́TD 1 : Principal components analysis (PCA)

Introduction

L'objectif de ce devoir est de découvrir l'Analyse en Composantes Principales (ACP). Ce travail explore deux rôles fondamentaux de l'ACP :

1. Trouver la meilleure projection (de faible dimension) des données.
2. Effectuer un changement de base où les nouvelles variables (composantes principales) ne sont pas corrélées.

On observe n échantillons $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$ d'un **vecteur** aléatoire $\mathbf{X}$ de dimension d. Ainsi, la matrice des données est : $\mathrm{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix} \in \mathbb{R}^{n \times d}$

1 Questions préliminaires

1. Déterminez l'estimateur empirique de la moyenne $\mathbb{E}(\mathbf{X})$ en fonction de la matrice X.
2. Déterminer un estimateur empirique de la matrice de covariance $\mathbb{V}(\mathbf{X})$ en fonction de n et X.
3. On suppose que $\mathbf{X}$ est centré, c-à-d $\mathbb{E}(\mathbf{X}) = 0$. Montrez que l'estimateur empirique de $\mathbb{V}(\mathbf{X})$ noté par Σ est égal à $\frac{1}{n}\mathrm{X}^\top\mathrm{X}$.
4. Écrire Σ en fonction des $\mathbf{x}_i$ en utilisant le produit matriciel colonnes-lignes.
5. En pratique, on n'a quasiment jamais $\mathbb{E}(\mathbf{X}) = 0$. Comment peut-on y remédier ?

2 ACP comme meilleure projection des données

On suppose dorénavant que $\mathbb{E}(\mathbf{X}) = 0$. Soit E un sous-espace vectoriel de dimension $k \leq d$ fixée. On note φ_E la projection orthogonale sur E. On cherche le sous-espace E telle que la projection qui minimise la perte d'information moyenne sur tous les $\mathbf{x}_i$. Ainsi, on cherche à résoudre :

$$\min_{\substack{E \\ \dim(E)=k}} \quad \frac{1}{n}\sum_{i=1}^n \|\mathbf{x}_i - \varphi_E(\mathbf{x}_i)\|^2 \tag{1}$$

2.1 Projection en dimension 1

On fixe $k = 1$. On cherche donc la meilleure projection sur une droite donnée par un vecteur $\mathbf{p}$ unitaire (que l'on cherche) :

$$\min_{\substack{\mathbf{p}\in\mathbb{R}^d \\ \|\mathbf{p}\|=1}} \quad \frac{1}{n}\sum_{i=1}^n \|\mathbf{x}_i - \varphi_{\mathbf{p}}(\mathbf{x}_i)\|^2 \tag{2}$$

1. Déterminer $\varphi_{\mathbf{p}}(\mathbf{x})$ en fonction de $\mathbf{p}$ et $\mathbf{x}$.
2. Montrez que le problème d'optimisation (2) est équivalent à :

$$\max_{\substack{\mathbf{p}\in\mathbb{R}^d \\ \|\mathbf{p}\|=1}} \frac{1}{n}\sum_{i=1}^n (\mathbf{p}^\top\mathbf{x}_i)^2 \tag{3}$$

3. Montrez que l'on peut réécrire ce problème d'optimisation (3) comme :

$$\max_{\substack{\mathbf{p}\in\mathbb{R}^d \\ \|\mathbf{p}\|=1}} \mathbf{p}^\top\Sigma\mathbf{p} \tag{4}$$

4. En déduire la meilleure projection des données en dimension 1.
5. Quel problème d'optimisation aurait-on à résoudre si les données n'étaient pas centrées ? Commenter.

2.2 Projection en dimension $k > 1$

Soit E un sous-espace vectoriel de $\mathbb{R}^d$ de dimension k. Soit $\mathbf{p}_1, \dots, \mathbf{p}_k$ une base orthonormale de E. On admet que l'on peut résoudre le problème (1) de proche en proche : en cherchant $\mathbf{p}_1$ puis $\mathbf{p}_2$ orthogonal à $\mathbf{p}_1$ puis $\mathbf{p}_3$ orthogonal à $\mathrm{Vect}(\mathbf{p}_1, \mathbf{p}_2)$ puis ... etc.

1. Déterminer la base orthogonale $\beta \stackrel{\text{def}}{=} (\mathbf{p}_1, \dots, \mathbf{p}_k)$.
2. Quelles sont les coordonnées de $\varphi(\mathbf{x}_i)$ dans la base β ?
3. On pose $\mathrm{P} \in \mathbb{R}^{d \times k}$ la matrice dont les colonnes sont les $\mathbf{p}_i$. En déduire la nouvelle matrice de taille $n \times k$ des données projetées sur le sous-espace E dans la base β en fonction de X et P.
4. On appelle les $\mathbf{p}_k$ des *composantes principales*. On suppose que l'on dispose d'une fonction qui permet de diagonaliser une matrice symétrique. Proposez un algorithme qui prend en argument des données X et une dimension k et effectue la projection orthogonale des données sur le meilleur sous-espace de dimension k.