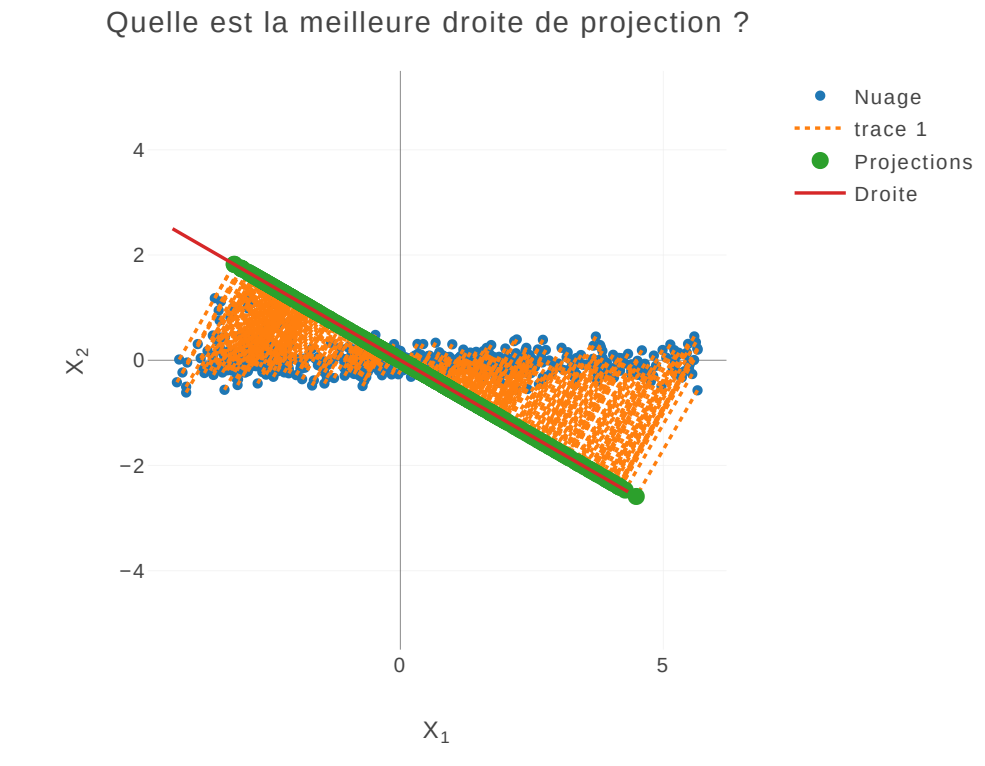
- La **matrice de corrélation** $S \in \mathbb{R}^{p\times p}$ contient toutes les corrélations:

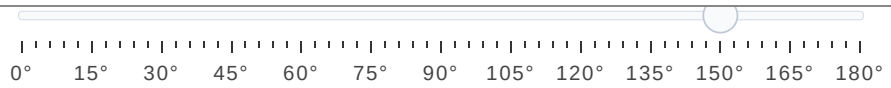
$$S = (\text{cor}(X_{:,j}, X_{:,k}))_{j,k=1^p} = \left(\frac{\text{cov}(X_{:,j}, X_{:,k})}{s_{X_{:,j}} s_{X_{:,k}}}\right)_{j,k=1}^{p}.$$

La matrice d'inertie (et de variance) des données centrées-réduites Z est la matrice de corrélation des données initiales X, et son inertie totale est égale à p (dimension de l'espace des individus).

2.5 Droite d'inertie maximale

On part d'un nuage qui vit dans $\mathbb{R}^2$, et on veut une image approchée en dimension 1.





