Projet : Comment comprendre et visualiser des données en très grande dimension ?

Stéphanie Allassonnière, stephanie.allassonniere@polytechnique.edu

Les données à étudier sont de plus en plus de grande dimension et / ou en grand nombre. Les exemples sont issus de domaines extraimement variés allant des sondages, de données issus de réseaux sociaux à l'analyse de données médicales (images, génôme, etc..). Plus leur dimension est grande plus il est difficile de visualiser, comprendre et interpréter les données.

Nous allons étudier ici deux modèles, l'un issu d'une étude géométrique (mais qui a une interprétation statistique très intuitive), l'autre purement statistique, qui permettent d'analyser des données de dimension quelconque. En particulier, ces deux méthodes permettent non seulement de comprendre la structure des données mais aussi d'en réduire la dimension pour voir que cette structure impose des contraintes qui font que le nuage de points ne "vit" en fait que dans un espace de dimension plus petite. Nous allons travailler sur quelques statistiques (appelé aussi facteurs) pour 49 pays du globe qui mettent en lumière les "causes" majeurs qui déterminent si un pays est développé ou en voie de développement. Vous trouverez les données sur ce site : http ://www.cmap.polytechnique.fr/~allassonniere/donneesProjetMAP311.txt.

Le premier tableau représente le taux de mortalité infantile, le nombre moyen d'habitants par médecin, la densité moyenne de population, la densité moyenne de population par surface de terres cultivables, le taux d'analphabétisme chez les plus de 15 ans, le taux d'étudiants dans le supérieur et le produit national brut.

Pour comprendre l'intérêt d'une méthode par rapport à une autre, nous travaillerons aussi sur ce jeux de données synthétiques là (https ://drive.google.com/open ?id=0BwgNRLo3udQIbHJXUVNBdWw2V3M) dont la représentation est la suivante :

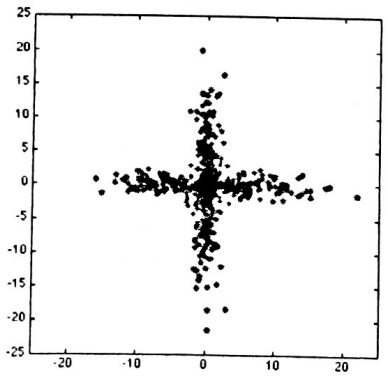

FIGURE 1 – Données synthétiques

1 Analyse en Composantes Principales (ACP)

La première idée est la plus intuitive, est d'essayer de représenter les données comme un nuage de points dans $\mathbb{R}^p$ où p est le nombre de facteurs (pour le premier jeu de données $p = 7$, pour le second $p = 2$). Comme il n'est pas facile de visualiser les données en dimension supérieure à 3, la solution est d'analyser le nuage sur le ou les meilleurs sous espaces sur lesquels la projection du nuage est la plus informative. Mais quel(s)

plan(s) choisir ?

Le principe de l'analyse en composante principale (ACP) consiste à traiter le nuage de points comme s'il était un solide (ou qu'il le représentait par échantillonnage), en isoler les axes d'inertie et se placer dans le repère propre.

1. Charger la première table de données dans la variable data. Essayer de donner des réponses intuitives aux questions soulevées dans le précédent paragraphe en regardant la matrice des données.

1.1 Description géométrique de la méthode

On note x_i la i^{eme} ligne de la matrice data pour i dans $[1, ..., n]$ (dans notre cas $n = 49$), qui correspond aux statistiques du i^{eme} pays ; c'est un élément de $\mathbb{R}^7$. On note $e_1, ..., e_7$ la base duale standard de $\mathbb{R}^7$ (en munissant $\mathbb{R}^7$ de sa structure euclidienne standard) de sorte que $e_j(x_i)$ est la j^{eme} statistique du i^{eme} pays.

2. On va effectuer l'analyse dans un repère centré au centre de gravité du nuage. Expliquez l'intérêt de ce changement de repère et explicitez l'opération matricielle qui permet d'obtenir à partir de data la matrice X qui représente les coordonnées du nuage centré.
   Ecrire ensuite l'opération qui consiste à renormaliser chaque variable de sorte qu'elle présente une variance empirique valant 1. Expliquer en quoi cela est important.

On travaillera maintenant avec la matrice X correspondant aux données centrées et renormalisée.

