Rapport de tp 1
Analyse en composantes principales

Paul EZVAN - Omar GIVERNAUD

9 octobre 2009

Table des matières

1 Introduction

Ce TP met en application les algorithmes de régression linéaire et d'analyse en composantes principales sur des données réelles. Le but est de comparer le taux d'emplois de différentes villes et donc de normaliser les données pour que la comparaison aie un sens.

2 Statistique descriptive

2.1 Fonctions de bases

```
load Emploi_eur
moy = mean(X)
ecart = std(X)
mediane = median(X)
min=min(X)
max=max(X)
```

X1 : Taux d'emploi

X2 : Temps de travail hebdomadaire (heures)

X3 : Taux de chômage

X4 : Taux d'emploi des 55-64 ans

X5 : Taux de chômage des 15-24ans

X6 : Productivite horaire (100=UE27)

```
>> moy = mean(X)
moy =
   66.4280   41.5680    6.2840   45.8520   14.3360   96.8760

>> ecart = std(X)

ecart =
    5.8507    1.1423    2.0307   11.4814    5.1064   35.0442

>> mediane = median(X)

mediane =
   67.8000   41.2000    6.1000   46.0000   14.3000   98.5000
```

```
>> min=min(X)

min =
    55.7000    39.9000     3.2000    28.3000     5.9000    47.1000

>> max=max(X)

max =
    77.1000    44.3000    11.1000    70.0000    22.9000   195.4000
```

A cette étape de l'analyse, il est difficile de comparer les pays entre eux. Par exemple nous aurions pu remarquer que le Portugal et la Slovénie avaient des résultats similaires pour les deux premiers paramètres, cependant différents pour les autres valeurs. Les différentes données sont réparties sur des plages et échelles différentes. Une représentation sous la forme d'un diagramme est une première solution facilitant la comparaison des individus.

