Exercise 1. Refresher on matrices

1. Let $A$ be a $n \times d$ matrix with real entries. Show that $\text{Im}(A) = \text{Im}(AA^T)$.

   Solution.

   First note that $\text{Im}(AA^T) = \text{Im}(A^TA)$. Therefore, $\text{Ker}(A^TA) = \text{Ker}(A^T)$. And using that $\text{Ker}(B^T) = (\text{Im}(B))^\perp$, we deduce that $\text{Im}(AA^T)^\perp = \text{Im}(A)^\perp$, which concludes the proof.

2. Let $\{U_k\}_{1 \leq k \leq r}$ be a family of $r$ orthonormal vectors of $\mathbb{R}^d$. Show that $\sum_{k=1}^r U_kU_k^T$ is the matrix associated with the orthogonal projection onto $H = \{ \sum_{k=1}^r \alpha_k U_k : \alpha_1, \ldots, \alpha_r \in \mathbb{R} \}$. Deduce that if $A$ is a $n \times d$ matrix with real entries such that each column of $A$ is in $H$, then, $\left( \sum_{k=1}^r U_kU_k^T \right) A = A$.

   Solution.

   Let $\pi_H(X)$ be the orthogonal projection of $X$ onto $H$. Since $\{U_k\}_{1 \leq k \leq r}$ is an orthonormal basis of $H$, $\pi_H(X) = \sum_{k=1}^r <X, U_k > U_k = \left( \sum_{k=1}^r U_kU_k^T \right) X$.

   This implies that for each $X \in H$, $X = \left( \sum_{k=1}^r U_kU_k^T \right) X$. Since all the column vectors of $A$ are in $H$, this yields $\left( \sum_{k=1}^r U_kU_k^T \right) A = A$.

Exercise 2. Kernel Principal Component Analysis

Principal Component Analysis

Principal component analysis is a multivariate technique which aims at analyzing the statistical structure of high dimensional dependent observations by representing data using orthogonal variables called principal components. Reducing the dimensionality of the data is motivated by several practical reasons such as improving computational complexity. Let $(X_i)_{1 \leq i \leq n}$ be i.i.d. random variables in $\mathbb{R}^d$ and consider the matrix $X \in \mathbb{R}^{n \times d}$ such that the $i$-th row of $X$ is the observation $X_i^T$. In this exercise, it is assumed that data are preprocessed so that the columns of $X$ are centered. This means that for all $1 \leq i \leq d$, $\sum_{i=1}^n X_{i,k} = 0$. Let $\Sigma_n$ be the empirical covariance matrix:

$$\Sigma_n = n^{-1} \sum_{i=1}^n X_iX_i^T.$$ 

Principal Component Analysis aims at reducing the dimensionality of the observations $(X_i)_{1 \leq i \leq n}$ using a compression matrix $U \in \mathbb{R}^{d \times p}$ with orthonormal columns with $p \leq d$ so that for each $1 \leq i \leq n$, $U^TX_i$ is a low dimensional representation of $X_i$. The original observation may then be partially recovered using $U \in \mathbb{R}^{d \times p}$. Principal Component Analysis computes $U$ using the least squares approach:

$$U_\ast \in \arg \min_{U \in \mathbb{R}^{d \times p}} \sum_{i=1}^n \|X_i - UU^TX_i\|^2,$$
1. Prove that for all $\mathbb{R}^{n \times d}$ matrix $A$ with rank $r$, there exist $\sigma_1 \geq \ldots \geq \sigma_r > 0$ such that

$$A = \sum_{k=1}^{r} \sigma_k u_k v_k^T,$$

where $\{u_1, \ldots, u_r\} \subset \mathbb{R}^n$ and $\{v_1, \ldots, v_r\} \subset \mathbb{R}^d$ are two families of orthonormal vectors. The vectors $\{u_1, \ldots, u_r\}$ (resp. $\{v_1, \ldots, v_r\}$) are the left-singular (resp. right-singular) vectors associated with $\{\sigma_1, \ldots, \sigma_r\}$, the singular values of $A$.

