SY09 Printemps 2017

TP 1
Statistique descriptive, Analyse en composantes principales

1 Statistique descriptive

1.1 Notes

Données

Le jeu de données contenu dans le fichier sy02-p2016.csv contient des informations relatives aux étudiants inscrits à l'UV SY02 au semestre de printemps 2016. On commencera par charger le jeu de données et déclarer les variables qualitatives comme telles :

```
> notes <- read.csv("sy02-p2016.csv", na.strings="", header=T)
> notes$nom <- factor(notes$nom, levels=notes$nom)
> notes$correcteur.median <- factor(notes$correcteur.median,
      levels=c("Cor1","Cor2","Cor3","Cor4","Cor5","Cor6","Cor7","Cor8"))
> notes$correcteur.final <- factor(notes$correcteur.final,
      levels=c("Cor1","Cor2","Cor3","Cor4","Cor5","Cor6","Cor7","Cor8"))
> notes$niveau <- factor(notes$niveau, ordered=T)
> notes$resultat <- factor(notes$resultat, levels=c("F","Fx","E","D","C","B","A"),
      ordered=T)
```

Analyse

On pourra adopter la stratégie d'étude suivante.

1. Faire une analyse descriptive générale des données, en présentant les informations disponibles (nature, domaine de définition, volume, etc). Identifier les valeurs manquantes et les expliquer si possible. Mentionner les variables qui sont a priori liées.
2. Étudier les liens statistiques entre variables. On pourra en particulier étudier si la réussite à l'UV est influencée par la formation d'origine des étudiants, leur branche, ou leur niveau. On pourra de même étudier l'influence du correcteur sur la note obtenue.

1.2 Données crabs

Données

Le jeu de données considéré, disponible dans la bibliothèque de fonctions MASS, est constitué de 200 crabes décrits par huit variables (trois variables qualitatives, et cinq quantitatives). Charger le jeu de données et sélectionner les variables quantitatives en utilisant le code R suivant :

```
> library(MASS)
> data(crabs)
> crabsquant <- crabs[,4:8]
```

Analyse

1. Effectuer dans un premier temps une analyse descriptive des données. On s'interrogera notamment sur les différences de caractéristiques morphologiques, en particulier selon l'espèce ou le sexe : semble-t-il possible d'identifier l'une ou l'autre à partir d'une ou plusieurs caractéristiques morphologiques ?
2. Dans un second temps, on étudiera la corrélation entre les différentes variables. Quelle en est vraisemblablement la cause ? Quel traitement est-il possible d'appliquer aux données pour s'affranchir de ce phénomène ?

1.3 Données Pima

Données

Le jeu de données considéré, disponible dans le fichier Pima.csv, est constitué de 532 individus de sexe féminin décrits par huit variables (dont une qualitative) :

— nombre de grossesses (npreg),
— taux plasmatique de glucose (glu),
— pression artérielle diastolique (bp),
— épaisseur du pli cutané au niveau du triceps (skin),
— indice de masse corporelle (bmi),
— fonction de pedigree du diabète (ped : mesure de l'influence génétique espérée des proches, affectés ou non par le diabète, sur le risque éventuel du sujet),
— âge (age),
— catégorie (z, diabétique si $z = 2$).

Charger le jeu de données en utilisant le code R suivant :

```
> Pima <- read.csv("Pima.csv", header=T)
> Pima$z <- factor(Pima$z)
```

Analyse

1. Effectuer dans un premier temps une analyse descriptive des données.
2. On tentera ensuite d'identifier les liens statistiques forts entre variables. En particulier, on s'intéressera au facteur « diabète », et à son influence potentielle sur les indicateurs numériques présents dans le jeu de données.

2 Analyse en composantes principales

2.1 Exercice théorique

On s'intéresse à présent plus particulièrement aux correcteurs du jeu de données notes présenté ci-dessus. On commencera par constituer un nouveau jeu de données au moyen du code suivant :

```
> moy.median <- aggregate(note.median~correcteur.median, data=notes, FUN=mean)
> names(moy.median) <- c("correcteur","moy.median")
> std.median <- aggregate(note.median~correcteur.median, data=notes, FUN=sd)
> names(std.median) <- c("correcteur","std.median")
> median <- merge(moy.median, std.median)
> moy.final <- aggregate(note.final~correcteur.final, data=notes, FUN=mean)
> names(moy.final) <- c("correcteur","moy.final")
> std.final <- aggregate(note.final~correcteur.final, data=notes, FUN=sd)
> names(std.final) <- c("correcteur","std.final")
> final <- merge(moy.final, std.final)
> correcteurs <- merge(median, final, all=T)
```

Ce jeu de données contient les moyennes et les écarts-types par correcteur, pour le médian et le final. On remarquera qu'il manque des informations pour deux correcteurs ; on les écartera donc dans un premier temps du jeu de données :

```
> corr.acp <- correcteurs[-c(2,3),]
```

On associera dans cet exercice les mêmes pondérations à tous les individus, et on munira $\mathbb{R}^p$ de la métrique euclidienne.

1. Calculer les axes factoriels de l'ACP du nuage de points défini par les quatre variables quantitatives. Quels sont les pourcentages d'inertie expliquée par chacun de ces axes ?
2. Calculer les composantes principales ; en déduire la représentation des six individus dans le premier plan factoriel.
3. Tracer la représentation des quatre variables dans le premier plan factoriel.
4. Calculer l'expression $\sum_{\alpha=1}^{k} \mathbf{c}_\alpha \mathbf{u}'_\alpha$ pour les valeurs $k = 1, 2$ et 3. À quoi correspond cette somme lorsque $k = 4$?
5. On souhaite représenter les individus initialement écartés de l'ACP. Remplacer chacune de leurs valeurs manquantes par la moyenne de la variable correspondante (imputation par la moyenne), puis représenter ces individus dans les deux premiers plans factoriels.

2.2 Utilisation des outils R

L'objectif de cet exercice est de se familiariser avec les fonctions R permettant d'effectuer une ACP, en particulier les fonctions `princomp`, `summary`, `loadings`, `plot` et `biplot`. Remarquons qu'il existe une autre fonction `prcomp` qui effectue les calculs de manière différente ; on ne l'utilisera pas ici.

— En utilisant ces fonctions, effectuer l'ACP du jeu de données notes étudiées en cours. Montrer comment on peut retrouver tous les résultats alors obtenus (valeurs propres, axes principaux, composantes principales, représentations graphiques, ...).

— On s'intéresse à l'affichage des résultats de la fonction `princomp`. Qu'affichent les fonctions `plot` et `biplot` ? Détailler plus particulièrement le fonctionnement de la fonction `biplot` redéfinie pour la classe `princomp` (accessible par `biplot.princomp`) et de ses différentes options.

2.3 Données Crabs

Cette étude vise à utiliser l'ACP pour trouver une représentation des crabes qui permettent de distinguer visuellement différents groupes, liés à l'espèce et au sexe.

1. Tester tout d'abord l'ACP sur `crabsquant` sans traitement préalable. Que constatez-vous ? Comment pouvez-vous expliquer ce phénomène à la lumière des analyses menées au paragraphe 1.2 ?
2. Trouver une solution pour améliorer la qualité de votre représentation en termes de visualisation des différents groupes.

2.4 Données Pima

On cherche ici à représenter les données de manière à distinguer visuellement les groupes de patientes diabétiques et non diabétiques.

Tenter une ACP sur les données. Que constate-t-on ? Semble-t-il possible de trouver une représentation simple qui permette de distinguer les deux catégories de patientes ?