2.2 Profile en étoile

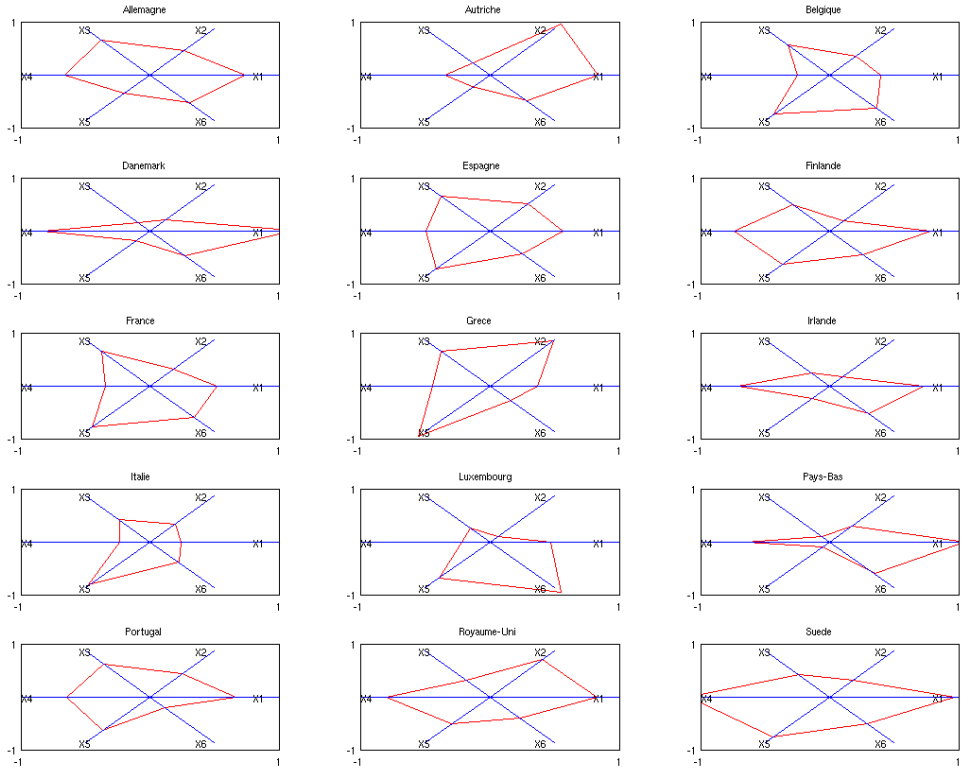

FIGURE 1 – Diagramme en étoile

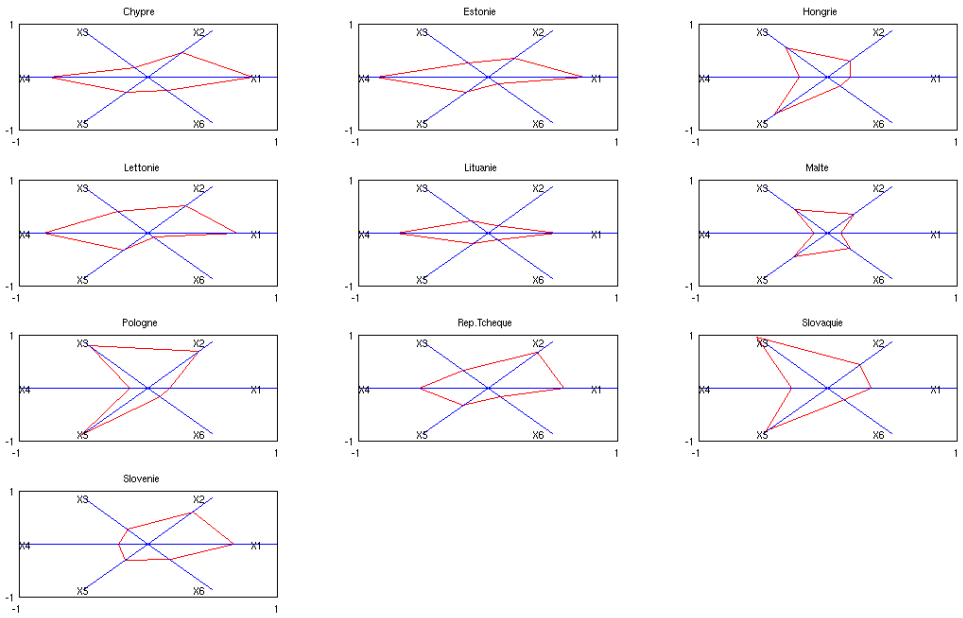

FIGURE 2 – Diagramme en étoile

Il apparaît ici que les deux pays que nous avions prédits comme « similaires » à la vue des données présentent des diagrammes en étoiles assez différents. Le problèmes de cette méthode est qu'elle repose sur une analyse subjective de l'opérateur : les diagramme se ressemblent-ils ? De plus l'opérateur doit comparer deux à deux les graphiques ou avoir une bonne mémoire, ce qui pour un grand nombre de données rend la méthode encore moins fiable.

2.3 Régression linéaire

La régression linéaire a pour but de montrer un lien de proportionnalité entre deux variables. Nous comparons le taux d'emploi global au taux d'emploi des personnes âgées de 55 à 64 ans.

Nous obtenons les résultats suivants :

```
>> [a, b , r] = agressionLineaire(X(:,1), X(:,4))
a =
    1.5613

b =
   -57.8603

r =
    0.7956
```

