Analyse en Composantes Principales

Michel Deudon, Laurent Piccinini

Université Paul Valéry, Département MIAp
prenom.nom@univ-montp3.fr

Abstract. L'analyse en composantes principales est l'une des techniques les plus utilisées de réduction de dimension en sciences des données. Dans ce chapitre, nous verrons comment appliquer la décomposition en valeur singulière pour réduire la dimensionalité et visualiser des données de grandes dimensions.

Keywords: Algèbre linéaire, factorisation de matrices, analyse en composantes principales (ACP), décomposition en valeur singulière (SVD), multi dimensional scaling (MDS), réduction de dimension

1 Définitions et notations

Vecteurs et espaces vectoriels Dans ce cours, $E = \mathbb{R}^d$ désignera un espace vectoriel de départ et $F \subset E$ un sous espace vectoriel de E, de dimension $d' < d$. Nous utiliserons $x \in \mathbb{R}^d$ pour désigner un vecteur d-dimensionnel de E, et x' pour désigner des coordonnés de ce point dans l'espace d'arrivé F.

Matrices Nous utiliserons $X = (x_{ij})_{1 \leq i \leq n, 1 \leq j \leq d,} \in \mathbb{R}^{nd}$ pour désigner une matrice rectangulaire à n lignes et d colonnes, correspondant à n observations (ou individus) en dimension d, i.e. chaque observation x_i est décrite par d variables réelles. On appelle matrice de covariance empirique de X la matrice carrée

$$C_{=}(X-m)^T(X-m) = \frac{1}{n}\sum_{i=1}^{n}(x_i-m)(x_i-m)^T \in \mathbb{R}^{dd}$$

où m est la moyenne empirique. Elle vérifie $C^T = C$ et donc est symmétrique. $C_{ij} > 0$ (resp. < 0) si X_i et X_j sont corrélées (resp. anti-corrélées) et nul sinon.

Produit scalaire Nous noterons $u.v = ||u||||v||cos(\widehat{u,v}) \in \mathbb{R}$ le produit scalaire entre deux vecteurs u et v. Il s'agit d'une forme bilinéaire, symmétrique, définie, positive.[1] Dans une base orthonormée, le produit scalaire canonique s'écrit

$$u.v = \sum_j u_j v_j$$

[1] Plus généralement, si Σ est une matrice diagonale dont les éléments diagonaux sont strictement positifs, alors $\psi(u,v) = u^T \Sigma v = \sum_j u_j \sigma_{jj} v_j$ définit un produit scalaire.

aussi noté $u^T v$ et np.dot(u,v) avec numpy [7]. On note $||u|| = \sqrt{u.u}$ la norme asssociée et $d(u, v) = ||u - v||$ la distance entre u et v.

Base orthonormée Soit E un espace vectoriel muni d'un produit scalaire. On dit que $u, v \in E^2$ sont orthogonaux si $u.v = 0$ et que u est normé si $||u|| = 1$. On dit que $(e_1...e_d)$ forment une base de E si tout vecteur $x \in E$ peut s'écrire comme une unique combinaison linéaire des vecteurs e_i, i.e., il existe d réels $(a_1...a_d)$ tels que $x = \sum_i a_i e_i$. Les réels a_i constituent les coordonnées de x dans la base $(e_1...e_d)$. Dans une base orthonormée, tout vecteur x peut s'écrire

$$x = \sum_i (x.e_i)e_i$$

2 L'ACP de 2D à 1D

Soit n points (quasi) alignés sur la droite d'équation $y = -x$, comme illustré sur la figure 1. Chaque point a pour coordonnées $(x, -x)$ dans la base canonique (e_i, e_j) de $E = \mathbb{R}^2$. Dans la base orthonormée (v_1, v_2), chaque point a pour coordonnées $(\sqrt{2}x, 0)$ par Pythagore, noté $x' = \sqrt{2}x$ en une dimension.