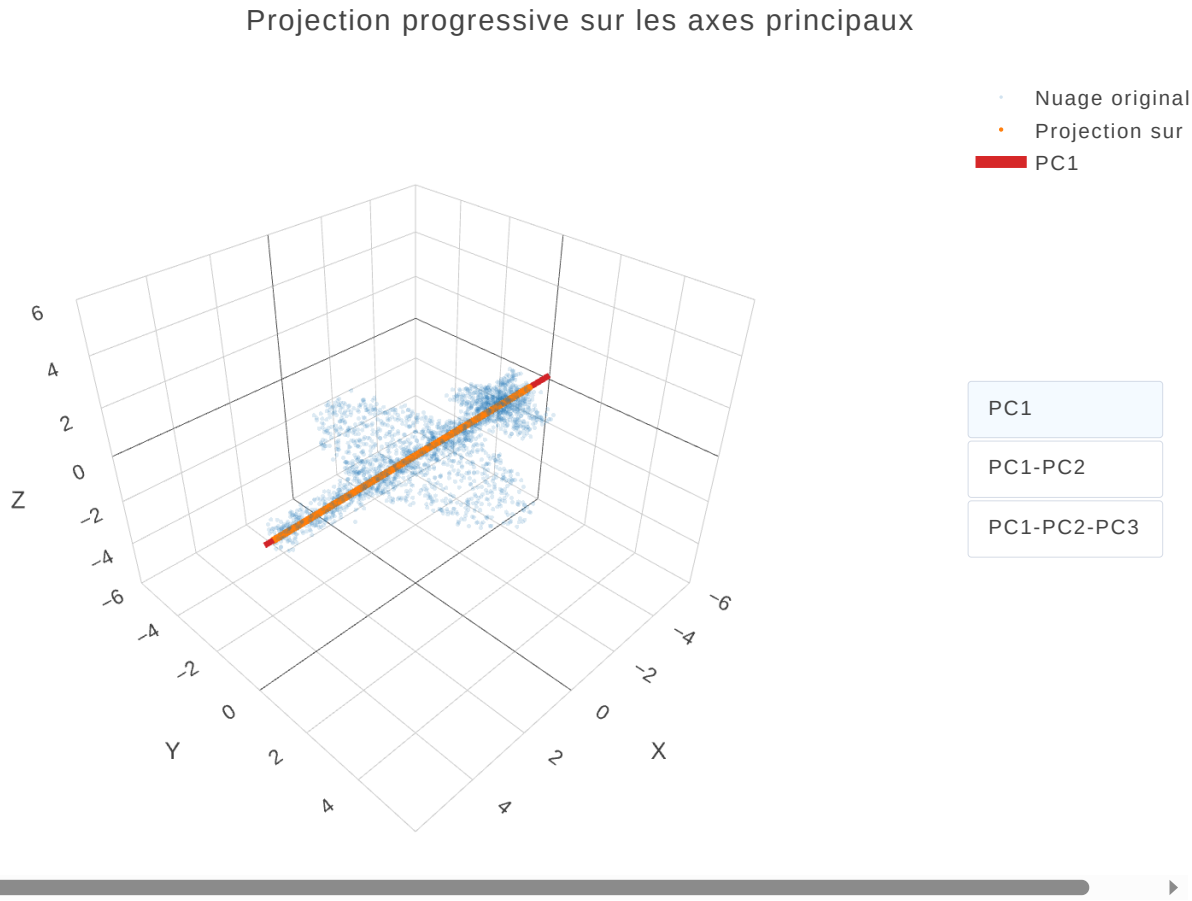
2.6 Exemple jouet

- La “meilleure” droite de projection est celle pour laquelle la structure du nuage est la mieux conservée après projection.
- Nous allons caractériser cette structure à l’aide de l’inertie du nuage de points.
- On cherche donc la droite de projection qui modifie le moins l’inertie totale du nuage de point après projection.
- Plus formellement, on cherche une droite (D), passant par l’origine O, telle que

$$\frac{1}{n}\sum_{i=1}^{n}\|OH_i\|^2$$

soit maximale, où H_i est le projeté orthogonal de l’individu $X_{i,:}$ sur (D).

2.7 Sous espaces d’inertie maximale



- La premiere direction principale est celle sur laquelle l’inertie des projections est maximale.
- Les directions suivantes maximisent la même quantité sous la contrainte d’être orthogonales aux directions déjà retenues.

2.8 Principe de l’ACP

- **Problème :** On considère un nuage de points dans $\mathbb{R}^p$, avec $p > 2$, et on cherche une bonne représentation de ce nuage dans un espace vectoriel de dimension $q \ll p$ (typiquement $q = 2$ ou 3). On cherche donc à identifier q directions orthogonales

$$u_1, \dots, u_q,$$

qui définissent un sous-espace de dimension q sur lequel l'inertie des projections est maximale.

- **Solution :** sous espaces emboîtés

1. u_1 est la direction d'inertie maximale ;
2. u_2 est la direction d'inertie maximale parmi les directions orthogonales à u_1 ;
3. u_3 est la direction d'inertie maximale parmi les directions orthogonales à u_1 et u_2 ;
4. $\dots$
5. u_q est la direction d'inertie maximale parmi les directions orthogonales à $u_1, \dots, u_{q-1}$.

L'hyperplan recherché est donc le sous-espace engendré par les vecteurs

$$\mathrm{Vect}(u_1, \dots, u_q).$$

Les vecteurs $u_1, \dots, u_q$ sont appelés les **directions principales** (ou **axes principaux**).

2.9 Matrice de covariance et axes principaux

On considère que les données ont été centrées. La **matrice de covariance** des données est alors

$$S = \frac{1}{n} X^\top X \in \mathbb{R}^{p \times p},$$

Il s'agit également de la matrice d'inertie du nuage des individus.

- La matrice S est **symétrique**, donc diagonalisable dans une base orthonormée[3] :

[3] Théorème spectral, https://www.youtube.com/watch?v=mhy-ZKSARxI

$$S = P \Lambda P^\top,$$

où P est une matrice **orthogonale** $P^\top P = P P^\top = I$ et $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_p)$ contient les valeurs propres de S.

- Comme S est **symétrique semi-définie positive**, ses valeurs propres vérifient

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0.$$

- Les colonnes de P sont les **vecteurs propres** de S et correspondent aux **directions principales**.

2.10 L'Analyse en Composantes Principales

Réaliser une Analyse en Composantes Principales s'effectue en **3 étapes** :

1. **Calculer et diagonaliser** la matrice de covariance (ou de corrélation) S.

2. **Ranger les valeurs propres** par ordre décroissant :

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0.$$

3. Pour chaque valeur propre λ_k, **identifier un vecteur propre** u_k associé.

Les vecteurs propres $u_1, \dots, u_p$ correspondent aux **directions principales** : u_k est la direction d'inertie maximale sous la contrainte d'être orthogonale à $u_1, \dots, u_{k-1}$.

Note

La valeur propre λ_k mesure l'**inertie portée par la direction principale** u_k.

2.11 Exemple classique

Moyennes des températures mensuelles sur la période 1991-2020 dans 14 villes françaises

