1.6.2 Analyse des composantes principales

❒ **Valeur propre, vecteur propre** – Étant donnée une matrice $A \in \mathbb{R}^{n\times n}$, λ est dite être une valeur propre de A s'il existe un vecteur $z \in \mathbb{R}^n\backslash\{0\}$, appelé vecteur propre, tel que :

$$\boxed{Az = \lambda z}$$

❒ **Théorème spectral** – Soit $A \in \mathbb{R}^{n\times n}$. Si A est symétrique, alors A est diagonalisable par une matrice réelle orthogonale $U \in \mathbb{R}^{n\times n}$. En notant $\Lambda = \textrm{diag}(\lambda_1,...,\lambda_n)$, on a :

$$\boxed{\exists\Lambda \textrm{ diagonal},\quad A = U\Lambda U^T}$$

Remarque : le vecteur propre associé à la plus grande valeur propre est appelé le vecteur propre principal de la matrice A.

❒ **Algorithme** – La procédure d'analyse des composantes principales (en anglais *Principal Component Analysis* ou *PCA*) est une technique de réduction de dimension qui projette les données sur k dimensions en maximisant la variance des données de la manière suivante :

— Étape 1 : Normaliser les données pour avoir une moyenne de 0 et un écart-type de 1.

$$\boxed{x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{\sigma_j}} \quad \textrm{where} \quad \boxed{\mu_j = \frac{1}{m}\sum_{i=1}^m x_j^{(i)}} \quad \textrm{and} \quad \boxed{\sigma_j^2 = \frac{1}{m}\sum_{i=1}^m (x_j^{(i)} - \mu_j)^2}$$

— Étape 2 : Calculer $\displaystyle\Sigma = \frac{1}{m}\sum_{i=1}^m x^{(i)}{x^{(i)}}^T \in \mathbb{R}^{n\times n}$, qui est symétrique avec des valeurs propres réelles.

— Étape 3 : Calculer $u_1, ..., u_k \in \mathbb{R}^n$ les k valeurs propres principales orthogonales de Σ, i.e. les vecteurs propres orthogonaux des k valeurs propres les plus grandes.

— Étape 4 : Projeter les données sur $\textrm{span}_{\mathbb{R}}(u_1,...,u_k)$. Cette procédure maximise la variance sur tous les espaces à k dimensions.