Apprentissage en grande dimension

Ex 1. Soit $\mathbf{X} = (X_1, X_2)'$, où $X_1 \sim \text{Bern}(p)$ et $X_2 = 1 - X_1$.

- Calculer la matrice de covariance $\mathbf{\Sigma}$ de $\mathbf{X}$.
- Déterminer les composantes principales et les interpréter.

Ex.2 Soit

$$\mathbf{\Sigma} = \sigma^2 \begin{pmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \dots & \rho & 1 \end{pmatrix},$$

avec $\sigma > 0$ et $0 \leq \rho \leq 1$.

- Pour chaque $k = 1, \dots, p$ déterminer la proportion de la variance expliquée par les k premières composantes principales.
- Déterminer la première composante principale si $\rho > 0$.
- Que se passe-t-il pour $\rho = 0$ et $\rho = 1$?

Ex 3. (utiliser le jeu de données iris de Fisher avec R "library(datasets) data(iris)") Pour la variété Setosa:

- Avec R, déterminer la matrice de covariance empirique $\mathbf{S}$ et trouver les composantes principales.
- Déterminer la proportion de variabilité contenue dans les trois premières composantes principales.

Ex4 Pour chaque individu de la variété Setosa, générer 46 nouvelles variables avec la loi $\mathcal{N}(2, 1)$ (mutuellement indépendantes). Réaliser ensuite une Analyse en Composante Principale sparse sur les données obtenues.

Ex5 Considérons le problème de classification avec deux populations π_1 et π_2, où $\pi_i = \mathcal{N}_2(\boldsymbol{\mu}_i, \mathbf{\Sigma})$, avec

$$\mathbf{\Sigma} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix},$$

et $\boldsymbol{\mu}_i = (i, i)$, $i = 1, 2$. Nous savons que les coûts de misclassification sont donnés par $c_{1|2} = 1$ et $c_{2|1} = 2$. A priori, les deux populations sont équiprobables. Determiner l'équation de la droite qui sera utilisée pour discriminer les deux populations.

Ex.6 (utiliser le jeu de données iris de Fisher avec R "library(datasets) data(iris)"). Construire la classification linéaire pour la comparaison entre la variété Setosa et la variété versicolor.