Master 2 Data Science
Statistique en Grande Dimension et Apprentissage

Quelques compléments sur l'ACP

1 L'ACP vue comme un problème de minimisation

Théorème 1.1. *Notons X un vecteur aléatoire à valeurs dans $\mathbb{R}^p$ (en pratique centré). Notons $C_X = \mathbb{E}[XX^T]$. Alors, pour tout $d \leq p$,*

$$\text{Argmin}_{\dim(V)\leq d}\mathbb{E}[\|X - \text{Proj}_V(X)\|^2]$$

est atteint en le sous-espace vectoriel $V^\star = \text{Vect}(u_1, \ldots, u_d)$ où $u_1, \ldots, u_d$ désignent des vecteurs propres orthonormés associés aux d plus grandes valeurs propres de C_X. Ce sous-espace peut être vu comme la résolution successive des problèmes d'optimisation suivants :

$$u_1 = \text{Argmax}_{u,\|u\|=1} u^T C_X u$$

puis, récursivement,

$$u_k = \text{Argmax}_{u\perp(u_1,\ldots,u_{k-1},\|u\|=1} u^T C_X u.$$

Remarque 1.2. *Notons que l'on peut utiliser la méthode de la déflation : pour tout k,*

$$u_k = \text{Argmax}_{u,\|u\|=1} u^T C_k u,$$

où $C_1 = C_X$ et $C_k = C_{k-1} - \lambda_{k-1} u_k u_k^T$. La matrice C_k est en effet par construction une matrice symétrique dont les valeurs propres sont $0, \ldots, 0, \lambda_k, \ldots, \lambda_p$.

Proof. Si $V = (v_1, \ldots, v_d)$ désigne une base orthonormée du sous-espace vectoriel V (on prend abusivement la même notation pour la matrice V et le sous-espace vectoriel engendré), alors

$$p_V(x) = V.V^T x.$$

En effet, $V.V^T x = \sum_{i=1}^d \langle v_i, x\rangle v_i$. Donc si $x \in V$, $p_V(x) = x$ et si $x \in V^\perp$, $p_V(x) = 0$. Ainsi,

$$\|x - p_V(x)\|^2 = \langle (I_p - VV^T)x, (I_p - VV^T)x\rangle = \langle (I_p - VV^T)x, x\rangle = \langle x, x\rangle - \langle VV^T x, x\rangle.$$

Ainsi, minimiser $\mathbb{E}[\|X - p_V(X)\|^2]$ sur toutes les systèmes orthonormés de d vecteurs revient à maximiser

$$F(V) = \mathbb{E}[\|V^T X\|^2].$$

Dans le cas où V est un sous-espace de dimension 1, cela revient à maximiser $F(v) = v^T A v$ où $A = \mathbb{E}[X.X^T]$. Il est alors bien connu qu'il faut choisir v comme le vecteur propre associé à la plus grande valeur propre de A (qui est une matrice symétrique positive). En effet, supposons que $(u_1, \ldots, u_p)$ soit une base orthonormée de vecteurs propres associée aux p valeurs propres $\lambda_1 \geq \ldots \geq \lambda_p$ de A, alors si pour tout vecteur v de norme 1, $v = \sum_{i=1}^p \alpha_i u_i$ avec $\sum_{i=1}^p \alpha_i^2$ de sorte que

$$v^T A v = \sum_{i=1}^p \alpha_i^2 \lambda_i \leq \lambda_1$$

et $v^T A v = \lambda_1$ avec $v = u_1$. Si maintenant $V = (v_1, v_2)$ avec $v_1 = \sum_{i=1}^p \alpha_i u_i$ et $v_2 = \sum_{i=1}^p \beta_i u_i$ avec $\langle v_1, v_2 \rangle = \sum_{i=1}^p \alpha_i \beta_i = 0$, alors

$$F(V) = v_1^T A v_1 + v_2^T A v_2 = \sum_{i=1}^p \alpha_i^2 \lambda_i + \sum_{i=1}^p \beta_i^2 \lambda_i,$$

et en notant $\alpha = (\alpha_i)_{i=1}^p$ et $\beta = (\beta_i)_{i=1}^p$

$$\max F(V) = \max_{\beta, \langle \beta, \alpha \rangle = 0} \max_{\alpha} \left(\sum_{i=1}^p \alpha_i^2 \lambda_i + \sum_{i=1}^p \beta_i^2 \lambda_i \right).$$

Or, le max en α est encore atteint pour $\alpha = (1, 0, \ldots, 0)$. Ainsi, $\beta = (0, \beta_2, \ldots, \beta_p)$ et il vient que $\sum_{i=1}^p \beta_i^2 \lambda_i$ est maximisé lorsque $\beta = (0, 1, 0, \ldots, 0)$. Le même raisonnement peut facilement se généraliser à un sous-espace de dimension d. □

L'ACP sur les données $(x^{(1)}, \ldots, x^{(n)})$ peut alors être comprise comme la solution du problème ci dessus lorsque

$$\mathbb{P}_X = \frac{1}{n} \sum_{k=1}^n \delta_{x^{(k)}}.$$

La matrice de covariance empirique peut alors être également vue comme $\mathbb{E}[XX^T]$ lorsque $X \sim \frac{1}{n} \sum_{k=1}^n \delta_{x^{(k)}}$. Par la loi des grands nombres, si $(X^{(1)}, \ldots, X^{(n)})$ désigne un n-échantillon associé à X, $\frac{1}{n} \sum_{k=1}^n \delta_{X^{(k)}}$ converge *p.s.* vers $\mathbb{P}_X$ de sorte que l'on retrouve asymptotiquement la solution du "vrai" problème de minimisation. Malheureusement, pour les raisons déjà évoquées dans le chapitre sur le fléau de la dimension, cette convergence n'est utilisable que si $p \ll n$. Sans expliquer en détail le "pourquoi", on pourra noter que la variance de $\|X - \text{Proj}_V(X)\|^2$ est "naturellement" de l'ordre de p puisque l'on a p variables.

2 L'ACP sparse

Pour pallier le problème de variance évoqué plus haut, l'idée est de résoudre un problème similaire où l'on restreint l'ensemble des vecteurs v. C'est le principe de l'ACP sparse dont l'algorithme mime celui de l'ACP construite de manière récursive en déterminant dans un premier temps

$$u_1 = \text{Argmax}_{u, \|u\|=1, \|u\|_0 \leq s} u^T C_X u$$

où $\|u\|_0$ désigne le nombre de coordonnées non nulles de u (Attention, même si la notation le suggère, il ne s'agit pas d'une norme !). On détermine ensuite u_2 en cherchant

$$\text{Argmax}_{u, \|u\|=1, \|u\|_0 \leq s, \langle u, u_1 \rangle = 0} u^T C_X u$$

et ainsi de suite.

La solution de ce problème supporte mieux la dimension (on l'expliquera dans le chapitre sur la régression linéaire pénalisée). Il "suffit" de l'appliquer la matrice de covariance empirique. Malheureusement, déterminer numériquement la solution de ce problème n'est pas possible. Il s'agit d'un problème *NP-hard*. On remplace alors la contrainte $\|u\|_0 \leq s$ par $\|u\|_1 \leq \alpha$ où $\|v\|_1 = \sum_{i=1}^p |v_i|$. C'est ce type d'ACP qui est par exemple programmé dans le module *sklearn* **SparsePCA**.

3 L'ACP à noyau

Ce type d'ACP sera expliqué à la fin du chapitre sur les SVMs. Il permet d'étendre l'ACP à des sous-ensembles non linéaires.