Températures mensuelles et coordonnées géographiques

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce	Lati	Long
Bordeaux	5.6	6.6	10.3	12.8	15.8	19.3	20.9	21.0	18.6	13.8	9.1	6.2	44.50	-0.34
Brest	6.1	5.8	7.8	9.2	11.6	14.4	15.6	16.0	14.7	12.0	9.0	7.0	48.24	-4.29
Clermont	2.6	3.7	7.5	10.3	13.8	17.3	19.4	19.1	16.2	11.2	6.6	3.6	45.47	3.05
Grenoble	1.5	3.2	7.7	10.6	14.5	17.8	20.1	19.5	16.7	11.4	6.5	2.3	45.10	5.43
Lille	2.4	2.9	6.0	8.9	12.4	15.3	17.1	17.1	14.7	10.4	6.1	3.5	50.38	3.04
Lyon	2.1	3.3	7.7	10.9	14.9	18.5	20.7	20.1	16.9	11.4	6.7	3.1	45.45	4.51
Marseille	5.5	6.6	10.0	13.0	16.8	20.8	23.3	22.8	19.9	15.0	10.2	6.9	43.18	5.24
Montpellier	5.6	6.7	9.9	12.8	16.2	20.1	22.7	22.3	19.3	14.6	10.0	6.5	43.36	3.53
Nantes	5.0	5.3	8.4	10.8	13.9	17.2	18.8	18.6	16.4	12.2	8.2	5.5	47.13	-1.33
Nice	7.5	8.5	10.8	13.3	16.7	20.1	22.7	22.5	20.3	16.0	11.5	8.2	43.42	7.15
Paris	3.4	4.1	7.6	10.7	14.3	17.5	19.1	18.7	16.0	11.4	7.1	4.3	48.52	2.20
Rennes	4.8	5.3	7.9	10.1	13.1	16.2	17.9	17.8	15.7	11.6	7.8	5.4	48.05	-1.41
Strasbourg	0.4	1.5	5.6	9.8	14.0	17.2	19.0	18.3	15.1	9.5	4.9	1.3	48.35	7.45
Toulouse	4.7	5.6	9.2	11.6	14.9	18.7	20.9	20.9	18.3	13.3	8.6	5.5	43.36	1.26

L'ACP va nous aider à simplifier les données de sorte à

- identifier les profils de villes en terme de températures
- étudier la variabilité de ces profils.

Moyennes des températures mensuelles sur la période 1991-2020 dans 14 villes françaises après réduction.

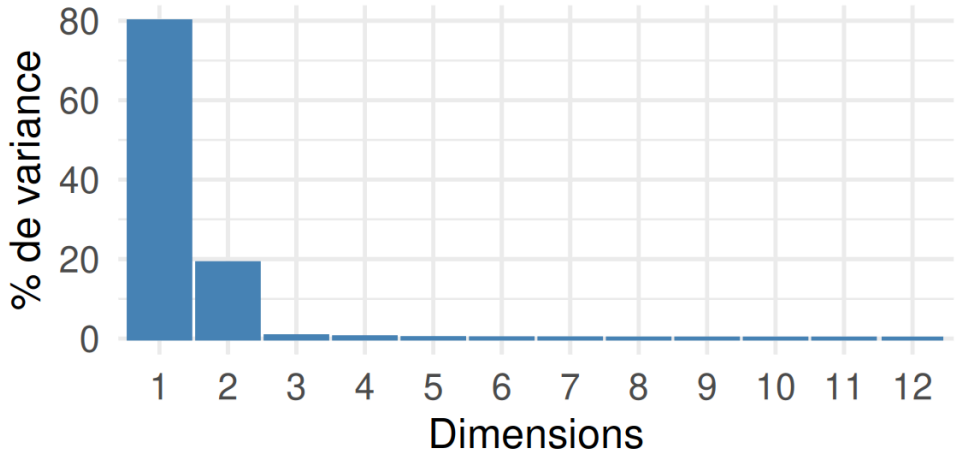
Températures mensuelles centrées-réduites

	Janv	Févr	Mars	Avri	Mai	Juin	juil	Août	Sept	Octo	Nove	Déce
Bordeaux	0.74	0.88	1.28	1.21	0.85	0.76	0.47	0.66	0.81	0.74	0.59	0.63
Brest	0.99	0.46	-0.33	-1.29	-1.87	-1.88	-1.94	-1.74	-1.24	-0.22	0.53	1.03
Clermont	-0.73	-0.65	-0.52	-0.53	-0.45	-0.32	-0.21	-0.25	-0.45	-0.65	-0.78	-0.68
Grenoble	-1.27	-0.92	-0.40	-0.32	0.00	-0.05	0.10	-0.06	-0.19	-0.55	-0.83	-1.33
Lille	-0.83	-1.07	-1.49	-1.50	-1.36	-1.40	-1.26	-1.21	-1.24	-1.08	-1.05	-0.73
Lyon	-0.98	-0.86	-0.40	-0.11	0.26	0.33	0.38	0.23	-0.08	-0.55	-0.72	-0.93
Marseille	0.70	0.88	1.09	1.35	1.50	1.58	1.56	1.53	1.50	1.39	1.19	0.98
Montpellier	0.74	0.93	1.02	1.21	1.11	1.20	1.28	1.29	1.18	1.17	1.08	0.78
Nantes	0.45	0.19	0.06	-0.18	-0.38	-0.37	-0.49	-0.49	-0.35	-0.12	0.10	0.28
Nice	1.68	1.88	1.60	1.56	1.43	1.20	1.28	1.39	1.71	1.93	1.90	1.63
Paris	-0.34	-0.44	-0.46	-0.25	-0.12	-0.21	-0.35	-0.44	-0.56	-0.55	-0.50	-0.33
Rennes	0.35	0.19	-0.27	-0.67	-0.90	-0.91	-0.89	-0.88	-0.72	-0.44	-0.12	0.23
Strasbourg	-1.81	-1.81	-1.75	-0.88	-0.32	-0.37	-0.40	-0.64	-1.03	-1.57	-1.70	-1.83
Toulouse	0.30	0.35	0.57	0.38	0.26	0.44	0.47	0.62	0.66	0.48	0.32	0.28

Note

(Lati et Long sont des variables supplémentaires → mises de côté)

2.12 Premier résultat : les valeurs propres



On cherche à

- Visualiser les données, idéalement en 2D → Le plan d'inertie maximale est-il pertinent ? Et les suivants ?
- Réduire la dimension des données → Jusqu'à combien de dimensions peut-on projeter le nuage de points ?