Si V désigne un sous-espace vectoriel de $\mathbb{R}^7$, on appelle inertie du nuage par rapport à V la quantité :

$$\mathcal{I}_V(X) = \sum_i \|p_V(x_i)\|^2 \tag{1}$$

où p_V désigne le projecteur orthogonal sur V et la norme $\|.\|$ désigne la norme euclidienne.
Cette quantité est une mesure de la dispersion du nuage projeté sur V .

3. Si u désigne un vecteur de $\mathbb{R}^7$ et D_u la droite qu'il engendre, calculer $\mathcal{I}_{D_u}(X)$. Pouvez vous identifier l'ensemble des vecteurs u pour lesquels cette quantité est maximale et discuter de son unicité ?

Cette maximisation répond au problème suivant : si on veut résumer le nuage par projection sur une droite et ramener ainsi chaque point qui le compose à un scalaire, quelle est la droite qui donne la projection la plus dispersée possible (où encore celle qui permet le meilleur discernement entre les points) ?

On repose la même question parmi les droites dans l'espace orthogonal de la droite trouvée à la question précédente puis dans l'orthogonal du plan formé par les deux droites, etc.
Les droites ainsi trouvées sont appelées composantes principales.

1.2 Résultats numériques

4. Représenter sur plusieurs figures les projections du nuage sur un plan extrait des coordonnées canoniques. Que remarquez vous en terme de dispersion du nuage.
   Faire la même chose sur des espaces tridimentionnels.
5. Calculer les composantes principales (CP) et les inerties associées au nuage représentant les données X. Classer les CP associées à une dispersion de la plus grande à la plus petite. Commenter.
6. Représenter la projection orthogonale du nuage dans le plan formé par les deux premières composantes principales ; Puis en 3D sur l'espace engendré par les trois premières CP. Interpréter les résultats : que discrimine le premier axe ? Cela vous paraît-il logique ? (rappelons qu'aucune connaissance a piori n'a été ajoutée dans l'algorithme !) Même question pour le second axe. Commentez les corrélations entre facteurs ainsi que les points isolés du nuage projeté.

1.3 Modèle statistique sous-jacent

Sous-jacent à cette analyse géométrique, nous pouvons proposer un modèle statistique. C'est le modèle le plus simple qui met en jeu une loi facilement décrite par peu de paramètre : la loi normale multidimentionnelle.

Soit X_i le vecteur de facteurs dans $\mathbb{R}^7$ pour le pays i. On suppose que les pays sont indépendants et de même loi, alors on pose :

$$X_i \sim \mathcal{N}(0, \Sigma) \tag{2}$$

où Σ est une matrice pleine de dimension 7×7.

7. Expliquer en quoi l'estimation de la matrice de covariance Σ permet l'estimation des composantes indépendantes de la matrice de données. Proposer un estimateur de cette matrice, c'est à dire une quantité ne dépendant que de la matrice X qui serait une bonne approximation de la matrice Σ.
8. Utiliser une fonction python permettant d'exhiber rapidement les composantes indépendantes des données. Comparer avec vos précédentes expériences.

2 Analyse en Composantes Indépendantes (ACI)

On va maintenant changer de modèle d'interprétation du nuage en utilisant un modèle plus complexe mais avec plus de degrés de liberté. Il s'agit de mimer la méthode de génération des données en supposant qu'elles proviennes de sources dont les effets sont mélangés avec des coefficients indépendants. Remarquez qu'il est évident que les données ne suive aucune distribution statistique identifiable mais on cherche à approcher cette distribution par des lois simples.

Dans le cas qui nous intéresse, on propose donc un modèle très simple appelé Analyse en Composante Indépendantes (ACI) comme suit :

Chaque vecteur d'observation X_i dans $\mathbb{R}^p$ est supposé être la combinaison linéaire de d variables aléatoire indépendantes $(Y_i^j)_{1 \leq j \leq d}$ et de même loi :

$$X_i^k = \sum_{j=1}^{d} a_{k,j} Y_i^j \,, \tag{3}$$

où les poids $a_{k,j}$ sont déterministes.

On peut réécrire matriciellement cette égalité en considérant la matrice $A = (a_{i,j})$ et le vecteur colonne $Y = (Y^1, ..., Y^d)$:

$$X_i = AY_i \tag{4}$$

pour chaque individu (ou pays) i.

