Université de Caen	Examen du 02 décembre 2021. Durée 1h30	**Analyse de données avancée**
UFR des Sciences	Par F. Chamroukhi	**M2 SAAD/DECIM**

Consignes : Documents interdits. Composer avec un stylo. La feuille double d'examen doit porter vos nom, prénom, et signature, cachés par collage.

Exercice 1 *(2 pts/question) :* On considère un échantillon indépendant $((\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n))$ d'un couple $(\boldsymbol{X}, Y)$ où $\boldsymbol{X}$ représente un vecteur de d descripteurs réels ($\boldsymbol{X} \in \mathbb{R}^d$) et $Y \in \{1, \ldots, K\}$ la classe. On cherche à prédire les classes d'un ensemble de test $((\boldsymbol{x}_{n+1}, y_{n+1}), \ldots, (\boldsymbol{x}_m, y_m))$ sur la base d'un modèle de paramètres $\boldsymbol{\theta}$ appris sur un ensemble d'apprentissage $((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n))$ par la règle de décision :

$$\widehat{y}_i = \arg \max_{k \in \{1,\ldots,K\}} \mathbb{P}(Y_i = k | \boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta}) \quad \forall i = n+1, \ldots, m. \tag{1}$$

Soient le modèle de régression softmax et celui d'analyse discriminante suivants : $\forall k \in \{1, \ldots, K\}$:

$$\text{Régression softmax} \quad : \quad \mathbb{P}(Y = k | \boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\theta}) = \frac{\exp(\alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{x})}{1 + \sum_{\ell=1}^{K-1} \exp(\alpha_\ell + \boldsymbol{\beta}_\ell^T \boldsymbol{x})} \tag{2}$$

$$\text{Analyse Discriminante} \quad : \quad \mathbb{P}(Y = k, \boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta}) = w_k \phi_d(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3}$$

où $w_k = \mathbb{P}(Y = k)$ $\forall k$ et $\phi_d(\boldsymbol{x}; \mathbf{m}, \mathbf{R}) = \frac{1}{(2\pi)^{\frac{d}{2}} |\mathbf{R}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \mathbf{m})^T \mathbf{R}^{-1} (\boldsymbol{x} - \mathbf{m})\right)$ est la densité d'un vecteur Gaussien de $\mathbb{R}^d$ d'espérance $\mathbf{m}$ et de matrice de covariance $\mathbf{R}$.

1. Montrer que la règle de décision entre une classe k et une classe h avec la régression softmax est de la forme $a + \boldsymbol{b}^T \boldsymbol{x}$ (i.e. linéaire en $\boldsymbol{x}$), dont on identifiera les valeurs de a et de $\boldsymbol{b}$ en fonction de celles de α_k, α_h, $\boldsymbol{\beta}_k$ et $\boldsymbol{\beta}_h$.
2. Montrer que la règle décision entre une classe k et une classe h avec l'analyse discriminante, dans le cas où $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots, = \boldsymbol{\Sigma}_K = \boldsymbol{\Sigma}$, est aussi de la forme $a + \boldsymbol{b}^T \boldsymbol{x}$ (i.e. linéaire en $\boldsymbol{x}$), dont on identifiera les valeurs de a et de $\boldsymbol{b}$ en fonction de celles de w_k, w_h, $\boldsymbol{\mu}_k$, $\boldsymbol{\mu}_h$ et $\boldsymbol{\Sigma}$.
3. Discuter le(s) lien(s) entre la régression softmax et l'analyse discriminante dans un tel cas.
4. Calculer $\widehat{\boldsymbol{\mu}}_k$ l'estimateur du maximum de vraisemblance de $\boldsymbol{\mu}_k$, $\forall k \in \{1, \ldots, K\}$
5. Calculer $\widehat{\boldsymbol{\Sigma}}_k$ l'estimateur du maximum de vraisemblance de $\boldsymbol{\Sigma}_k$, $\forall k \in \{1, \ldots, K\}$

Exercice 2 *(2 pts/question) :* On considère un jeu de données multivariées $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ où $\boldsymbol{x}_i \in \mathbb{R}^p$ est un vecteur de p variables réelles. On cherche à réduire la dimension en projetant ce jeu de données dans un espace de dimension M plus réduite et ce au sens de l'ACP. Soient $\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i$ le vecteur de la moyenne empirique et

$$\mathbf{S} = \frac{1}{n} \sum_{i=1}^n (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T \tag{4}$$

la matrice de variance-covariance empirique.

1. On suppose que l'on projette les données dans un espace de dimension 1, i.e., une droite de vecteur directeur $\boldsymbol{u}_1$, que l'on suppose normé (i.e.,$\boldsymbol{u}^T \boldsymbol{u} = 1$). Donner l'expression de la variance des données projetées dans cet espace et exprimer là en fonction de $\mathbf{S}$ et $\boldsymbol{u}_1$. On la notera v_1.
2. Sachant que l'ACP maximise la variance dans l'espace projeté v_1, montrer que le vecteur $\boldsymbol{u}_1$ définissant cet espace correspond au vecteur propre associé à la plus grande valeur propre de $\mathbf{S}$.
3. On suppose maintenant que l'on projette les données dans un espace de dimension $M \leq p$ déterminé par le sous espace définie par la base orthonormée $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M)$, i.e.,

   $$\boldsymbol{u}_j^T \boldsymbol{u}_j = 1 \ \forall j \text{ et } \boldsymbol{u}_j^T \boldsymbol{u}_k = 0 \ \forall j \neq k.$$

   Montrer que les vecteurs de cette base sont les vecteurs propres de $\mathbf{S}$ organisés selon l'ordre décroissant des valeurs propres associées.
4. Donner un critère pour choisir la dimension M.
5. Donner en une phrase une interprétation de l'ACP.