2.13 Qualité de la représentation

On a vu que l'inertie totale du nuage est $\sum_{j=1}^{p} \lambda_j$. La **part d'inertie expliquée** par les k premiers axes est :

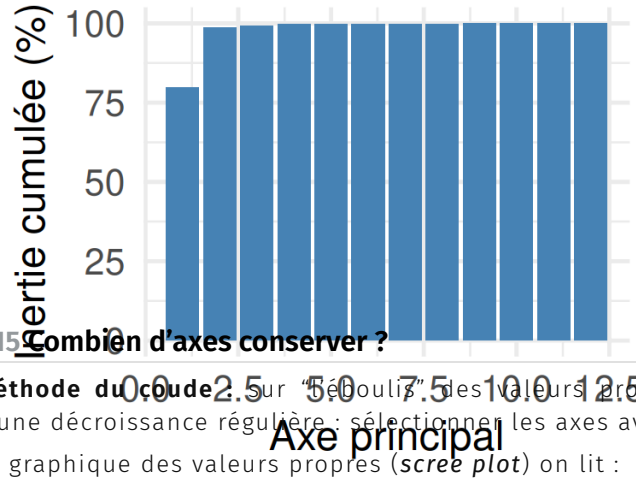
$$i_k = \frac{\sum_{j=1}^{k} \lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$

Choix du nombre de dimension : se fait souvent sur des critères empiriques… Cela dépend aussi de la finalité de l'ACP :

- pour de la description de données on a souvent $q \leq 4$ (au delà difficultés d'interprétation),
- pour de la compression q peut être beaucoup plus grand

2.14 Combien d'axes conserver ?

▸ Température des villes



2.15 Combien d'axes conserver ?

Méthode du coude : Sur "éboulis" des valeurs propres, on observe un décrochement (coude) suivi d'une décroissance régulière : sélectionner les axes avant le décrochement.

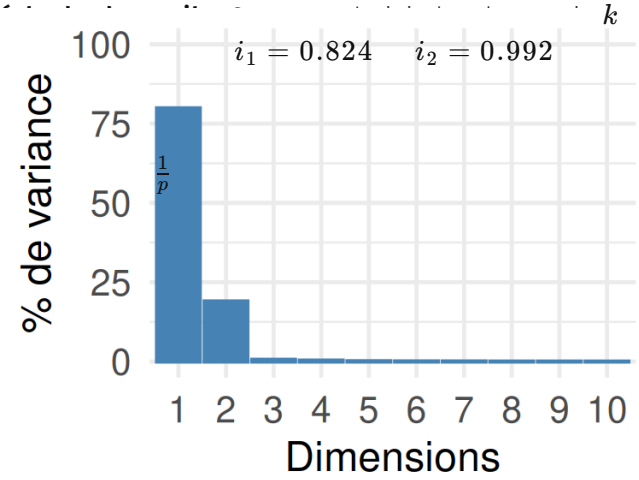
Le graphique des valeurs propres (***scree plot***) on lit :

▸ Température des villes

M k our lequel i_k dépasse un seuil. Si on prend 0.90, d lonne 2 axes.

M l'on peut interpréter...

M cipales qui expliquent une proportion de l'inertie su $q = 1$. Mais mieux vaut couper avant/après un d



Ici, on peut tout à fait ne garder que les 2 premières composantes principales.

3 Étude du nuage des individus

3.1 Résultats de l'ACP

- Les coordonnées des individus sur les axes principaux sont les **composantes principales** :

$$C = XP \in \mathbb{R}^{n \times p},$$

où $P = (u_1, \ldots, u_p) \in \mathbb{R}^{p \times p}$ est la matrice dont les colonnes sont les vecteurs propres.

- La colonne $C_{:,k} \in \mathbb{R}^n$ donne les coordonnées des n individus sur l'axe k.

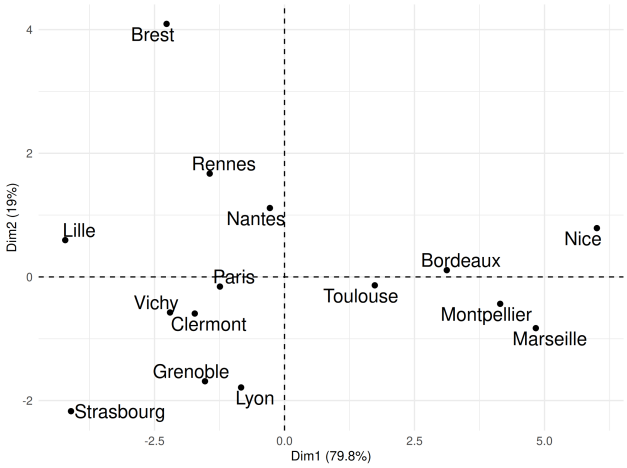
$$C_{:,k} = \begin{pmatrix} \langle X_{1,:}, P_{:,k} \rangle \\ \vdots \\ \langle X_{n,:}, P_{:,k} \rangle \end{pmatrix}.$$

- Les premières colonnes de C portent le plus d'**inertie** : les valeurs propres sont classées par ordre décroissant,

> **Note**
>
> On peut représenter le nuage des individus dans le plan défini par les deux premiers axes principaux en considérant les deux premières colonnes de C.

3.2 Visualisation des individus

Représentation du nuage de points par rapport aux 2 premières composantes principales.



- Axe 1 : oppose villes les plus chaudes aux plus froides (Nice et Lille).
- Axe 2 : oppose les climats océaniques aux climats plus continentaux, amplitude saisonnière (Brest et Strasbourg).

3.3 Qualité de représentation

La **qualité de représentation d'un individu** i sur un axe principal k est mesurée par le cosinus carré de l'angle entre l'axe factoriel et le vecteur $X_{i,:}$ (ou $Z_{i,:}$ si centré-réduit) :

$$0 \leq \cos^2_{i,k} = \frac{C^2_{i,k}}{d^2(i,G)} \leq 1,$$

où $d^2(i,G)$ est la distance au carré entre l'individu i et le centre de gravité.