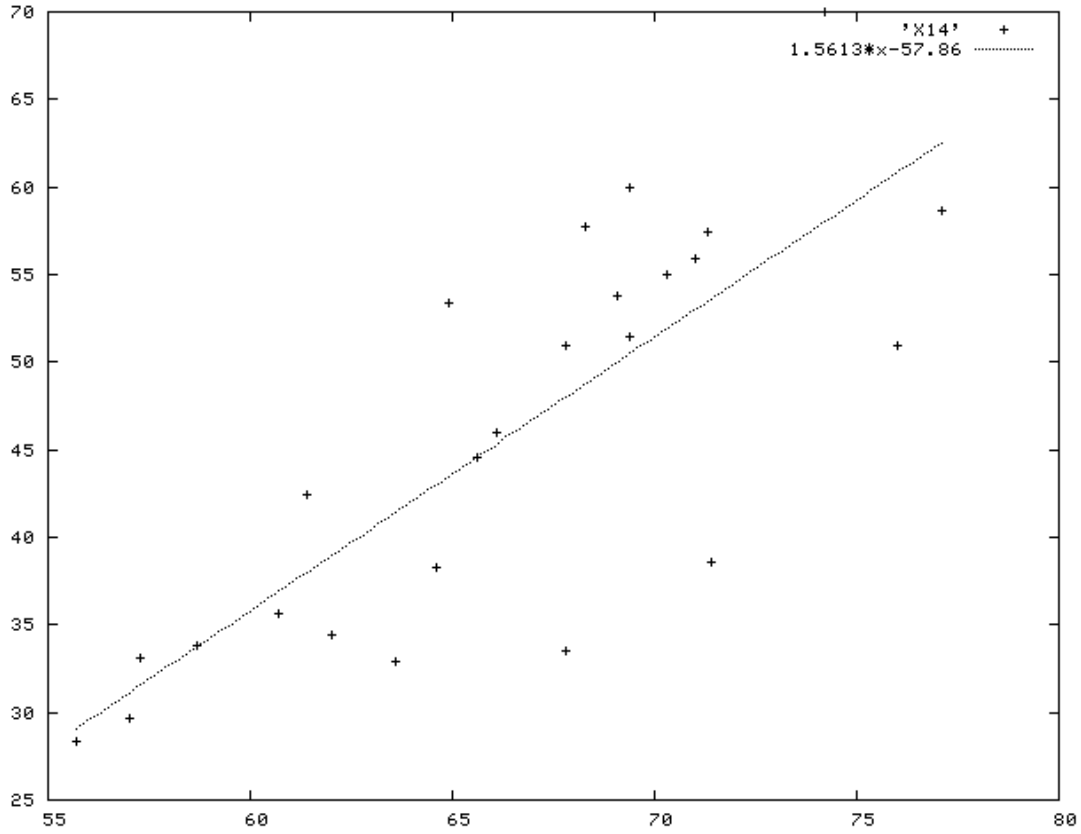

FIGURE 3 – Régression linéaire de X4 par rapport à X1

Nous obtenons un taux de corrélation de 0.8 ce qui montre une forte corrélation des deux paramètres. Encore une fois, ramener les données dans un ensemble de dimension moindre facilite la comparaison. Graphiquement, il nous était difficile de dire si les données étaient corrélées.

Une études graphique des résidu pourrait être intéressante.

```
%regressionLineaire(X,Y)
%éralise la érgression élinaire de Y
%par rapport à X
%a et b coefficients de la érgression
%r coefficient de écorrlation

function [ a, b, r] = regressionLineaire( X, Y )
    Sxx = std(X)*std(X);
    Syy = std(Y)*std(Y);
    rc    = corrcoef([X,Y]);
    r = rc(1,2);
    Sxy = r*sqrt(Sxx*Syy);
    a = Sxy/Sxx;
    b = mean(Y) - (Sxy/Sxx)*mean(X);
end
```

Nous avons choisi d'utiliser la fonction `corrcoef` qui calcule le coefficient de corrélation des variables afin d'obtenir la valeur S_{xy}. Cette méthode étant plus rapide puisqu'elle utilise une fonction compilée (et optimisée) du logiciel.