9. Sous quelles conditions sur p, d et A a-t-on une solution à ce problème inverse ?

Il reste une chose à faire pour que le modèle soit complet : il faut donner la loi des variables aléatoires indépendantes Y_i^j. L'une des hypothèse les plus courante est de choisir la loi logistique définie par sa fonction de répartition : pour tout $t \in \mathbb{R}$,

$$P(Y_i^j \leq t) = \frac{1}{1 + exp(-2t)}. \tag{5}$$

pour tout $1 \leq i \leq n$ et $1 \leq j \leq d$.

10a. Proposer une méthode de simulation selon la densité logistique. Tracer sur une même courbe la densité ainsi qu'un histogramme de la distribution donnée par 1000 échantillons.

10b. Choisissez une matrice de votre choix en petite dimension (2 ou 3) puis générez un échantillon de points selon le modèle proposé.

La matrice $W = A^{-1}$ is appelée matrice de décomposition : appliquée aux données initiales, elle fournit un "filtrage" des données en composantes indérendantes.

Etant donné ce modèle, nous allons estimer la matrice A ou son inverse W. Pour cela nous allons calculer les paramètres les plus vraisemblables. On appelle vraisemblance de A au vu des observations $(X_1, ..., X_n)$ d'un n-échantillon indépendamment et identiquement distribué selon la loi de densité $f_A(X)$, le nombre :

$$L(X_1, ..., X_n; A) = f_A(X_1) \times ... \times f_A(X_n) = \prod_{i=1}^{n} f_A(X_i)\,. \tag{6}$$

Pour estimer la valeur la plus vraisemblable de A, on maximise la quantité précédente (Il est plus simple de maximiser en W, nous allons dans la suite se placer dans ce cas et tout expliciter en W).

11. Que pouvez vous dire du point qui maximise L par rapport à celui qui maximise $\log L$? Justifier.
12. Soit ψ la fonction de répartition de la distribution logistique. Calculez $L(X_1, ..., X_n; A)$. Montrez que la matrice W doit maximiser :

$$E(W) = \log|det(W)| + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d} \log(\psi'[(WX_i)_j])\,, \tag{7}$$

où $(WX_i)_j = \sum_{m=1}^{d} W_{j,m}X_{i,m}$.

13. Comment trouver la matrice W optimale ?
14. Montrer alors que le gradient de cette expression est :

$$\langle \nabla_W E(W), H\rangle = trace(HW^{-1}) + \frac{1}{n}\sum_{i=1}^{n}\sum_{j,m=1}^{d} \chi'((WX_i)_j)H_{j,m}X_{i,m} \tag{8}$$

avec $\chi = \frac{\psi''}{\psi'}$ (On verifiera que pour la loi logistique $\chi(t) = \frac{2}{1+exp(-2t)}$.)
On pourra calculer $f'(0)$ où $f(\varepsilon) = E(W + \varepsilon H)$ pour toute matrice inversible H. (On pourra accepter l'égalité suivante : Soit $g(B) = \log|det(B)|$ alors la différentielle de g en B appliquée à H est $d_H g(B) = trace(HB^{-1})$).

Il est cependant mieux d'utiliser une définition du gradient plus adaptée à la recherche de matrices inversibles. On définit alors :

$$f'(0) = \langle HW^{-1}; \nabla_W E(W)W^{-1}\rangle \tag{9}$$

15. En déduire que

$$\nabla_W E(W) = W + \Gamma(W)W'W \tag{10}$$

où W' est la matrice transposée de W et $\Gamma(W)$ est la matrice telle que $\Gamma(W)_{j,m} = \frac{1}{n}\chi'((WX_i)_j)X_{i,m}$.

16. Proposer une méthode de minimisation de votre choix (utiliser une commande existante en python ou développer votre propre algorithme type descente de gradient avec le gradient ci-dessus).
17. On travaille maintenant sur la seconde matrice du fichier de données pour des raisons de visualisation. Dessiner sur une figure le nuage de points en 2 dimensions (penser à recentrer le nuage).
Tracer les deux droites engendrées par les vecteurs colonnes de la matrice A obtenue à partir des données précédentes sur le même graphe : $\tilde{p}_j(Y_i) = A_{i,j}Y_i^j$.
18. Tracer maintenant sur le même graphe les deux directions principales obtenues avec votre algorithme de la partie 1.
19. Commenter les différences, avantages et inconvénients de chacune des représentations. En particulier, pour mieux comprendre la différence entre les deux modèles, représenter les deux directions principales et les deux sources de l'ACI du nuage de la figure 1.
20. Reprenez les données mondiales de la partie 1 et comparer les axes de l'ACP et ceux de l'ACI.