TP7: KERNEL PRINCIPAL COMPONENT ANALYSIS

COURS D'APPRENTISSAGE, ECOLE NORMALE SUPÉRIEURE

Raphaël Berthier
raphael.berthier@inria.fr

This exercise sheet was composed using [1, Exercises 1.6.4, 9.5.6].

1. Plain Principal Component Analysis (PCA)

Let $X_1, \dots, X_d \in \mathbb{R}^p$ be a sequence of samples. The goal of the PCA is to compute a low-dimensional representation of the data. For a given dimension $d \le p$, the PCA computes the subspace V_d such that

$$V_d \in \operatorname*{argmin}_{V \subset \mathbb{R}^p \,:\, \dim(V) \le d} \sum_{i=1}^n \|X_i - \operatorname{Proj}_V X_i\|_2^2 \,, \tag{1}$$

where Proj_V denotes the orthogonal projection onto V. Let us denote X the data matrix, whose lines are $X_1, \dots, X_n$, and $X = \sum_{k=1}^r \sigma_k u_k v_k^T$ its singular values decomposition (SVD), with $\sigma_1 \ge \dots \ge \sigma_r > 0$.

We have the following theorem

Theorem 1. *Let $d \le r$. We have*

$$\min_{Y : \operatorname{rank}(Y) \le d} \|Y - X\|_F^2 = \sum_{k=d+1}^r \sigma_k^2 \,,$$

where $\|.\|_F$ denotes the Frobenius norm of a matrix, $\|Z\|_F^2 = \operatorname{tr}(ZZ^T) = \sum_{i,j} Z_{i,j}^2$. Moreover, the minimum is reached if

$$Y = \sum_{k=1}^d \sigma_k u_k v_k^T \,.$$

1) Check that

$$\sum_{i=1}^n \|X_i - \operatorname{Proj}_V X_i\|_2^2 = \|X - X \operatorname{Proj}_V\|_F^2 \,.$$

Conclude that $V_d = \{v_1, \dots, v_d\}$ minimizes (1).

2) Prove that the coordinates of $\operatorname{Proj}_V X_i$ in the orthonormal basis $(v_1, \dots, v_d)$ of V_d are given by $(\sigma_1 \langle e_i, u_1 \rangle, \dots, \sigma_d \langle e_i, u_d \rangle)$.

The right-singular vectors $v_1, \dots, v_d \in \mathbb{R}^p$ are called the principal axes. There represent the directions where the data varies the most. The vectors $c_k = X v_k = \sigma_k u_k \in \mathbb{R}^n, k = 1, \dots, d$ are called the principal components. The principal component c_k gathers the coordinates of $X_1, \dots, X_n$ on v_k.

Note that in practice, since V_d is a linear span and not an affine span, it is highly recommended to first center the data point $\tilde{X}_i = X_i - \frac{1}{n} \sum X_i$ before applying the PCA.

2. Kernel principal component analysis

The Reproducing Kernel Hilbert Space (RKHS) framework allows us to "delinearize" some linear algorithm. We show here how it can be applied to PCA. Let us consider that we have n points $X_1, \dots, X_n \in \mathcal{X}$. Note that in the kernel setting, $\mathcal{X}$ does not need to be a vector space anymore. Let $\phi : \mathcal{X} \to \mathcal{F}$ be a mapping from $\mathcal{X}$ to some RKHS $\mathcal{F}$ associated to some positive definite kernel k on $\mathcal{X}$. The principal of kernel PCA is to perform a PCA on the points $\phi(X_1), \dots, \phi(X_n)$ without computing the points $\phi(X_1), \dots, \phi(X_n)$ explicitly.