3 Analyse en composantes principales

3.1 Normalisation des données

Comme nous l'avions fait remarquer précédemment, comparer des données qui ne sont pas sur les même échelles peut être problématique. Nous allons donc les normaliser :

$$x'_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}$$

Ce qui a pour effet de centrer le nuage de données.

```
for i=1:size(X,2)
    Y(:,i)=(X(:,i)-mean(X(:,i)))/std(X(:,i),1);
end
```

3.2 Boîte à moustaches

Afin d'étudier l'impact de la normalisation des données sur leur répartition, nous avons tracé les boîtes à moustaches dans les deux cas (normalisé ou non).

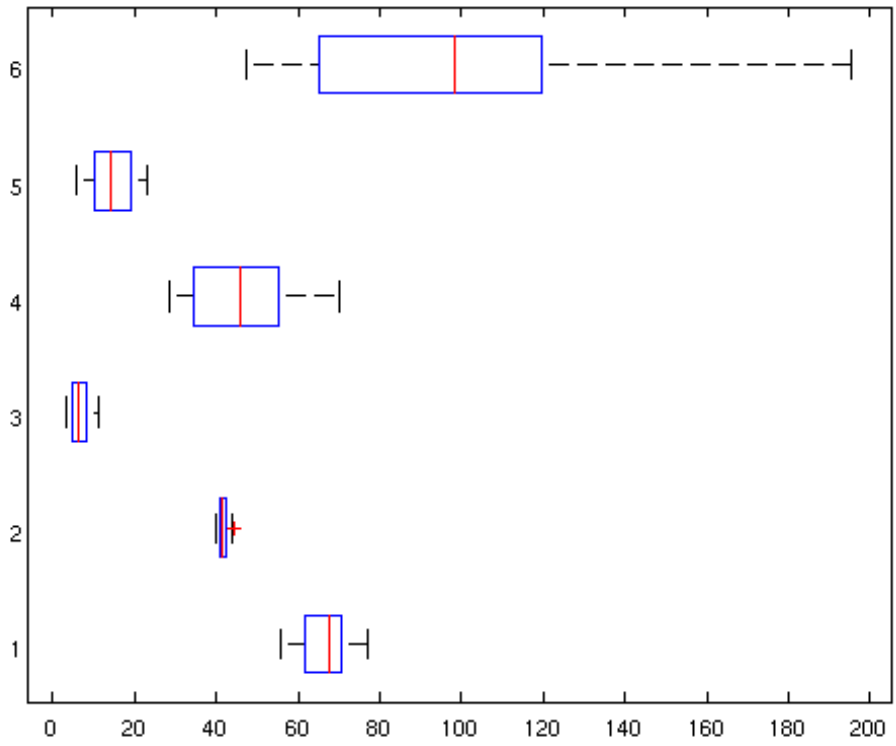

FIGURE 4 – Boîte à moustache (données non normalisée

Nous remarquons des écarts importants entre les différentes informations que nous donnent les boîtes à moustache (médiane, quartiles, min et max). Il serait alors incohérent de comparer les données entre elles. Néanmoins, ce digramme nous apporte une information quant à la densité des données. Pour le caractère 6, nous observons que les données sont étalée ce qui n'est pas le cas pour le caractère 2.

