SY19 Automne 2011
TP 0 : Analyse Factorielle Discriminante linéaire

1 Principe de la méthode

L'analyse factorielle discriminante (AFD) linéaire est une technique statistique essentiellement descriptive. Contrairement à l'analyse en composantes principales (ACP), les données sont ici séparées en classes ; l'objectif est de trouver un espace de représentation dans lequel ces classes sont séparées autant que possible.

2 Questions théoriques

Soit $\omega_1, \ldots, \omega_K$ l'ensemble des classes de points considérées. Soit n_k l'effectif de la classe ω_k ; on note μ la moyenne empirique calculée à partir de tous les points et μ_k la moyenne empirique calculée à partir des points de la classe ω_k :

$$\begin{aligned} \mu &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \\ \mu_k &= \frac{1}{n_k}\sum_{\mathbf{x}_j\in\omega_k}\mathbf{x}_j. \end{aligned}$$

Soit S la matrice de variance-covariance du nuage Ω, et S_k la matrice de variance-covariance des points de la classe ω_k :

$$\begin{aligned} S &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\mu)(\mathbf{x}_i-\mu)^\top, \\ S_k &= \frac{1}{n_k}\sum_{j:\mathbf{x}_j\in\omega_k}(\mathbf{x}_j-\mu_k)(\mathbf{x}_j-\mu_k)^\top. \end{aligned}$$

On supposera par la suite que l'espace est muni de la métrique euclidienne (associée à la matrice identité).

1. En utilisant le théorème de Huygens, montrer que la matrice de variance-covariance S se décompose en deux termes :
$$S = S_W + S_B,$$
où S_W est appelé terme de variance intra-classes et S_W terme de variance inter-classes.
2. Supposons que l'on n'ait que deux classes ω_1 et ω_2, et que l'on cherche un unique axe discriminant Δu_1 pour les distinguer. Le principe de l'AFD linéaire est alors de déterminer Δu_1 de vecteur directeur $\mathbf{u}_1$ en maximisant $\mathbf{u}_1^\top S_B\mathbf{u}_1/\mathbf{u}_1^\top S_W\mathbf{u}_1$. Expliquer le choix de ce critère.

En pratique, la recherche des p axes factoriels discriminants se fait par la résolution de l'équation $S_W^{-1}S_B\mathbf{u} = \lambda\mathbf{u}$. Les valeurs propres et vecteurs propres de $S_W^{-1}S_B$ fournissent donc une solution au problème de l'AFD.

3 Exercices pratiques

Charger la bibliothèque de fonctions ade4. Effectuer une analyse factorielle discriminante des données iris et comparer au résultat obtenu par ACP.

Charger les données crabs (charger pour cela la bibliothèque de fonctions MASS). Effectuer l'AFD et comparer avec les résultats obtenus par ACP, dans un premier temps sur les données brutes, et dans un second temps sur les données pré-traitées :

```
> library(MASS)
> data(crabs)
> crabsquant <- crabs[,4:8]
> crabsquant <- crabsquant/matrix(rep(crabsquant[,4],4),
  nrow=dim(crabsquant)[1],byrow=F)
```