Let d be a fixed positive integer. We seek for the space $\mathcal{V} \subset \mathcal{F}$ such that

$$\mathcal{V}_d \in \operatorname*{argmin}_{\mathcal{V} \subset \mathcal{F} \,:\, \dim(\mathcal{V}) \le d} \sum_{i=1}^n \|\phi(X_i) - \operatorname{Proj}_{\mathcal{V}} \phi(X_i)\|_{\mathcal{F}}^2 \,.$$

where $\mathrm{Proj}_{\mathcal{V}}$ denotes the orthogonal projection on $\mathcal{V}$ in the Hilbert space $\mathcal{F}$. In the following, we denote by $\mathcal{L}$ the linear map

$$\begin{aligned} \mathcal{L}\, &: \mathbb{R}^n \to \mathcal{F} \\ \alpha &\mapsto \mathcal{L}\alpha = \sum_{i=1}^n \alpha_i \phi(X_i)\,. \end{aligned}$$

3) Prove that $\mathcal{V}_d = \mathcal{L}V_d$, with V_d fulfilling

$$V_d \in \underset{V \subset \mathbb{R}^n \,:\, \dim(V) \leq d}{\operatorname{argmin}} \sum_{i=1}^n \|\phi\left(X_i\right) - \mathrm{Proj}_{\mathcal{L}V}\, \phi\left(X_i\right)\|_{\mathcal{F}}^2\,. \tag{2}$$

4) We denote by K the $n \times n$ matrix with entries $K_{i,j} = k(\phi(X_i), \phi(X_j))$ for $i, j = 1, \ldots, n$. We assume in the following that K is non-singular. Prove that for any $\alpha \in \mathbb{R}^n$, $\|\mathcal{L}K^{-1/2}\alpha\|_{\mathcal{F}}^2 = \|\alpha\|_2^2$.

5) Let V be a subspace of $\mathbb{R}^n$ of dimension d and denote by $(b_1, \ldots, b_d)$ an orthonormal basis of the linear span $K^{1/2}V$. Prove that $(\mathcal{L}K^{-1/2}b_1, \ldots, \mathcal{L}K^{-1/2}b_d)$ is an orthonormal basis of $\mathcal{L}V$.

6) Prove the identities

$$\begin{aligned} \mathrm{Proj}_{\mathcal{L}V}\, \mathcal{L}\alpha &= \sum_{k=1}^d \langle \mathcal{L}K^{-1/2}b_k, \mathcal{L}\alpha\rangle_{\mathcal{F}} \mathcal{L}K^{-1/2}b_k \\ &= \mathcal{L}K^{-1/2}\, \mathrm{Proj}_{K^{1/2}V}\, K^{1/2}\alpha\,. \end{aligned}$$

7) Let us denote $(e_1, \ldots, e_n)$ the canonical basis of $\mathbb{R}^n$. Check that

$$\begin{aligned} \sum_{i=1}^n \|\phi(X_i) - \mathrm{Proj}_{\mathcal{L}V}\, \phi(X_i)\|_{\mathcal{F}}^2 &= \sum_{i=1}^n \|\mathcal{L}e_i - \mathcal{L}K^{-1/2}\, \mathrm{Proj}_{K^{1/2}V}\, K^{1/2}e_i\|_{\mathcal{F}}^2 \\ &= \sum_{i=1}^n \|K^{1/2}e_i - \mathrm{Proj}_{K^{1/2}V}\, K^{1/2}e_i\|_2^2 \\ &= \|K^{1/2} - \mathrm{Proj}_{K^{1/2}V}\, K^{1/2}\|_F^2\,. \end{aligned}$$

8) Using Theorem 1, show that $V_d = \mathrm{Span}\{v_1, \ldots, v_d\}$, where $v_1, \ldots, v_d$ are eigenvectors of K associated to the d largest eigenvalues of K, is a minimizer of (2).

9) We set $f_k = \mathcal{L}K^{-1/2}v_k$ for $k = 1, \ldots, d$. Check that $(f_1, \ldots, f_d)$ is an orthonormal basis of $\mathcal{V}_d$ and

$$\mathrm{Proj}_{\mathcal{V}}\, \phi(X_i) = \sum_{k=1}^n \langle v_k, K^{1/2}e_i\rangle f_k\,.$$

Thus in the basis $(f_1, \ldots, f_k)$, the coordinates of the orthogonal projection of the point $\varphi(X_i)$ onto $\mathcal{V}$ are $(\langle v_1, K^{1/2}e_i\rangle, \ldots, \langle v_d, K^{1/2}e_i\rangle)$.

The take-home message is that the kernel PCA can be computed in the original space $\mathcal{X}$ as a eigenvalue decomposition of the Gram matrix K.

References

[1] Christophe Giraud. *Introduction to high-dimensional statistics.* Chapman and Hall/CRC, 2014.