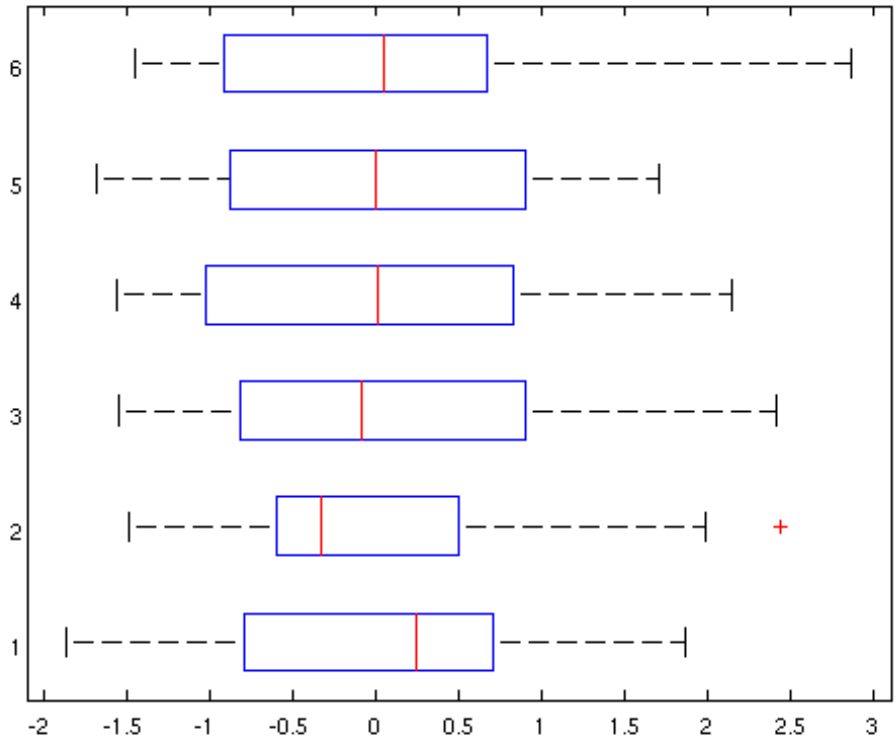

FIGURE 5 – Boîte à moustache (données normalisée)

L'information que nous la boîte à moustache normalisée est la répartition des données par rapport aux quartiles. Plus ils sont proportionnellement petit par rapport au reste des données, plus la densités est importante. Par exemple, nous remarquons que pour le caractère 2 le deuxième quartile est relativement court. Nous en déduisons une densité importante des données dans cette zone.

3.3 Calcul des axes principaux du nuage

Calcul des valeurs propres triées :

```
R=corrcoef(X)
[V, D] = eig(R)
[L, I]=sort(diag(D))
V(:,I)
```

Matrice de corrélation :

```
R =
   1.000000  -0.065637  -0.585620   0.795594  -0.619314   0.243957
  -0.065637   1.000000   0.205262  -0.184615   0.115429  -0.308364
  -0.585620   0.205262   1.000000  -0.395584   0.777352  -0.177344
   0.795594  -0.184615  -0.395584   1.000000  -0.416836  -0.093358
  -0.619314   0.115429   0.777352  -0.416836   1.000000   0.057758
   0.243957  -0.308364  -0.177344  -0.093358   0.057758   1.000000
```

Valeurs propres (dans la version utilisée elles étaient ordonnées) :

```
L =
   0.055486
   0.206534
   0.772965
   0.811631
   1.287825
   2.865559
```

Vecteurs propres :

```
V =
  -0.676691   0.234221  -0.398463   0.216225  -0.054277   0.527975
   0.182663  -0.099743  -0.672266  -0.311365  -0.616930  -0.164903
   0.097391   0.707599  -0.138123   0.475628  -0.043439  -0.492577
   0.532571  -0.221733  -0.059167   0.651877  -0.191190   0.449672
  -0.331056  -0.617208  -0.239342   0.427385   0.177299  -0.487933
   0.325589   0.066216  -0.556257  -0.149919   0.739308   0.105516
```

Le fait de projeter des données dans un ensemble plus restreint engendre fatalement une perte de données. Il est donc nécessaire de vérifier que la distorsion lié à ce changement n'est pas trop importante. Pour cela il est possible de calculer la fidélité du nuage qui mesure le rapport entre le nuage originale et celui projeté.

Calcul de la fidélité de la projection sur deux et trois axes :

```
f2 = 1/size(L,1)*(L(6,1) +L(5,1))
f3 = 1/size(L,1)*(L(6,1) +L(5,1)+L(4,1))

f2 =  0.69223
f3 =  0.82750
```

La fidélite de la projection sur deux axes est satisfaisante, mais celle sur trois axes permettrait un gain d'information certain.

