Projet Data Mining

Naoufali MADI

Darlène NIBOGORA

Ariinui TERIITEHAU

Master 1 EKAP

Le but de cette étude est de réaliser une analyse discriminante sur des données de satisfaction clientèle par de deux manières, PLS 2 et PLS 1 (Partial Least Square), c'est-à-dire la régression des moindres carrés partiels. L'agorithme emprunte la même démarche que l'analyse en composantes principales (ACP). PLS2 est une version de régression cherchant à expliquer, modéliser un exemple de q variables Y par un ensemble de p variables explicatives X. En revanche, PLS1 cible qu'une variable Y par p variables explicatives X.

Dans un premier temps, on va réaliser l'étude avec la PLS2 sur un jeu de données de Marketing. On a les variables Y qui représentent l'Usage et la Satisfaction et on a le bloc X : Price2, Price1, Speed, Service, Image2, Image1 et Quality. Puis une PLS1 sur chacune des variables à expliquer

PLS2

Tout d'abord, on commence à réaliser l'ACP avec les deux variables à expliquer (Usage et Satisf) afin de comparer les résultats sur le choix des composantes avec PLS2. Dans le graphique (la figure 1), ils sont représentés les valeurs propres de chaque composante, il nous permet de voir la variance de chaque composante. La 1er composante représente la valeur d'inertie la plus élevée parmi les valeurs propres, elle explique à elle seul 36.08% de la dispersion totale. Tandis que la seconde composante représente 30.32% de la dispersion. Tout en se basant, sur ces deux composantes, il est préférable de faire une analyse dans un nuage de dimension 2 car les deux composantes sont capables d'expliquer 66.4% de la dispersion totale du nuage.

Dans la figure 2, on représente toutes les variables X et Y, ainsi les variables Satis et Usage sont colorées en bleue afin de les distinguer et on remarque que les deux variables se trouvent au-dessous de l'axe 1 et à gauche de l'axe 2. Alors cela signifie que la Satis et Usage va augmenter tout en se déplaçant vers le haut de l'axe 2 et on remarque aussi que les deux variables sont très proches de l'axe 2, elles contribuent fortement à l'axe 2. En comparant entre l'axe 1 et 3, les deux variables contribuent très faiblement (la figure 1.A, annexe).

Figure 1

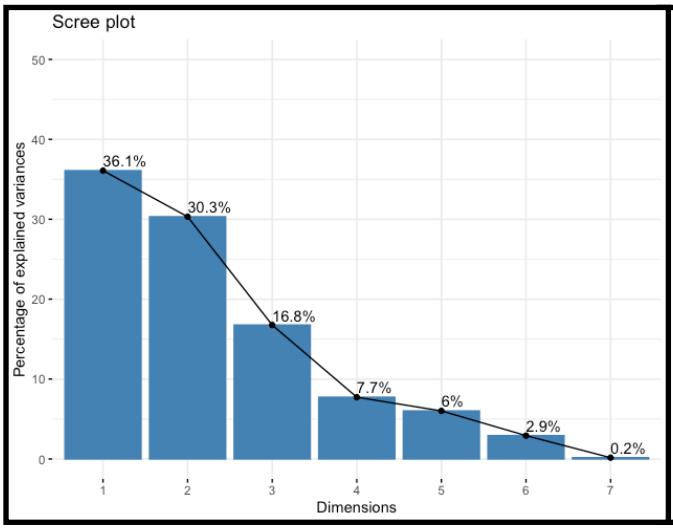
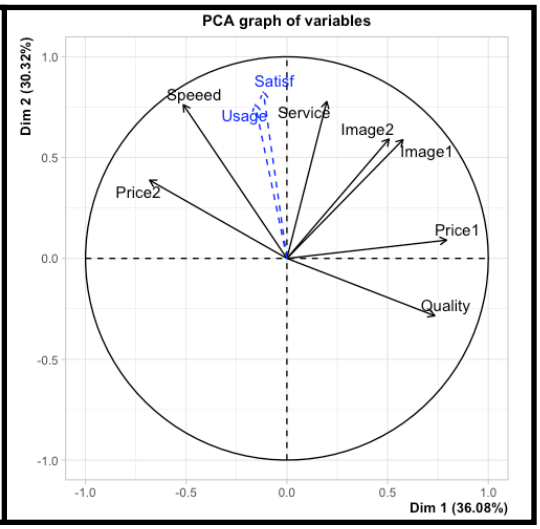


