TP 1 - Calcul scientifique et ACP

Rivière Alexandre - Lacroix Yann - Mangé Valérian

Département Sciences du Numérique - Première année
2019-2020

Partie 1 : Visualiser les données

Question 1 : Dans le TP1 nous avons appliqué l'ACP sur une image. Le tableau de données X dans ce TP correspondait à la vectorisation des niveaux de Rouge Vert et Bleu de notre image (105600x3). Nos données étaient de dimension 264x400.

Question 2 : Voir code.

Partie 2 : L'Analyse en Composantes Principales

Question 3 : Sur la figure 2 on peut voir que les points rouges projetés dans la base canonique sont très regroupés. Au contraire les points bleus projetés sur les axes principaux sont quant-à eux très dispersés, on peut donc en tirer plus d'information : on dit qu'on a maximisé l'étalement et la dispersion des points, ce qui est exactement le but recherché par l'utilisation de l'ACP.

Question 4 : L'information est d'autant plus élévée que les covariances entre les q premières composantes sont basses. Ainsi la matrice Σ nous renseigne sur la qualité de l'information.

Partie 3 : L'ACP et la classification de données

Commenter la figure 1 classification.m : On peut voir grâce à l'ACP qu'on peut dénombrer deux classes au sein des points bleus. Ce qu'on ne pouvait pas voir sans l'ACP (eg les points rouges).

Question 5 : Dans le plan on peut distinguer trois classes (les points bleus et magentas sont confondus). Cependant dans l'espace on peut clairement distinguer quatre classes.

Commenter fin classification.m : Des figures 2 et 6 on peut supposer que l'information pour N classes reste élevée sur les N-1 premiers axes principaux.

Question 6 : D'abord nous commençons par tracer le pourcentage d'information contenue dans chaque axe. On se rend compte que le taux d'information chute au delà du $7^{ème}$ axe. Or, d'après l'hypothèse précèdente on pourrait penser qu'il y a 7 classes distinctes d'individus dans ce jeu de données. Nous projetons d'abord ces individus sur une droite : on peut distinguer 4 classes distinctes. Ce n'est pas en accord avec notre hypothèse, ainsi, nous projetons nos individus dans le plan, cette fois on distingue 6 classes. Ensuite, on projete nos données dans l'espace sur les trois premiers axes principaux et on voit encore 6 classes différentes. Cependant, en projetant nos données sur trois autres axes : le troisième, le quatrième et le sixième, on s'aperçoit qu'il y a 8 classes d'individus. Finalement, cela ne respecte pas notre hypothèse mais on peut conclure qu'il y a bien 8 clusters d'individus dans ce jeu de données.

Question 7 : On procède de la même manière qu'à la question précédente. On voit finalement 6 clusters. Ces derniers se voient très bien dès la figure 3.

Partie 4 : L'ACP et la méthode de la puissance itérée

Question 8 : Pour x un vecteur propre de $H^T H$, on a :

$$H^T Hx = \lambda x \Rightarrow H(H^T H)x = \lambda Hx \Rightarrow (HH^T)Hx = \lambda Hx \tag{1}$$

On a donc λ et Hx qui sont des éléments propres de HH^T.
De plus puisque $H^T H$ est de taille n, que HH^T est de taille p et que $n > p$, on aura donc la totalité des éléments propres de HH^T.

Question 9 : Voir code. (Pas de question posé dans les commentaires du code ?)

Question 10 : La fonction eig de Matlab permet de calculer tous les élements propres d'une matrice Σ. Cependant, dans le cas de l'ACP, on souhaite souvent projeter nos données sur une faible quantité d'axes, la fonction eig qu'offre Matlab consomme donc plus de temps de calcul et d'espace de stockage. Donc l'algorithme de puissance itéré semble plus utile dans le cas de l'ACP.

Question 11 : Pour minimiser le temps de calcul et la mémoire utlisée, il est préférable d'effectuer la méthode de la puissance itérée sur la matrice carrée de plus faible dimension. Cependant, en effectuant cette méthode sur la plus petite matrice, le résultat sera moins précis.