3.4 Cercle des corrélations et projection

```
% Matrice des correlations
C(:,1)=sqrt(D(d))*V(:,d);
C(:,2)=sqrt(D(d-1))*V(:,d-1);
figure
hold on
% Creation du cercle
circle([0,0],1,1000,'.');
% Representation de la matrice dans le cercle
plot(C(:,1),C(:,2),'+')
text(C(:,1)+0.01,C(:,2),q)
hold off
```

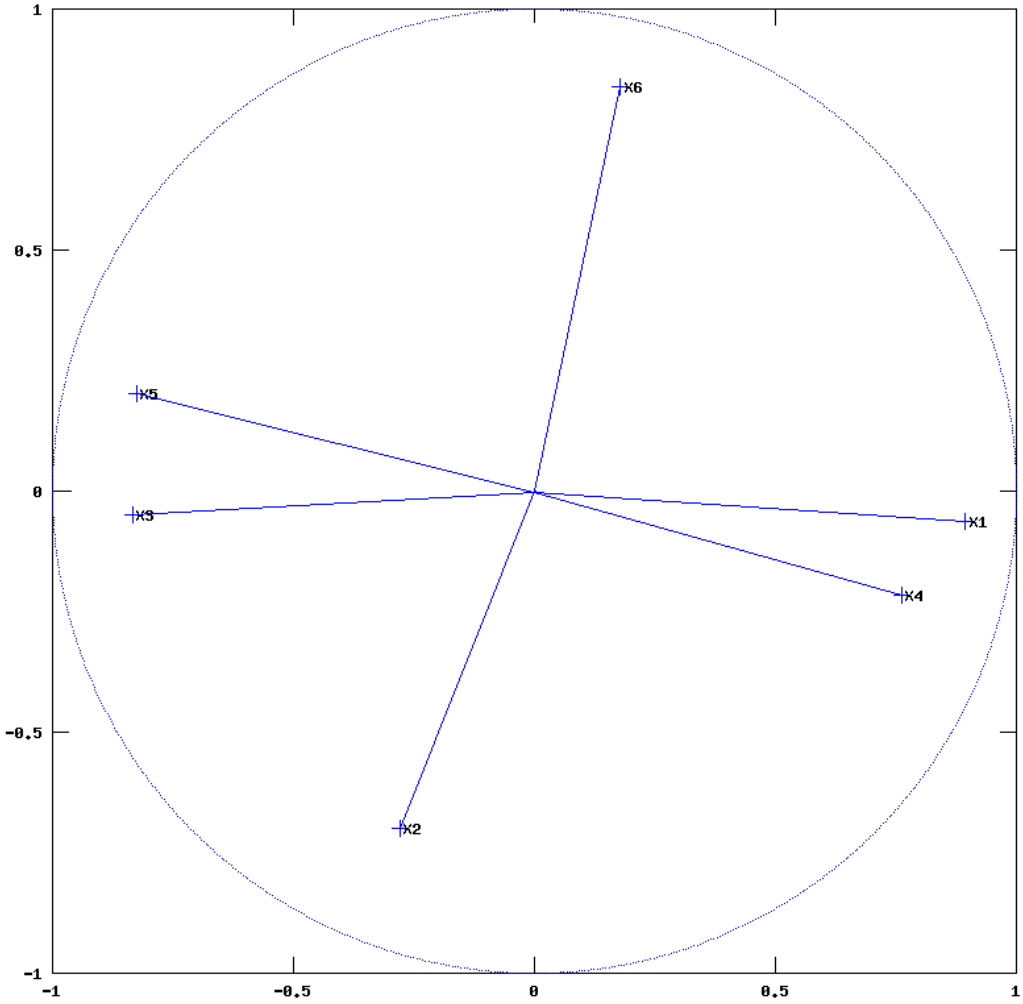

FIGURE 6 – Cercle de corrélation

Les paramètres X3 et X5 sont opposés aux paramètres X4 et X1 sur l'axe principal, cet axe oppose donc les pays au taux d'emploi élevé au pays ayant un taux de chômage important. Les paramètres X2 et X6 sont opposés sur l'axe secondaire, ce qui sépare les pays ayant une forte productivité horaire aux pays ayant un temps de travail hebdomadaire élevé.

```
% Matrice de projection en dimension 2
A=[V(:,size(X,2)),V(:,size(X,2)-1)]
% Projection
P=X*A
figure
```

```
hold on
plot(P(:,1),P(:,2),'x')
text(P(:,1),P(:,2)+2,q)
hold off
```