Figure 2



PLS2

L'idée principale de l'algorithme PLS2 est de rechercher des composantes maximisant le critère de covariance $cov(X, Y)$ sous des contraintes de norme et d'orthogonalité entre les composantes. La figure 3 montre les résultats de la PLS2, on choisit le nombre de composantes par validation croisée. La composante est retenue si $[Q^2_{cumule}]_h$ est nettement supérieur à $[Q^2_{cumule}]_{h-1}$ puis elle est acceptable si $[Q^2_{cumule}]_h > 0.5$. On la ratio Q qui se calcule comme $Q^2 = 1 - \frac{PRESS_h}{RESS_{h-1}} \geq 0.05$ avec

$PRESS_h = \sum_i (y_{(h-1),i} - \hat{y}_{(h-1),-i})^2$ la prédiction de de la somme des résidus au carrés

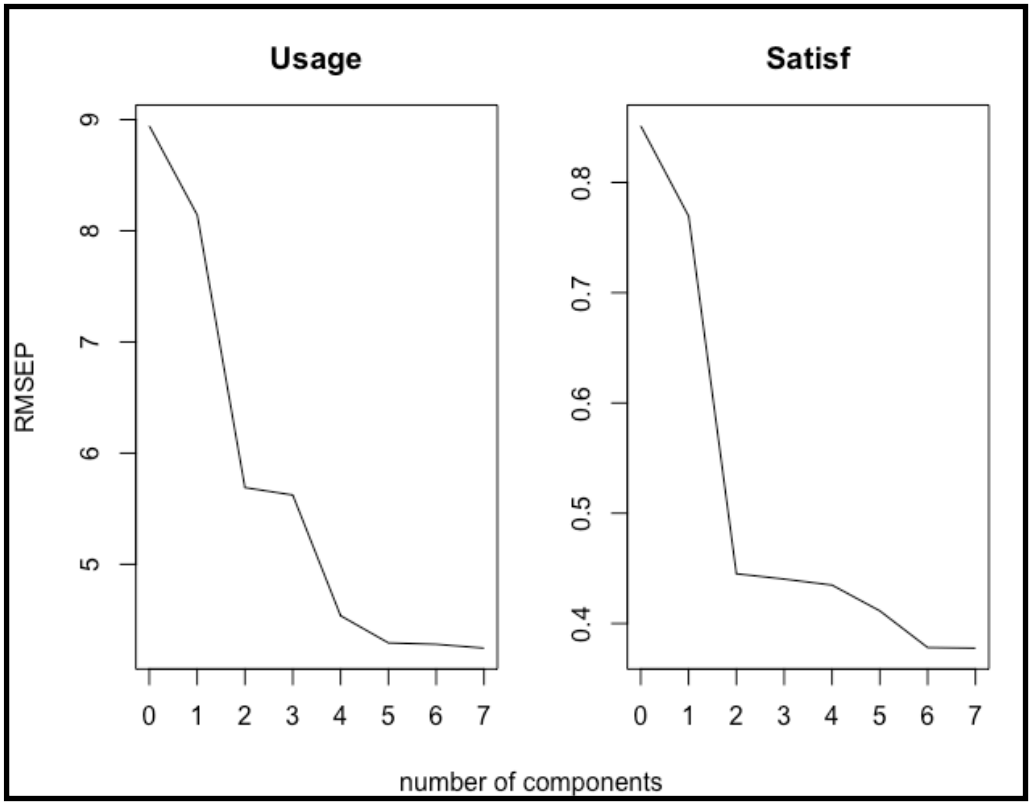
et $RESS_h = \sum_i (y_{(h-1),i} - \hat{y}_{(h-1),-i})^2$, la somme des résidus au carrés.

La figure 4 montre clairement que le choix de nombre de composante se porte sur 6 car l'erreur des résidus se stabiliseront à 5 ou 6 composante avec le critère de RMSEP (qualité de prédiction). De plus, les 6 composantes pour les deux variables expliquent à plus de 50% pour que le ratio cumulé soit acceptable, on trouve que les 6 composantes expliquent 99.90% de la variance de X, 77.10% de la variance de Usage et 80.27% de la variance Satisf. Ce résultat montre la qualité des composantes en PLS qui permet d'expliquer aussi bien les variables explicatives que les variables à expliquer.

