UNIVERSITE DE LORRAINE
Ecole Nationale Supérieure d'Electricité et de Mécanique

Introduction à l'Analyse en Composantes Principales

Didier Maquin – Didier.Maquin@univ-lorraine.fr

On se propose, pour cette séance de TD, de manipuler, sur des données simulées que l'on maîtrise complètement, les concepts élémentaires d'analyse en composantes principales introduits lors des deux séances de cours précédentes.

1 Génération des données (1er cas)

La fonction Matlab© suivante permet de générer un signal constitué d'une succession de créneaux d'amplitude et de durée aléatoires (amplitude comprise entre 0 et 10 et durée comprise entre 0 et 50). Le nombre de valeurs générées est égal à la dimension du vecteur `t` passé en argument.

```
function u = com(t);
u = [];
while length(u) < length(t)
   for i=1:5+rand*5
       u = [ u ; rand*10*ones(fix(rand*50),1)];
   end
end
u = u(:);
u = u(1:length(t));
```

On se propose de générer une matrice de données X comportant 500 observations de 7 variables définies de la manière suivante :

$$\begin{aligned} x_1 &= u_1 + \varepsilon_1 \\ x_2 &= u_1 + \varepsilon_2 \\ x_3 &= u_1 + \varepsilon_3 \\ x_4 &= u_2 + \varepsilon_4 \end{aligned} \qquad \begin{aligned} x_5 &= u_2 + \varepsilon_5 \\ x_6 &= 3u_1 + 2u_2 + \varepsilon_6 \\ x_7 &= 2u_1 + u_2 + \varepsilon_7 \end{aligned}$$

où u_1 et u_2 sont des signaux aléatoires générés à l'aide de la fonction `com` précédente et ε_i sont des bruits aléatoires gaussiens d'espérance nulle (fonction Matlab© `randn`). Leurs variances respectives seront choisies de façon à ne pas "noyer" complètement les signaux dans le bruit.

2 Analyse en composantes principales

2.1 ACP normée

On se propose d'effectuer une analyse en composantes principales normée de la matrice de données X.

- Centrer et réduire la matrice X (fonction Matlab© `mean` et `std`). Soit X_c la matrice X centrée réduite, vérifier dans ce cas que la matrice $X_c^T X_c$ correspond bien à la matrice des corrélations entre les variables (fonction Matlab© `corrcoef`).

- Calculer et tracer les composantes principales en ayant préalablement déterminé les vecteurs propres et valeurs propres de la matrice $X_c^T X_c$ (fonction Matlab© `eig`).
- Après avoir tracé la courbe de décroissance des valeurs propres, choisir le nombre d'axes qui vous semble pertinent pour représenter correctement l'ensemble des données.
- Compte tenu de la manière dont les données ont été engendrées, le choix précédent est-il facilement interprétable ?
- Reprendre le calcul des composantes principales en utilisant cette fois une décomposition en valeurs singulières de la matrice X_c (fonction Matlab© `svd`).
- Calculer et visualiser une estimation $\hat{X}$ de la matrice de données initiale en utilisant le nombre de composantes principales retenu précédemment.
  Rappel : si $X = VSU^T = \begin{pmatrix} V_1 & V_2 \end{pmatrix} \begin{pmatrix} S_1 & 0 \\ 0 & S_2 \end{pmatrix} \begin{pmatrix} U_1^T \\ U_2^T \end{pmatrix} = V_1 S_1 U_1^T + V_2 S_2 U_2^T$, $\hat{X} = V_1 S_1 U_1^T$.
- Evaluer le taux de réduction de l'information engendré par cette "compression".

3 Génération des données ($2^{ème}$ cas)

On se propose maintenant de générer une matrice de données incluant des dépendances non linéaires entre variables. La matrice X est alors constituée de 7 variables ainsi définies (on a supprimé les redondances directes) :

$$\begin{aligned} x_1 &= u_1 + \varepsilon_1 \\ x_2 &= u_2 + \varepsilon_2 \\ x_3 &= u_2^2 + \varepsilon_3 \\ x_4 &= u_1 + 6u_2 + \varepsilon_4 \end{aligned} \qquad \begin{aligned} x_5 &= 5u_1^2 + 3u_2^2 + \varepsilon_5 \\ x_6 &= 2u_1 + 3u_2 + \varepsilon_6 \\ x_7 &= u_1^2 + u_2 + \varepsilon_7 \end{aligned}$$

où les variables u_1, u_2 et ε_i ont la même signification que dans le cas précédent. Dans un premier temps, afin de n'engendrer que des non-linéarités "faibles", la plage d'excursion des deux commandes u_1 et u_2 sera limitée à l'intervalle $[0 \;\; 0.25]$. On fixera ensuite cette plage à $[-0.25 \;\; 0.25]$.

4 ACP dans un cas non linéaire

Comme précédemment, effectuer une ACP normée de la matrice de données.

- Reprendre les questions de la section 2.
- Interpréter les résultats obtenus. En particulier, discuter le nombre de composantes principales retenues pour la reconstruction.
- Observez-vous des différences dans la qualité de l'estimation en fonction de la plage d'excursion des variables u_1 et u_2 ? Donnez en une interprétation.

5 Pour aller plus loin...

Considérer de nouveau les données du premier cas (cas linéaire). Elaborer le modèle ACP sur les 400 premières observations. Ajouter un défaut sur l'une des variables (par exemple, remplacer la variable x_1 sur l'intervalle d'observation $[440 \;\; 470]$ par $1.2x_1$; il s'agit là d'un défaut multiplicatif qui peut correspondre à un problème de gain de capteur). Calculer, à partir du modèle ACP précédent et sur les 500 observations, les différentes reconstructions $\hat{z}_i$ des variables en supprimant leur mesure respectives x_i et analyser les écarts entre ces reconstructions et les mesures pour localiser le défaut.