$$\boxed{s=\frac{b-a}{\max(a,b)}}$$

❒ **Index de Calinski-Harabaz** – En notant k le nombre de partitions, B_k et W_k les matrices de dispersion entre-partitions et au sein d'une même partition sont définis respectivement par :

$$B_k=\sum_{j=1}^k n_{c^{(i)}}(\mu_{c^{(i)}}-\mu)(\mu_{c^{(i)}}-\mu)^T,\qquad W_k=\sum_{i=1}^m(x^{(i)}-\mu_{c^{(i)}})(x^{(i)}-\mu_{c^{(i)}})^T$$

l'index de Calinski-Harabaz $s(k)$ renseigne sur la qualité des partitions, de sorte à ce qu'un score plus élevé indique des partitions plus denses et mieux séparées entre elles. Il est défini par :

$$\boxed{s(k)=\frac{\textrm{Tr}(B_k)}{\textrm{Tr}(W_k)}\times\frac{N-k}{k-1}}$$

2.3 Réduction de dimension

2.3.1 Analyse des composantes principales

C'est une technique de réduction de dimension qui trouve les directions maximisant la variance, vers lesquelles les données sont projetées.

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

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

❒ **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 *PCA - Principal Component Analysis*) 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.

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

— 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 et aux 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.