Introduction au data mining
L3 MIS - STA 1616 - 2010
V. Monbet
Préparation des données

L'objectif de ce TD est de montrer les possibilités offertes par des choix de métriques particuliers pour la détection de points atypiques multivariés par une ACP et comment une ACP curvilinéaire apparait, dans certains cas, comme une version robuste de l'ACP.

1 Points atypiques

Lire les données décrivant les villes de province de plus de 100000 habitants :

```
data province ;
infile 'h:/.../province.dat' dlm='09'x ;
input ville $ dep $ crent bac chom salr indus serv impot crim
        jeun cine sport pollu ;
run ;
```

Exécuter le programme ciŋdessous en tentant de comprendre la démarche suivie. L'essentiel des difficultés vient de ce que la procédure `princomp` ne connait que la métrique euclidienne usuelle dans l'espace des individus et, en dehors du module `SAS/IML`, les calculs matriciels (inversion, produit) ne sont pas immédiats. Le programme calcule successivement la métrique de Mahalanobis, les pondérations des individus par une fonction décroissante (avec choix de k) de la norme des individus au sens de cette métrique puis l'estimation robuste de la matrice de covariance à l'aide de ces pondérations. Enfin, les données sont transformées de sorte que la macro usuelle d'ACP calcule l'ACP recherchée, c'estŋàŋdire avec la bonne métrique. Naturellement ce programme pourrait être inclus dans une macro mais cela occulterait largement la structure des calculs !

```
/* paramètres*/
%global dataset ident listev;
%let dataset = province;
%let listev = crent--pollu;
%let p=12;
%let k=0.5;
```

```
%let ident= ville;
/* sélection des variables */
data donnees;
set sasuser.&dataset (keep=&ident &listev);
run;
/* Centrage et réduction des données */
proc standard data=donnees out=un mean=0 std=1 vardef=n;
var &listev ;
run;
/* Calcul des pondérations des individus */
/* calcul de V*LamdaŞ(ŋ1/2)*/
%vlambda(un,vpmsrl);
/* calcul de X*V*LamdaŞ(ŋ1/2) */
proc score nostd out=xmvpsrl data=un score=vpmsrl;
var &listev;
run;

 /* Calcul des poids*/
data deux;
set xmvpsrl;
poids=exp(ŋ&k*sqrt(uss(of scor1ŋscor&p)));
run;
/* calcul de l'estimation robuste Psi de la covariance */
proc corr cov outp=trois vardef=wgt data=deux noprint;
var &listev;
weight poids;
run;
/* Calcul de l'acp de (X,Psi^ŋ1,1/nI) */
/* par celle de (XPsi^(ŋ1/2),I,1/nI) */
/* calcul de PsiŞ(ŋ1/2) */
%vlambda(trois,psimsrl);
/* calcul de X*Psi^(ŋ1/2) */
proc score nostd out=sasuser.xtrans data=un score=psimsrl;
var &listev;
id &ident;
run;
%acp(xtrans,&ident,scor1ŋscor&p,red=cov);
%gacpix;
```

Lister les points atypiques apparus et interpréter les résultats obtenus. Il serait possible de tracer un graphe des variables mais cela reste à faire...

2 Application aux données des départements

Adapter le programme précédent, et donc les macros variables apparaîssant sous le commentaire `\* Paramètres *\`, afin d'appliquer ce programme, pour différentes valeurs de k (entre 0.5 et 1), aux données déjà étudiées : `depart`. Quels sont les départements atypiques apparaîssant sur les premiers axes ?

3 Mise en oeuvre sous R

Aidez vous du programme écrit en SAS pour développer un outil équivalent sous R. Appliquer les commandes R sur les mêmes jeux de données qu'en SAS et comparer les résultats.