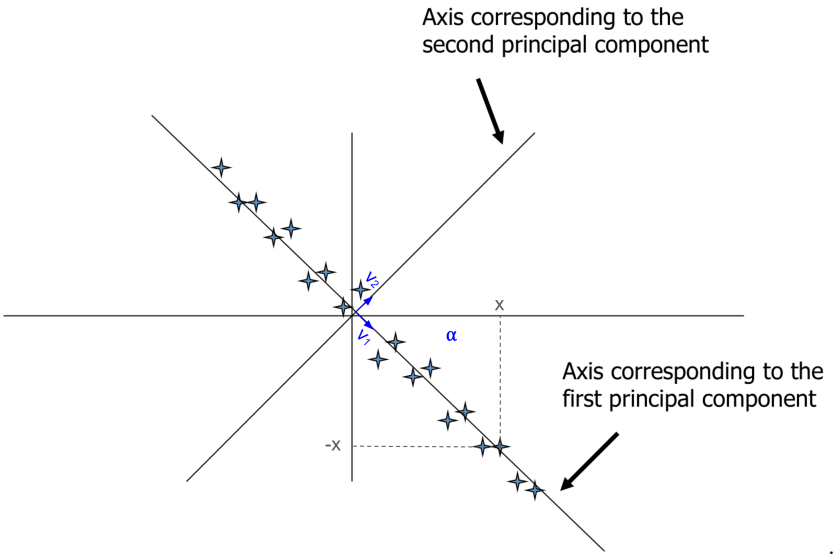

Fig. 1. Géométriquement, l'ACP correspond à une rotation d'angle α puis une projection orthogonale (ici sur une droite). Comment trouver l'axe qui maximise la variance?

Plus généralement, pour tout angle α, $V = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} cos(\alpha) & sin(\alpha) \\ -sin(\alpha) & cos(\alpha) \end{pmatrix}$ forme une base orthonormée de $\mathbb{R}^2$, pratique pour réduire la dimension! En effet, on a $cos^2 + sin^2 = 1$ et $(a, b).(-b, a) = 0$.

3 L'ACP en 3D et plus

En grandes dimensions, certaines coordonnées peuvent être fortement corrélés ou ne pas contribuer à la représentation des données. Le plus proche voisin est déformé. Les intuitions de basse dimension ne s'appliquent pas [1].

3.1 Formulation du problème

Soit n observations de E, supposées indépendantes et identiquements distribués (hypothèse iid).[2] Par exemple, chaque observation peut décrire des caractéristiques d'un pays (surface, revenu moyen, population active, etc) ou d'un jeu (est éducatif, est aléatoire, note moyenne, prix, etc). Nous souhaitons apprendre une fonction ϕ (un *mapping*) de l'espace de départ E vers un espace d'arrivé F, qui préserve la *structure* et *l'information*, intuitivement qui maximise la variance des données. Nous faisons l'hypothèse que ϕ est linéaire. L'ACP peut être interprétée comme une rotation suivie d'une projection sur une base orthonormée de $F \subset E$. En termes d'écriture matricielle, cela correspond à une factorisation de matrice $X = (U\Sigma)V$ que nous détaillerons dans la section suivante.

3.2 Décomposition de matrices

Valeurs propres, vecteurs propres Soit C une matrice carrée. On dit que λ est une valeur propre de C, associée au vecteur propre v, si $Cv = \lambda v$. Les valeurs propres de C, sont solutions de det(C -λI) = 0. Le rang d'une matrice correspond au nombre de lignes ou colonnes linéairement indépendantes, de manière équivalente au nombre de valeurs propres non nulles.

Diagonalisation Si $C \in \mathbb{R}^{dd}$ est une matrice carré, réelle, symmétrique, alors en notant $\lambda_1 \leq ... \leq \lambda_d$ les valeurs propres de C associées aux vecteurs propres V, on a

$$C = V\Sigma V^T = \lambda_1 v_1 v_1^T + ... + \lambda_d v_d v_d^T$$

où $v_i v_i^T$ représente la projection via v_i, associée à λ_i et $\Sigma = diag(\lambda_i)$.

[2] Dans le cas non iid, on peut appliquer des poids différents pour chaque observation.

Décomposition en valeur singulière (SVD) La décomposition en valeurs singulières généralise le résulat précédent à des matrices rectangulaires $X \in \mathbb{R}^{nd}$:

$$X = U\Sigma V^T$$

$$U \in \mathbb{R}^{nn},\ \Sigma \in \mathbb{R}^{nd},\ V \in \mathbb{R}^{dd}\ /\ U^TU = I_n,\ VV^T = I_d$$

i.e. U, V définissent une base orthonormée de $\mathbb{R}^n$ et $\mathbb{R}^d$. Les colonnes de V sont les vecteurs propres de $C = X^TX$ et Σ contient les racines carrées des valeurs propres associées. En effet, si $X = U\Sigma V^T$, alors $C = V\Sigma^2V^T$. De même, on vérifie que les colonnes de U sont les vecteurs propres de $G = XX^T$.

Résultat fondamental Avec les d' plus grandes valeurs propres de Σ, SVD donne la meilleure approximation de rang d' d'une matrice X, minimisant la norme $||Y - X||_F = \sum_i \sum_j (Y_{ij} - X_{ij})^2$, avec

$$Y^*_{d'} = U_{d'}\Sigma_{d'}V^T_{d'}$$

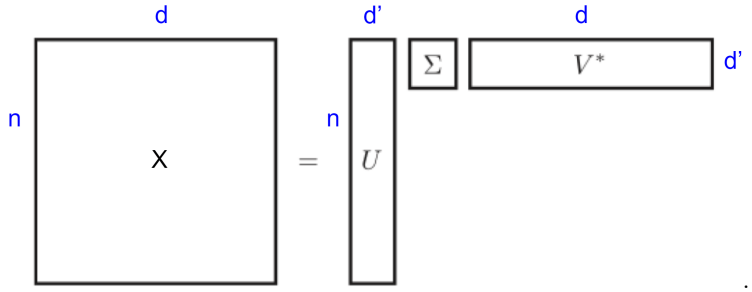

Fig. 2. L'ACP comme un SVD partiel (matrices tronquées).