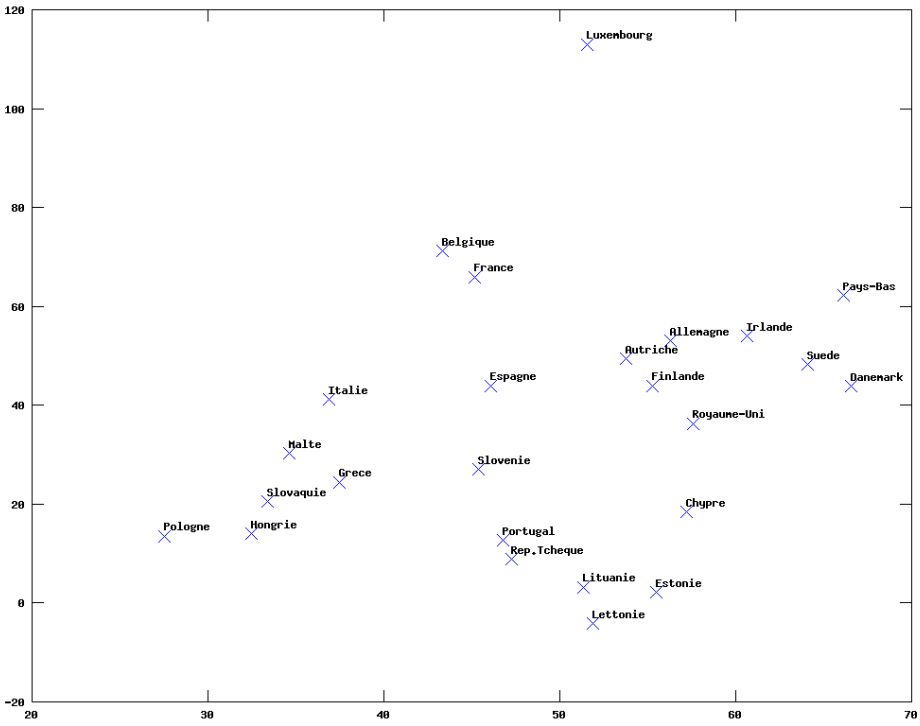

FIGURE 7 – Projection des données

La projection sur le plan principal nous permet visualiser de les caractères du pays selon les axes précédents. Nous notons que le Luxembourg est à part avec une forte productivité horaire et un faible temps de travail hebdomadaire. Les pays de l'est de l'Europe ont temps de travail hebdomadaire important avec une productivité horaire plus faible. Les pays présentant un taux d'emploi élevé sont les pays nordiques : Suède, Danemark, Finlande, Pays-Bas. La France et la Belgique ont un profil similaire, ainsi que les pays baltes.

La première comparaison que nous avions faite entre le Portugal et la Slovénie

4 Conclusion

L'analyse descriptive nous permet de précisément comparer deux pays entre eux mais rend difficile une comparaison globale. L'analyse en composantes principales nous permet de représenter ces caractères plus ou moins corrélés entre eux selon des axes non corrélés afin de minimiser la perte d'information liée d'un passage de six critères à deux. Celle-ci nous a donc permis de compresser les données afin de pouvoir les représenter en deux dimensions en minimisant la perte d'information. De cette représentation nous avons pu déduire des observations sur les caractéristiques de l'emploi des pays selon leur position géographique.