Figure 3

```
> summary(pls2)
Data:   X dimension: 100 7
        Y dimension: 100 2
Fit method: svdpc
Number of components considered: 7
TRAINING: % variance explained
        1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
X         46.47    67.67    80.10    90.79    98.25    99.90   100.00
Usage     17.10    59.53    60.48    74.26    76.96    77.10    77.47
Satisf    18.29    72.66    73.25    73.91    76.64    80.27    80.32
```

Figure 4



Maintenant, on s'intéresse sur la prédiction de PLS2, en utilisant la fonction PLSR dans R, la figure 7 montre un graphique qui représente la prédiction des deux variables Satisf et Usage et on remarque le nuage de point se trouve proche de la droite ce qui indique la prédiction des variables Y peut être de bonne qualité. De plus, on remarque que entre 5 et 6 composantes se stabilise, ce qui nous pousse à choisir le nombre de 6 composantes pour minimiser le taux d'erreur lors de la prédiction des variables à expliquer. Une fois le modèle de prédiction déterminer, il nous reste à juger de la qualité du modèle pour le valider. Pour cela, on va s'intéresser à l'erreur entre la valeur des variables Y et les prédictions correspondantes, on calcul la moyenne au carré de la différence entre celle-ci, la figure 8 montre les valeurs suivantes : 18.02 pour Usage et 0.14 pour satifs. De plus, en calculant RMSEP de la validation croisés (la figure 9), on constate que la moyenne de

prédiction que celle de la moyenne des variables Satisf et Usage avec une erreur d'environ de 1%. On peut donc dire que la qualité du modèle est très bonne avec le choix de 6 composantes. Pour vérifier les valeurs prédites, il faudrait réaliser d'autres test pour confirmer la pertinences des résultats.

Figure 5

```
> summary(plsr2)
Data:   X dimension: 100 7
        Y dimension: 100 2
Fit method: kernelpls
Number of components considered: 7
TRAINING: % variance explained
        1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
X         36.98    65.66    77.93    85.98    91.25    99.86   100.00
Usage     60.64    74.52    76.97    77.07    77.12    77.16    77.47
Satisf    62.47    72.95    74.49    75.47    79.83    80.32    80.32
```

Figure 6

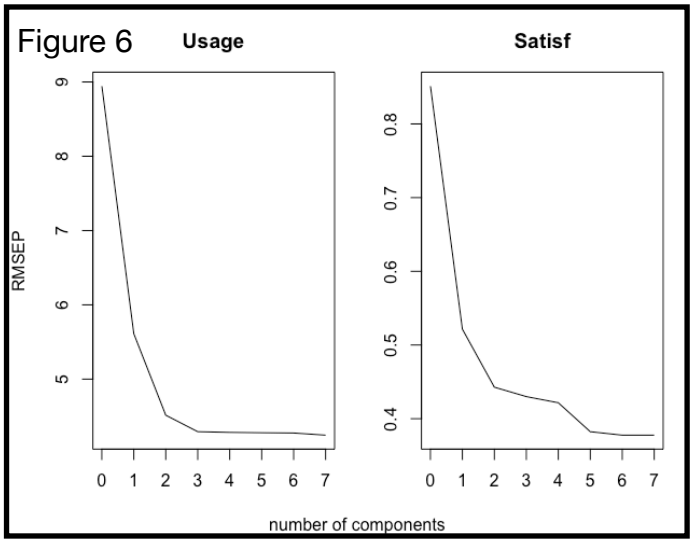


Figure 7

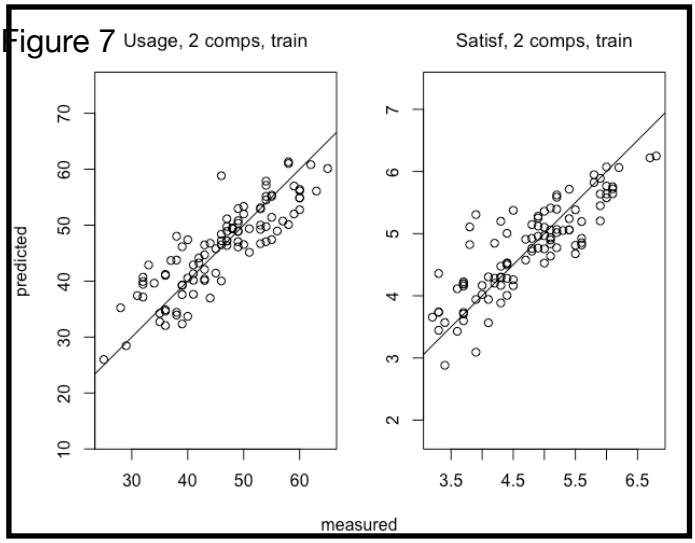


Figure 8

```
> mean(qualite_prediction$Usage^2)
[1] 18.02456
> mean(qualite_prediction$Satisf^2)
[1] 0.1426139
```

