TP Analyse en Composantes Principales (ACP)

6 octobre 2021

Résumé

Le but de ce TP est de mettre en oeuvre l'ACP :

1. d'abord sur des données simulées pour mieux comprendre le lien entre la forme du nuage de points, c'est-à-dire la distribution des données, et les résultats de l'ACP,
2. puis sur des tableaux de données réelles en utilisant les fonctionnalités de la librairie `Scikit-Learn`.

1 ACP canonique sur des données simulées

L'objectif est d'effectuer une ACP sur des données simulées pour mieux comprendre le lien entre la forme du nuage de points correspondant, c'est-à-dire la distribution des données, et les résultats (vecteurs propres et valeurs propres) fournis par l'ACP.

1.1 Génération des données

Générez une matrice $\mathbf{X}$ de 100 individus caractérisés par un couple de variables, qui suit une loi normale de vecteur moyenne $\begin{bmatrix}10 & 20\end{bmatrix}^T$ et de matrice de covariance $\begin{bmatrix}25 & -12\\ -12 & 9\end{bmatrix}$ à l'aide de la fonction `multivariate_normal` du module `numpy.random`.

Représentez le nuage de points correspondant en utilisant la fonction `scatter` du module `matplotlib.pyplot`. Que pensez-vous de la corrélation entre les 2 variables ?

1.2 Calcul des valeurs propres et vecteurs propres

Centrez les données avec la fonction `mean()` de `numpy` pour obtenir la matrice $\mathbf{Y}$ (selon les notations du cours).

Calculez la matrice de covariance empirique $\mathbf{V}$ des données centrées (rappel : la multiplication de matrices s'effectue avec la fonction `dot` de `numpy`). Comparez avec la matrice de covariance de la loi normale.

Calculez les valeurs propres et vecteurs propres de $\mathbf{V}$ à l'aide de la fonction `eig` du module `scipy.linalg`.

1.3 Affichage et interprétation des résultats

Affichez les valeurs propres, ainsi que le pourcentage d'inertie expliquée par chaque axe.

Représentez les vecteurs propres sur le nuage de points centré en utilisant la fonction `arrow` ou `quiver` de `matplotlib.pyplot`.

Commentez les résultats.

Etudiez l'impact de la distribution des données, c'est-à-dire de la forme du nuage de points, sur les valeurs propres et vecteurs propres obtenus en modifiant la matrice de covariance de la loi normale. En particulier, comparez les résultats obtenus pour différentes valeurs de covariance entre les 2 variables, par exemple les valeurs extrêmes : 0 et 15.

1.4 Utilisation de la librairie `Scikit-Learn`

Une autre façon d'obtenir les valeurs propres et vecteurs propres de la matrice de covariance empirique consiste à effectuer une ACP sur les données centrées $\mathbf{Y}$. En effet, l'ACP fournit les axes principaux qui sont les vecteurs propres et l'inertie (on parle aussi de variance) sur chaque axe qui correspond à chaque valeur propre.

Etudiez la classe `PCA` du module `sklearn.decomposition` :
- paramètres d'appel, en particulier `n_components`,
- attributs, en particulier `components_`, `explained_variance_`, `explained_variance_ratio_`,
- méthodes, en particulier `fit()` et `transform()`.

Effectuez une ACP sur les données centrées pour obtenir les axes principaux et l'inertie sur chaque axe. Retrouvez les valeurs propres et vecteurs propres de la matrice de covariance.

2 ACP normée sur des données réelles

L'objectif est d'effectuer une ACP normée sur les données d'un fichier disponible sur MyLearningSpace : `activites.txt`.

Commencez par visualiser le fichier. Il s'agit d'une enquête menée auprès de différentes populations sur le temps consacré à des activités données au cours de la journée.

La première colonne intitulée `POP` permet d'identifier des groupes de personnes. Le code utilisé est le suivant : H : Hommes, F : Femmes, A : Actifs, N : Non actifs, M : Mariés, C : Célibataires, U : USA, W : Pays de l'ouest, E : Pays de l'est, Y : Yougoslavie.

Les colonnes suivantes correspondent aux variables.
Les 10 variables numériques sont le temps passé en : `PROFession`, `TRANsport`, `MENAge`, `ENFAnts`, `COURses`, `TOILette`, `REPas`, `SOMMeil`, `TELEvision` et `LOISirs`.
Les 4 variables catégorielles sont : `SEXe` (1 = Hommes, 2 = Femmes), `ACTivité` (1= Actifs, 2 = Non Actifs, 9 = Non précisé), `état CIVil` (1 = Célibataires, 2 = Mariés, 9 = Non précisé), PAYs (1 = USA, 2 = Pays de l'ouest, 3 = Pays de l'est, 4 = Yougoslavie).

Les temps sont indiqués en centièmes d'heures. La 1ère case en haut à gauche indique que les hommes actifs aux USA passent en moyenne 6 heures et 6 minutes (10/100 ièmes d'heure) dans leur activité professionnelle. Le total sur les 10 activités est de 2400 (24 heures).