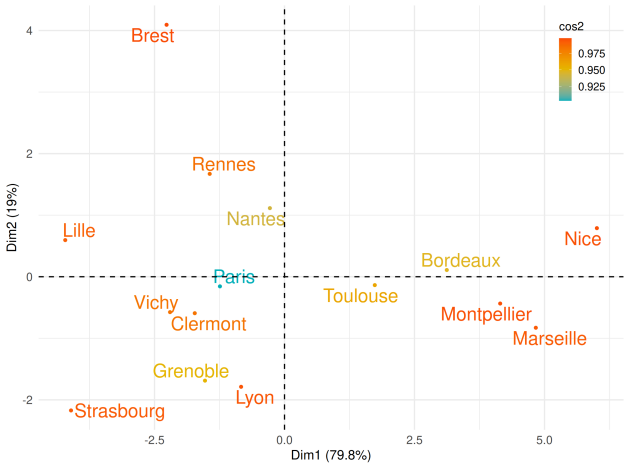
- Plus le cosinus carré est proche de 1, plus l'individu sera proche de l'axe factoriel et donc sera bien représenté sur cet axe ($\cos^2$ proche de 0 : l'axe représente **peu** l'individu).
- Qualité de la représentation sur un espace E_q : on ajoute simplement les q cosinus carrés. Dans le plan (k,l), on a:

$$\cos^2_{i,(k,l)} = \cos^2_{i,k} + \cos^2_{i,l}\,.$$

  Un point dont le $\cos^2$ est élevé dans le plan est donc **bien représenté** par ce plan.

> **Astuce**
>
> Il est préférable d'interpréter les points **bien représentés**.

3.4 Représentation des individus



- Tous les individus sont correctement représentés sur le premier plan principal.
- Brest, Lyon et Nantes sont quasiment orthogonaux à C_1.

	Dim.1	Dim.2
Bordeaux	0.947	0.001
Brest	0.234	0.763
Clermont	0.880	0.104
Grenoble	0.429	0.523
Lille	0.972	0.019
Lyon	0.178	0.817
Marseille	0.964	0.028
Montpellier	0.986	0.011
Nantes	0.056	0.886
Nice	0.980	0.017
Paris	0.889	0.014
Rennes	0.420	0.567
Strasbourg	0.776	0.217
Toulouse	0.953	0.006
Vichy	0.922	0.063

3.5 Contribution

La contribution de l'individu i à l'axe k est :

$$\text{contrib}_{i,k}^{(ind)} = \frac{C_{i,k}^2}{n\lambda_k} = \frac{\text{inertie de } i \text{ sur l'axe } k}{\text{inertie totale de l'axe } k}.$$

On a

$$0 \leq \text{contrib}_{i,k}^{(ind)} \leq 1$$

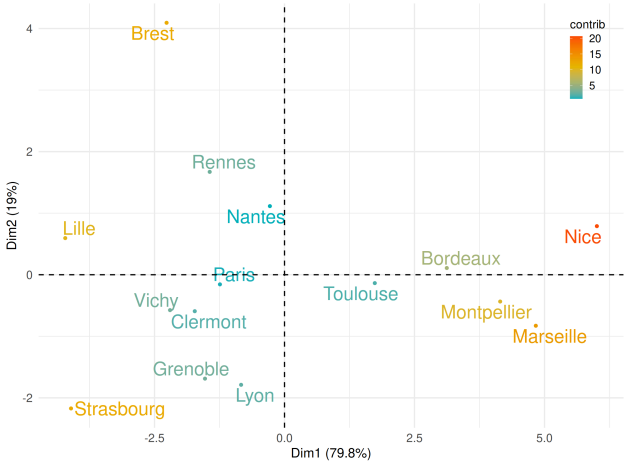
et

$$\sum_{i=1}^{n} \text{contrib}_{i,k}^{(ind)} = 1.$$

Ou parfois en pourcentage (somme à 100)

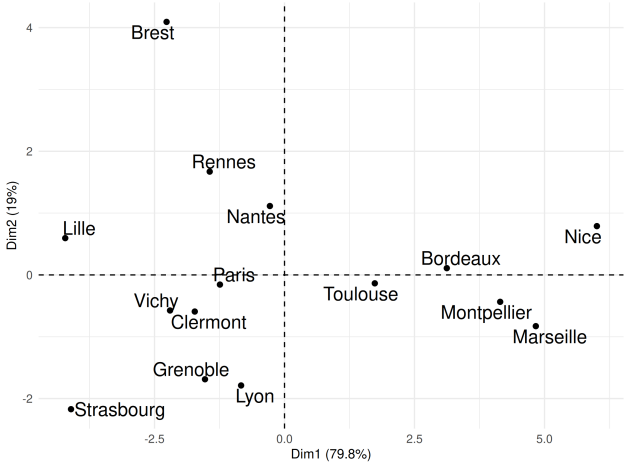
Interprétation : un individu qui a une grande coordonnée en valeur absolue sur l'axe k contribue fortement à la construction de cet axe.