3.3 Retour à l'ACP et approximation de rang inférieur

PCA est la même chose qu'un SVD partiel. $X \approx X'V^T_{d'}$ avec $X' = U_{d'}\Sigma_{d'} \in \mathbb{R}^{nd'}$ et où $V_{d'} \in \mathbb{R}^{dd'}$ définit une base orthonormée de F, formé de vecteurs propres correspondant aux plus grandes valeurs propres de C.[3] L'axe principal est une droite dirigée par le vecteur unitaire $v_1 \in \mathbb{R}^d$ qui maximise $\frac{1}{n}v^TCv$. λ_1 correspond à la variance expliquée par v_1. On appelle composantes principales la projection $x' = V^T_{d'}x$ (après normalisation) de x sur F. La qualité de la représentation obtenue par d' valeurs propres est la proportion de variance expliquée : $\frac{\sum_{j=1}^{d'}\lambda_j}{\sum_{j=1}^{d}\lambda_j}$. En pratique on choisit d' de facon à retenir 80% ou plus de la variance initiale.

[3] Les v_i sont aussi appelés facteurs principaux. v_i n'est pas unique ($-v_i$ solution).

4 L'ACP en pratique

4.1 Prétraitement des données

Afin que toutes les entrées de X aient une amplitude similaire, nous normalisons les données avant de faire l'ACP sur des données centrées réduites

$$x_{ij} = \frac{x_{ij} - m_j}{\sigma_j}$$

où $m_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$ et $\sigma_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - m_j)^2$. En test, on normalise les données avec la même moyenne et écart type que pendant la phase d'entrainement!

4.2 Pseudo code

Data: $X \in \mathbb{R}^{nd}$, $k \in [1..n]$
Result: X' en dimensions réduites
Calculer la moyenne empirique $m = \frac{1}{n}\sum_{i=1}^{n} x_i$;
Centrer les colonnes de $X = (X - m)$;
Calculer la matrice de covariance empirique dd, $C = X^T X$;
Calculer les valeurs et vecteurs propres de $C = V\Sigma V^T$;
Conserver les k plus grandes valeurs et vecteurs propres ;
Projecter X dans le nouvel espace de dimension réduit $X' = XV_k$;
return X';
Algorithm 1: Pseudo code de l'analyse en composantes principales

4.3 Notion de similarité

On peut mesurer la similarité entre u et v avec le cosinus de l'angle qui les sépare.

$$cos(\widehat{u,v}) = \frac{u.v}{||u||||v||}$$

Numériquement, il s'agit d'un nombre dans $[-1, 1]$, plus facile à interpréter que la distance euclidienne $d(u, v)$. Lorsque $cos(\widehat{u,v}) = 1$ (resp.-1), u et v sont colinéaires, de même sens (resp. sens opposés). Lorsque $cos(\widehat{u,v}) = 0$, les vecteurs sont orthogonaux. Aussi appelée *cosine similarity*, cette fonction est souvent utilisées en recherche d'informations.

5 Autres méthodes de factorisation matricielle

L'ACP minimise une erreur de reconstruction entre une fonction linéaire ϕ de E dans F, et $\tilde{\phi}$ de F dans E:

$$\phi(x) = V^T(x - m), \quad \tilde{\phi}(x') = m + Vx'$$

L'erreur entre x et sa reconstruction $\tilde{\phi}(\phi(x))$ peut s'écrire

$$\frac{1}{n}\sum_{i=1}^{n}||x_i - \tilde{\phi}(\phi(x_i))||^2 = \frac{1}{n}\sum_{i=1}^{n}||x_i - (m + VV^T(x_i - m))||^2$$