Figure 9

```
> sqrt(colMeans((qualite_prediction - Y)^2))
    Usage    Satisf
46.767248  4.831613
> colMeans(Y)
 Usage Satisf
46.100  4.771
```

Ainsi, il faut comparer les résultats précédent avec l'ACP. On remarque qu'en utilisant l'ACP, les variables à expliquer contribuer fortement avec l'axe 2 dont on devait choisir deux compensantes. Tandis qu'en utilisant PLS2, le choix de composante était 6 lequel

on a essayé de tester d'autres méthode pour valider le nombre de composantes avec la fonction PLSR (Partial least Square Regression). En effet, ACP et PLS2 se différencient par rapport au calcul de la covariance dont PLS2 maximisent le critère de covariance sous des contraintes de norme et d'orthogonalité.

PLS1

Ici, on va réaliser la méthode PLS1 pour chacune des variables à expliquer Usage et Satisf, .

Premièrement, on réalise un modèle PLSR avec la variable Usage afin de choisir le nombre de composante qui nous intéresse. D'après les résultat du graphique (Figure 10) et du tableau de résumé (Figure 11), on se doit de choisir 4 composantes qui peut être satisfaisante. Les 4 composantes expliquent 86.07% de la variance de X et 77.08% de la variance de Y. Ce choix de composante peut être confirmer avec la stabilisation de la variance des résiduels, ainsi on utilise la fonction RMSEP sous R, et on voit que la valeur se stabilise vers 4.2. En comparant avec les résultats de la PLS2, le nombre de composantes à diminuer, cela peut s'expliquer lors du calcul de matrice variance-covariance dont on prenait en compte aussi de la variable Satifs. Lors de la prochaine estimation avec PLS1 de Satifs, on s'attend d'avoir des résultats complètement différents de PLS2.

Deuxièmement, on doit vérifier la qualité de prédiction, dans la figure 11 qui montre la qualité du modele avec R^2, on voit que R^2 est proche de 80% ce qui nous pousse à dire que le modele est de très bonne qualité.

Ensuite, on réalise un nouveau modele de PLS1 avec la variable Satisf, d'où les résultats dans la figure 12 et 13. On remarque qu'on doit choisir 5 composantes car d'après le résultat de la fonction RMSEP, la valeur se stabilise vers 4.27 sur la cinquième composantes. Ainsi, c'est un résultat attendu par rapport à PLS2 comme dit précédemment mais le nombre de composante est plus élevé que la variable Usage. En vérifiant la qualité du modele PLS1 de Satifs avec R^2, on trouve que la valeur se stabilise aussi vers les 80% ce qui indique que le modele est de bonne qualité aussi comme pour la première variable Y.