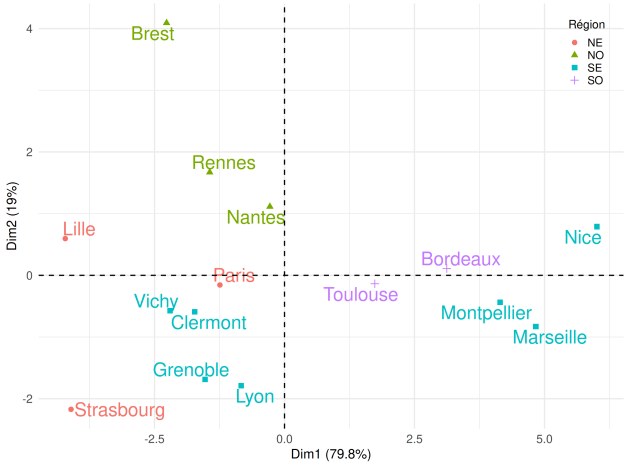
3.6 Visualisation des individus



	Dim.1	Dim.2
Bordeaux	6.776	0.035
Brest	3.579	49.069
Clermont	2.073	1.028
Grenoble	1.627	8.344
Lille	12.372	1.037
Lyon	0.485	9.365
Marseille	16.250	2.012
Montpellier	11.967	0.555
Nantes	0.055	3.638
Nice	25.106	1.825
Paris	1.073	0.072
Rennes	1.440	8.178
Strasbourg	11.728	13.819
Toulouse	2.097	0.054
Vichy	3.372	0.969

3.7 Visualisation des individus (ajout de variables supplémentaires)





- Ajout d'une variable supplémenaire qualitative (Région).

4 Étude du nuage des variables

4.1 Résultats de l'ACP

- Par définition, la matrice de covariance de $C \in \mathbb{R}^{n\times p}$ est :

$$S_C = \frac{1}{n} C^\top C = \Lambda.$$

Ainsi,

$$\mathrm{Cov}(C_{:,k}, C_{:,k'}) = \frac{1}{n}\langle C_{:,k}, C_{:,k'}\rangle \begin{cases} 0 & \text{si } k \neq k', \\ \lambda_k & \text{si } k = k'. \end{cases}$$

Les **composantes principales sont donc non corrélées deux à deux**.

- De plus, la variance de la composante principale $C_{:,k}$ est égale à la valeur propre associée :

$$\mathrm{Var}(C_{:,k}) = \lambda_k.$$

- La valeur propre λ_k correspond donc à l'**inertie portée par l'axe principal** k.

4.2 Résultats de l'ACP

La corrélation entre une variable $X_{:,j} \in \mathbb{R}^n$ et une composante principale $C_{:,k} \in \mathbb{R}^n$ est donnée par :

$$\mathrm{Cor}(X_{:,j}, C_{:,k}) = \frac{\mathrm{Cov}(X_{:,j}, C_{:,k})}{\sqrt{\mathrm{Var}(X_{:,j})}\sqrt{\mathrm{Var}(C_{:,k})}} = \frac{(SP_{:,k})_j}{\sqrt{\mathrm{Var}(X_{:,j})}\sqrt{\lambda_k}}$$

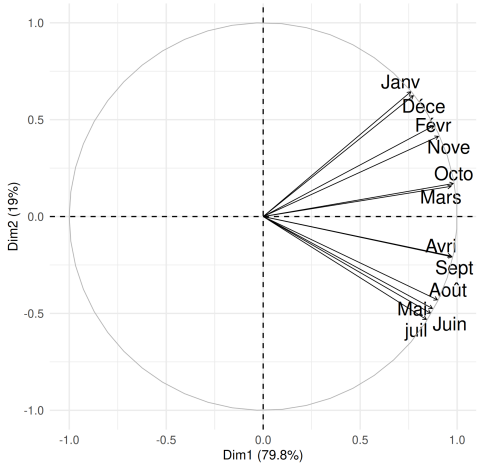
Or, puisque $SP_{:,k} = \lambda_k P_{:,k}$, on a

$$W = (\mathrm{Cor}(X_{:,j}, C_{:,k}))_{j,k=1}^{p} = \Sigma_X^{-1/2} P \Lambda^{1/2}$$

où $\Sigma_X = \mathrm{Diag}\left(\mathrm{Var}(X_{:,1}), \ldots, \mathrm{Var}(X_{:,p})\right) \in \mathbb{R}^{p\times p}$ est la matrice diagonale des variances des variables (si les données sont centrées-réduites, alors $\Sigma_X = I_p$).

- **Interprétation des composantes principales :** On mesure la relation entre C_k et la variable $X_{:,j}$ à l'aide de la corrélation $\mathrm{Cor}(C_k, X_{:,j})$.
- **Représentation** des variables dans le plan défini par les deux premiers axes en considérant les deux premières colonnes de W. Étant donnés C_1 et C_2, on trace le "cercle de corrélations" : la variable j est représentée par un point de coordonnées $(\mathrm{Cor}(C_1, X_{:,j}), \mathrm{Cor}(C_2, X_{:,j}))$ dans le disque.

4.3 Visualisation des variables: le cercle des corrélations



- Le coefficient de corrélation entre `Oct` et C_1 vaut $\sim 0{,}95$
- Tous les mois sont fortement et positivement corrélés à C_1.
- Les mois `Mars`, `Oct`, `Avri`, `Sept` (inter saison) sont les moins fortement corrélés à C_2.
- Correlation de signe opposé entre C_2 et les mois `Janv`, `Dece`, `Fevr` (hiver) et `Aout`, `Mai`, `Juin`, `Juil` (été).

- Les variables proches ont des corrélations similaires avec les deux axes.
- Des directions opposées indiquent des corrélations opposées.

4.4 Qualité de la représentation

- La qualité de représentation d'une variable j sur le k-ème axe factoriel dual est exprimée par le coefficient de corrélation linéaire[4] $\mathrm{Cor}(X_{:,j}, C_{:,k})$ entre la variable initiale $X_{:,j}$ et la nouvelle variable C_k.
- La valeur de cette corrélation sera également très importante pour interpréter ces nouveaux axes factoriels en fonction des variables initiales.
- Les variables les mieux expliquées correspondent aux points les plus proches du cercle unité.