Dans le cas de l'ACP où nous avons la contrainte $V^TV = I$, nous avons une solution explicite: m est la moyenne empirique et les colonnes de V toute base orthonormée de E, couverte par les d' premiers vecteurs propres de la matrice de covariance C. SVD apparaît comme la solution optimale. D'autres contraintes sur V ou les coordonnées dans F sont possibles et reportés dans le tableau ci-dessous. L'Analyse en Composantes Indépendantes (ICA) [2] [4] [5] est motivé par la séparation de signals, dont l'exemple le plus classique est le problème de séparation de sources dans un cocktail, où plusieurs personnes parlent simultanément. La Factorisation Matricielle non Négative (NMF) [3] est motivé par l'interprétation de la matrice Y reconstruite par SVD. NMF est utilisé pour extraire des co-occurences de mots ou des catégories sur des corpus linguistiques, en recherche d'informations, data mining, etc. en approximant X par $X \approx UV^T$, $U \in \mathbb{R}^{nd'}$ et $V \in \mathbb{R}^{d'd}$ non négatives. Le codage parcimonieux [6] est basé sur le principe qu'une entrée est approximativement la somme de nombreux petits motifs élémentaires. Pour apprendre à la fois un dictionnaire ainsi que les poids, on fixe alternativement un des paramètres pour entrainer l'autre à l'aide d'une descente de gradient stochastique.

Table 1. Hypothèses et contraintes de modèles linéaires $x' = V^T(x - m)$.

Modèle	Hypothèses / Contrainte
PCA	$V^TV = I$ (colonnes orthonormées)
ICA	Coordonnées x'_i et x'_j indépendents
NMF	V et x' non-negatifs
Codage parcimonieux	x' *sparse* (beaucoup de 0)

6 Préservation des distances

La mise à l'échelle multidimensionnelle (MDS) est une autre technique de réduction de dimension qui vise à préserver les distances entre les points dans E et F. Un critère naturel est $\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|d(x_i, x_j) - d'(x'_i, x'_j)|^2$. Si $d(x; y) = ||x - y||^2$, MDS est identique à PCA, sur une matrice différente, appelée matrice de similarité des données ou matrice de Gram,

$$G = (X - m)(X - m)^T \in \mathbb{R}^{nn}$$

La diagonalisation de $G = ULU^{-1}$ permet de calculer $X' = U_{d'}L_{d'}^{\frac{1}{2}}$ avec les d' premiers vecteurs propres. On peut appliquer MDS sans connaitre X. Sa complexité est $O(d'n^2)$ vs $O(d'd^2)$ pour PCA.

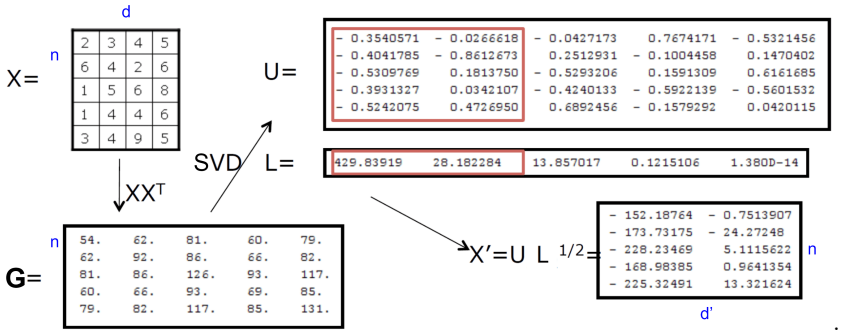

Fig. 3. Illustration de l'algorithme de mise à l'échelle multidimensionnelle.

En TP vous implémenterez PCA, MDS et SVD en binome sur un jeu de données de vins italiens, avec Python, Anaconda, Jupyter notebook et Numpy [7].

Conclusion

Maximiser la variance du jeu de données original, préserver les distances entre les points... différentes formulations peuvent mener au même résulat sous certaines conditions. La décomposition en valeur singulière (SVD) donne l'approximation optimale de rang k de X comme un produit de matrices $X'V$ sous la contrainte $V^TV = I$. ICA, NMF, le codage parcimonieux donnent la solution optimale pour d'autres hypothèses et contraintes sur X' et V, fonction de la nature du problème. L'ACP correspond géométriquement à une rotation, puis une projection. Elle est obtenue en diagonalisant la matrice de covariance $C = X^TX$ (après avoir normalisé X) puis en projetant X sur d' vecteurs propres, ce qui correspond à un SVD partiel. MDS cherche à préserver les distances entre les paires de points dans E et F, en diagonalisant une matrice de similarité ou matrice de Gram G. Lorsque la norme euclidenne L2 est utilisée, $G = XX^T$ et MDS est équivalent à l'ACP.

References

1. R. Bellman. The curse of dimensionality. Princeton. 1961.
2. P. Comon. Independent component analysis, a new concept? Signal processing. 1994.
3. P. Pentti et U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics. 1994.
4. A. J. Bell et T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. Neural computation. 1995.
5. A. Hyvärinen et O. Erkki. Independent component analysis: algorithms and applications. Neural networks. 2000.
6. J. Mairal, F. Bach et al. Online learning for matrix factorization and sparse coding. Journal of Machine Learning Research. 2010.
7. C.R. Harris, K.J. Millman et al. Array programming with NumPy. Nature. 2020.