Figure 10

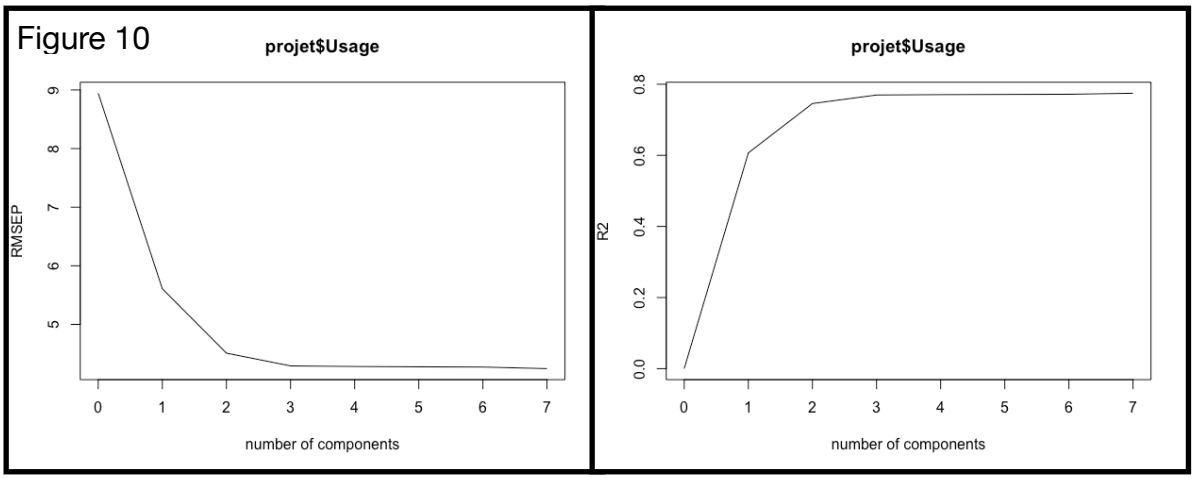


Figure 11

```
> summary(plsr1)
Data:   X dimension: 100 7
        Y dimension: 100 1
Fit method: kernelpls
Number of components considered: 7
TRAINING: % variance explained
              1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
X               36.97    65.64    77.91    86.07    92.94    99.83   100.00
projet$Usage    60.67    74.57    76.97    77.08    77.13    77.18    77.47
> RMSEP(plsr1) #qualité de prédiction
(Intercept)      1 comps      2 comps      3 comps      4 comps      5 comps      6 comps
      8.944        5.609        4.510        4.292        4.282        4.277        4.273
    7 comps
      4.246
```

Figure 12

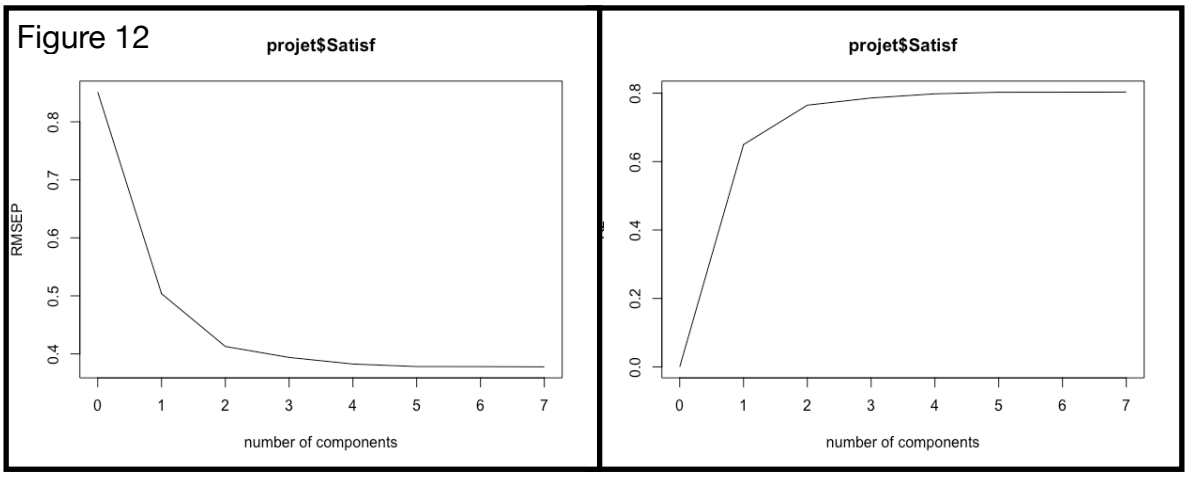


Figure 13

```
> summary(plsr11)
Data:   X dimension: 100 7
        Y dimension: 100 1
Fit method: kernelpls
Number of components considered: 7
TRAINING: % variance explained
              1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
X               36.67    67.26    75.56    81.90    88.97    99.89   100.00
projet$Satisf   64.98    76.49    78.59    79.81    80.27    80.28    80.32
> RMSEP(plsr11) #qualité de prédiction
(Intercept)      1 comps      2 comps      3 comps      4 comps      5 comps      6 comps
     0.8513       0.5038       0.4128       0.3939       0.3825       0.3781       0.3781
    7 comps
     0.3776
```

Pour conclure brièvement, PLS est une méthode d'analyse des données bien adaptée à la modélisation de ce type de données de marketing. Il s'agit d'effectuer une analyse en composantes principales (ACP), sous la contrainte que les composantes principales des X soient fortement corrélées aux composantes principales des Y. En effet, dans la première ACP, les variables Y n'étaient pas pris en compte lors de la modélisation. On peut aussi dire qu'il y a une différence entre le nombre de choix des composantes entre PLS1 et PLS2 à cause du calcul de la matrice de variance-covariance qui est différentes.

Annexe

Figure 1.A