[4] Quand les données sont centrée-réduites, on peut en prendre le carré et retrouver le $\cos^2$

Remarque : bien qu'ayant retenu un sous-espace factoriel expliquant une part importante de l'inertie totale, il est possible que certaines variables ou certains individus d'intérêt soient mal représentés dans ce sous-espace. Il sera alors intéressant de compléter le sous-espace factoriel en ajoutant des axes factoriels supplémentaires de sorte que ces variables ou individus d'intérêt soient bien représentés.

4.5 Contribution

Pour une variable j, sa contribution à l'axe k est :

$$\text{contrib}_{j,k}^{(\text{var})} = \frac{W_{j,k}^2}{p\lambda_k}.$$

On a également

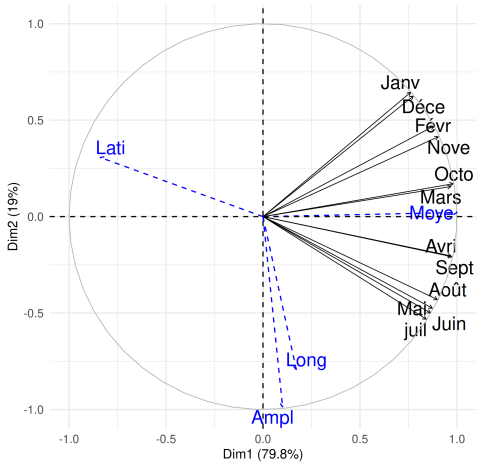
$$0 \leq \text{contrib}_{j,k}^{(\text{var})} \leq 1$$

et

$$\sum_{j=1}^{p} \text{contrib}_{j,k}^{(\text{var})} = 1.$$

Interprétation : une variable qui a une forte corrélation avec l'axe k contribue fortement à sa construction.

4.6 Ajout de variables supplémentaires



- La corrélation entre C_1 et `Moye` (température moyenne) est très proche de 1.
- La latitude `Lati` est également fortement corrélée à cet axe.
- L'amplitude annuelle `Ampl` est très fortement corrélée à C_2 .
- La longitude `Long` est également fortement corrélée à cet axe.

4.7 Exploiter sa connaissance des données

- On remarque dans les données que, tout au long de l'année, il fait plus froid à Lille et plus chaud à Nice relativement aux autres villes de France.
- On peut vérifier que la **température moyenne annuelle** est très fortement corrélée à C_1.
- On remarque également que Brest est la ville où les **amplitudes thermiques** sont les plus faibles au cours de l'année, contrairement à Strasbourg.
- On peut vérifier que l'**amplitude thermique annuelle** est très fortement corrélée à C_2.

Astuce

La richesse des interprétations vient surtout de la connaissance des données.

Il n'est pas toujours aussi aisé de faire des liens directs avec des variables supplémentaires !

4.8 La dualité individus / variables

- L'ACP peut être vue de deux façons : deux nuages sont construits à partir de la **même matrice de données** X
- **Nuage des individus** : les n individus sont des points de $\mathbb{R}^p$.
- **Nuage des variables** : les p variables sont des points de $\mathbb{R}^n$.
- **Deux matrices, un même problème.** Pour les individus, on étudie :

$$S_X = \frac{1}{n} X^\top X$$

Pour les variables, on étudie :

$$S_V = \frac{1}{n} X X^\top.$$

Or $X^\top X$ et $X X^\top$ ont les **mêmes valeurs propres non nulles**.

> **Astuce**
>
> **C'est la dualité de l'ACP :** étudier les individus ou étudier les variables conduit aux **mêmes valeurs propres** et donc aux **mêmes axes principaux**. Voir la SVD.

4.9 La dualité individus / variables

- **Pourquoi retrouve-t-on les mêmes axes ?**

Si

$$X^\top X u_k = \lambda_k u_k,$$

alors

$$X^\top (X u_k) = \lambda_k u_k.$$

En posant

$$v_k = \frac{X u_k}{\sqrt{\lambda_k}},$$

on obtient

$$X X^\top v_k = \lambda_k v_k.$$

Ainsi, les vecteurs propres de $X^\top X$ et de $X X^\top$ sont reliés directement par X.
La représentation des individus et celle des variables sont donc deux faces du même objet.

4.10 Vocabulaire

- **Espaces factoriels directs :** ce sont les p sous espaces vectoriels $\mathrm{Vect}\{u_k\}$ engendrés par les $u_k \in \mathbb{R}^p$ (droite/directions de plus grande inertie).
- **Espaces factoriels duals :** ce sont les sous espaces vectoriels engendrés par les composantes principales. La k-ème composante principale est le vecteur $C_k \in \mathbb{R}^n$ des coordonnées des n individus sur l'axe factoriel $\mathrm{Vect}\{u_k\}$. En ACP normée on a:

$$C_k = X u_k$$

La composante principale C_k est donc une combinaison linéaire des variables $X_{:,j}$ et peut être vue comme une nouvelle variable dans le sous-espace factoriel dual.

4.11 Nuage des variables

- **Problème :** chaque variable est un point de $\mathbb{R}^n$, et on souhaite obtenir une représentation dans un espace de dimension $m < n$, typiquement $m = 2$.

On cherche à nouveau m directions orthogonales

$$v_1, \dots, v_m,$$

mais cette fois en maximisant la **corrélation globale avec les variables**.