2.1 Lecture des données

Importez le fichier en mémoire à l'aide de la fonction `read_csv()` de `pandas` pour pouvoir manipuler les données. Pour rappel, la fonction `read_csv()` renvoie un objet de la classe `DataFrame`. Assurez-vous que l'importation est correcte en utilisant les fonctionnalités offertes pour les objets `DataFrame`, notamment l'affichage des informations sur les variables, ainsi que l'affichage des premières lignes : `info()`, `describe()`, `shape`, `head()`...

Quelles variables proposez-vous de conserver pour effectuer l'ACP ?

Stockez la première colonne à part : pour les représentations graphiques sur les plans principaux, il faudra préciser les noms des groupes d'individus.

Eliminez les colonnes qui ne seront pas utilisées avec la méthode `drop()` de la classe `DataFrame`. Les variables mises de côté apportent-elles une information supplémentaire ?

Stockez les noms des colonnes restantes de façon à conserver les noms des variables utilisées pour l'ACP avec l'attribut `columns` de la classe `DataFrame`.

2.2 Examen des données

Après avoir importé et sélectionné les données pour l'ACP, vous pouvez faire une première analyse des données en recherchant les relations qui existent entre les variables.

Calculez le coefficient de corrélation entre chaque couple de variables numériques en utilisant la méthode `corr()` de la classe `DataFrame`.

Il est aussi possible de visualiser graphiquement les corrélations entre variables grâce à la fonction `scatter_matrix()` de `pandas` ou `pandas.plotting`. Utilisez cette fonction qui croise deux à deux les variables numériques et affiche les nuages de points correspondants.

Quelles sont les variables les plus corrélées positivement ? négativement ? Quelles sont les variables les moins corrélées ?

2.3 Transformation des données

On va appliquer une ACP normée, donc il faut d'abord centrer et réduire les données. Il existe différentes façons de réaliser ces opérations. La plus simple est la suivante :

Commencez par convertir l'objet `DataFrame` en tableau de type `array` de la librairie `numpy`, que l'on notera $\mathbf{X}$ (selon les notations du cours). Suivant votre version de la librairie `pandas`, vous utiliserez la méthode `to_numpy()` (introduite à partir de la version v0.24.0) ou bien la méthode plus ancienne `values`.

Ensuite pour centrer et réduire les données, utilisez les fonctionnalités de la classe `StandardScaler` de `sklearn.preprocessing`. Etudiez cette classe, en particulier les méthodes `fit()` et `transform()`, puis effectuez la normalisation des données pour obtenir la matrice $\mathbf{Z}$ (toujours selon les notations du cours).

Vérifiez que la moyenne et la variance des variables centrées réduites valent bien 0 et 1 respectivement.

2.4 ACP

Effectuez l'ACP sur les données centrées réduites $\mathbf{Z}$ en utilisant la classe `PCA` du module `sklearn.decomposition`. Calculez les axes principaux et les valeurs propres.

Affichez les valeurs propres (en ordre décroissant), la part d'inertie expliquée par chaque axe (ou pourcentage de variance expliquée), ainsi que la part d'inertie cumulée à l'aide de la fonction `cumsum()` de `numpy`.

Tracez la courbe de décroissance des valeurs propres ainsi que la courbe de croissance de la part d'inertie cumulée avec la fonction `plot` du module

`matplotlib.pyplot`.

Déterminez le nombre d'axes à conserver avec chacun des 3 critères vus en cours.

Dans la suite, on s'intéressera aux 4 premiers axes principaux. Dans ces conditions, quelle est la qualité globale de la représentation ?

2.5 Représentation des individus et interprétation des résultats

Effectuez la projection de la matrice des données **Z** sur les axes principaux avec la méthode `transform()` de la classe `PCA`.

A quoi correspondent les valeurs obtenues ? Quelle formule du cours permet d'obtenir ces valeurs ?

Représentez le nuage des individus projeté dans le premier plan principal défini par les axes 1 et 2, puis dans le deuxième plan principal défini par les axes 3 et 4. Pour chaque point, précisez le nom du groupe d'individus correspondant avec la fonction `text()` du module `matplotlib.pyplot`.

Calculez les contributions des individus à chacun des 4 axes. A l'aide de ces valeurs et des 2 figures, identifiez les individus qui contribuent le plus à chaque axe (en positif et en négatif).

Etudiez les proximités et oppositions entre des groupes d'individus.

2.6 Représentation des variables (cercle des corrélations) et interprétation des résultats

Récupérez les vecteurs propres correspondant aux valeurs propres. Déduisez-en les corrélations entre les variables et les 4 premiers facteurs principaux.

Représentez les variables projetées et le cercle des corrélations sur les 2 plans factoriels considérés. A l'extrémité de chaque "trait", précisez le nom de la variable correspondante avec la fonction `text()` du module `matplotlib.pyplot`. Utilisez la fonction `Circle()` du module `matplotlib.pyplot` pour tracer le cercle.

Quelles sont les variables qui déterminent chacun des 4 axes (en positif et en négatif) ?

En croisant les résultats obtenus sur les individus et sur les variables, donnez une interprétation des différents axes.

2.7 Extension à d'autres données

Vous pouvez compléter ce travail en effectuant une ACP normée sur d'autres données, par exemple celles du fichier `decathlon_JO.txt` étudié en cours et disponible sur MyLearningSpace.