Exercice - Session 5

Un groupe industriel commercialisant du café souhaite élaborer une règle automatique de détermination du lieu de production d'un café à partir de spectrométrie proche infra-rouge. Pour cela, il dispose d'un tableau de données contenant les spectres proches infra-rouge et le lieu de production, codé par un entier entre 1 et 7, de 240 échantillons de café : 50 du lieu 1, 26 du lieu 2, 26 du lieu 3, 13 du lieu 4, 22 du lieu 5, 84 du lieu 6 et 19 du lieu 7.

Dans la suite, Y désigne le lieu de production d'un café et $x = [x(\lambda_1), \ldots, x(\lambda_m)]'$ son spectre proche infra-rouge après transformation SNV (ici, $m = 1050$ et $\lambda_1 < \lambda_2 < \ldots < \lambda_m$ est la séquence des nombres d'onde supports du spectre). Dans un premier temps, on choisit de construire un modèle de la probabilité qu'un café ait été produit en un lieu donné par son spectre proche infra-rouge : $\mathbb{P}_x(Y = j),\ j = 1, \ldots, 7$.

1. Comment s'appelle ce modèle ? Donner l'expression du modèle et son nombre de paramètres.

A l'aide de la fonction `multinom` du package `nnet` de `R`, on tente d'ajuster le modèle de la question 1.

2. Quel critère de qualité d'ajustement la fonction `multinom` *a-t'elle tenté d'optimiser ? Expliquez le problème d'ajustement que l'on observe.*

Afin de réduire la dimension du profil de variables explicatives, on approche chaque spectre proche infra-rouge par une fonction spline. Le postulat de cette méthode est que chaque valeur spectrale $x(\lambda_i)$ peut être décomposée de la façon suivante :

$$x(\lambda_i) \quad = \quad a_1 b_1(\lambda_i) + a_2 b_2(\lambda_i) + \ldots + a_k b_k(\lambda_i) + e(\lambda_i), \tag{1}$$

où les fonctions $b_j(\lambda)$ sont connues, appelées B-splines, les coefficients a_j sont des paramètres inconnus et $e(\lambda)$ est une erreur résiduelle. De manière équivalente, si $\boldsymbol{x} = (x(\lambda_1), \ldots, x(\lambda_m))'$ désigne le vecteur des valeurs observées à chaque nombre d'onde, alors :

$$\boldsymbol{x} \quad = \quad \boldsymbol{B}\boldsymbol{a} + \boldsymbol{e},$$

où $\boldsymbol{B}$ est la matrice $m \times k$ dont le terme (i, j) est $b_j(\lambda_i)$, $\boldsymbol{a} = (a_1, \ldots, a_k)'$ et $\boldsymbol{e} = (e(\lambda_1), \ldots, e(\lambda_m))'$.

3. Quelle est la nature mathématique des fonctions $b_j(\lambda)$? Comment le paramètre k influence-t'il la qualité d'approximation de $x(\lambda)$ par sa décomposition linéaire (1) ?

Dans la suite, on choisit arbitrairement $k = 30$.

4. Donnez l'expression du vecteur $\hat{\boldsymbol{a}} = (\hat{a}_1, \ldots, \hat{a}_k)'$ de l'estimateur de $\boldsymbol{a} = (a_1, \ldots, a_k)'$ par la méthode des moindres carrés, pour un spectre $\boldsymbol{x} = [x(\lambda_1), \ldots, x(\lambda_m)]'$.

5. Donnez une estimation de la variance de $e(\lambda)$ pour le spectre associé au 1er individu de l'échantillon.

6. Construire ainsi la matrice A dont la $i$ème ligne est le vecteur $\hat{a}$ associé au spectre proche infra-rouge du $i$ème café.

Afin de mieux comprendre comment les coefficients $\hat{a}$ peuvent permettre de caractériser les différents lieux de production de café, on choisit d'utiliser l'analyse discriminante linéaire.

7. Construire le graphique représentant la répartition des échantillons de cafés décrits par leurs 2 premiers scores discriminants.

8. Quel score discriminant permet le mieux de différencier les lieux de production 4 et 5 ? Même question pour les lieux de production 3 et 6 ? Quelle conséquence sur le nombre de scores discriminants à retenir dans le modèle d'analyse discriminante linéaire de Fisher ?

9. Donnez la matrice de confusion associée à la règle d'affectation de la question précédente obtenue par validation croisée (10-segments).