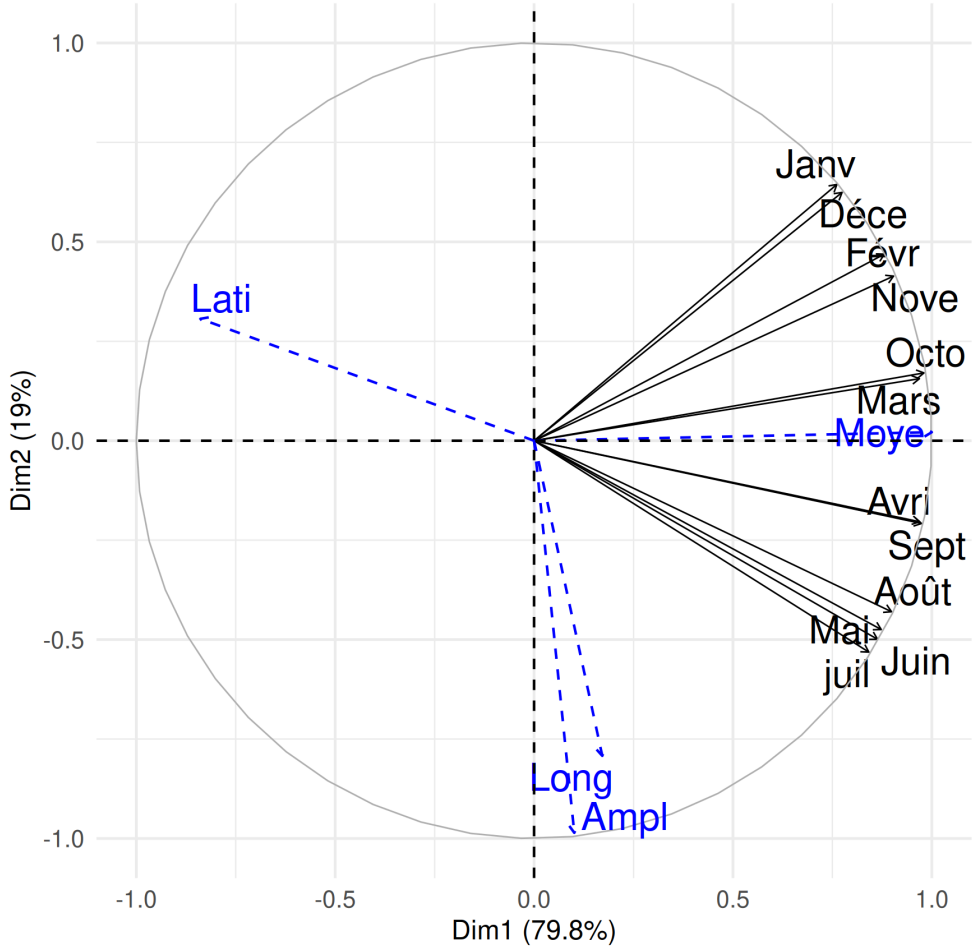
- **Solution :** sous espaces vectoriels emboîtés

1. v_1 est la direction maximisant $\sum_{j=1}^{p} \mathrm{Cor}^2(v, X_{:,j})$.
2. v_2 est la direction orthogonale à v_1 qui maximise $\sum_{j=1}^{p} \mathrm{Cor}^2(v, X_{:,j})$.
3. $\dots$

où $\mathrm{Cor}(U, V)$ désigne le **coefficient de corrélation** entre U et V.

4.12 Plot twist

On retrouve **exactement la même représentation** que lors de l'étude du nuage des individus !



4.13 Exemple : temperatures de villes francaises

Pour des temperatures mensuelles mesurées dans plusieurs villes francaises, une ACP sur les douze mois peut mettre en évidence :

- un premier axe lié au niveau moyen de température, opposant notamment les villes les plus chaudes aux plus froides ;

- un second axe lié à l'amplitude saisonnière, opposant par exemple des climats océaniques aux climats plus continentaux.

La temperature moyenne annuelle, la latitude, l'amplitude annuelle et la longitude peuvent être ajoutées comme variables supplémentaires pour aider à nommer les axes, sans modifier leur construction.

4.14 Realiser une ACP dans R

Le package `FactoMineR` fournit la fonction `PCA()`. Les colonnes actives doivent être quantitatives ; les variables supplémentaires peuvent être déclarées séparément.

```
> library(FactoMineR)
>
> resultat <- PCA(
+   X,
+   ncp = 5,
+   scale.unit = TRUE,
+   graph = TRUE
+ )
```

L'objet retourné contient notamment :

- `eig` : valeurs propres, pourcentages et pourcentages cumulés d'inertie ;
- `ind$coord`, `ind$contrib`, `ind$cos2` : résultats pour les individus ;
- `var$coord`, `var$contrib`, `var$cos2` : résultats pour les variables.

4.15 Autres implémentations de l'ACP en R

R propose aussi `prcomp()` et `princomp()`.

4.15.1 Avec `prcomp()`

```
> X.acp <- prcomp(
+   X,
+   center = TRUE,
+   scale. = TRUE
+ )
```

Principaux résultats :

- `sdev` : écarts-types des composantes ;
- `rotation` : axes principaux ;
- `x` : coordonnées des individus.

Valeurs propres :

```
> valeurs_propres <- resultat$sdev^2
```

4.16 Réaliser une ACP en Python

On recommande l'ACP implémentée dans `scikit-learn`. Voir la **documentation**

Centrage et réduction:

```
> from sklearn.preprocessing import StandardScaler
>
> scaler = StandardScaler()
> X_scaled = scaler.fit_transform(X)
```

Calculer l'ACP

```
> from sklearn.decomposition import PCA
>
> pca = PCA(n_components=5)
> resultat = pca.fit_transform(X_scaled)
```

Principaux résultats :

`pca.explained_variance_` : valeurs propres ; `pca.explained_variance_ratio_` : parts d'inertie ; `pca.components_` : axes principaux ; `resultat` : coordonnées des individus.

4.17 À retenir

L'ACP fournit une représentation de dimension réduite qui conserve au mieux l'inertie du nuage des individus. Elle aide à comparer des profils, à résumer les liaisons entre variables et à construire une interprétation informée par les données et leur contexte.

4.18 Références