Solution.

Since the matrix $AA^T$ is positive semidefinite, its spectral decomposition is given by

$$AA^T = \sum_{k=1}^{r} \lambda_k u_k u_k^T,$$

where $\lambda_1 \geq \ldots \geq \lambda_r > 0$ are the nonzero eigenvalues of $AA^T$ and $\{u_1, \ldots, u_r\}$ is an orthonormal family of $\mathbb{R}^n$. For all $1 \leq k \leq r$, define $v_k = \lambda_k^{-1/2}A^T u_k$ so that

$$\|v_k\|^2 = \lambda_k^{-1}(A^T u_k; A^T u_k) = \lambda_k^{-1} u_k^T A A^T u_k = 1,$$

$$A^T A v_k = \lambda_k^{-1/2} A^T A A^T u_k = \lambda_k u_k.$$

On the other hand, for all $1 \leq k \neq j \leq r$, $\langle v_k; v_j \rangle = \lambda_k^{-1/2} \lambda_j^{-1/2} u_k^T A A^T u_j = \lambda_k^{-1/2} \lambda_j^{-1/2} u_k^T u_j = 0$. Therefore, $\{v_1, \ldots, v_r\}$ is an orthonormal family of eigenvectors of $A^T A$ associated with the eigenvalues $\lambda_1 \geq \ldots \geq \lambda_r > 0$. Define, for all $1 \leq k \leq r$, $\sigma_k = \lambda_k^{1/2}$ which yields

$$\sum_{k=1}^{r} \sigma_k u_k u_k^T = \sum_{k=1}^{r} u_k u_k^T A = \left( \sum_{k=1}^{r} u_k u_k^T \right) A.$$

As $\{u_1, \ldots, u_r\}$ is an orthonormal family, $UU^T = \sum_{k=1}^{r} u_k u_k^T$ is the orthogonal projection onto the range($AA^T$) = range($A$) which implies

$$\sum_{k=1}^{r} \sigma_k u_k u_k^T = \left( \sum_{k=1}^{r} u_k u_k^T \right) A = A.$$

\[\square\]

If $U$ denotes the $\mathbb{R}^{n \times r}$ matrix with columns given by $\{u_1, \ldots, u_r\}$ and $V$ denotes the $\mathbb{R}^{d \times r}$ matrix with columns given by $\{v_1, \ldots, v_r\}$, then the singular value decomposition of $A$ may also be written as

$$A = UD_s V^T,$$

where $D_s = \text{diag}(\sigma_1, \ldots, \sigma_r)$. Then, $A^T A$ and $AA^T$ are positive semidefinite such that

$$A^T A = V D_s^2 V^T$$

and

$$AA^T = U D_s^2 U^T.$$

In the framework of this exercise, $n \Sigma_n = X^T X$ so that diagonalizing $n \Sigma_n$ is equivalent to computing the singular value decomposition of $X$.

2. Prove that solving the PCA least squares optimization problem boils down to computing

$$U_* \in \arg \max_{U \in \mathbb{R}^{n \times r}, U^T U = I_r} \{\text{trace}(U^T \Sigma_n U)\}.$$

Solution.
3. Let \( \{\vartheta_1, \ldots, \vartheta_p\} \) be orthonormal eigenvectors associated with the eigenvalues \( \lambda_1 \geq \ldots \geq \lambda_d \) of \( \Sigma_n \). Prove that a solution to this problem is given by the matrix \( U_n \) with columns \( \{\vartheta_1, \ldots, \vartheta_p\} \).

**Solution.**

Let \( \Sigma_n = V D_n V^T \) be the spectral decomposition of \( \Sigma_n \) where \( D_n = \text{Diag}(\lambda_1, \ldots, \lambda_d) \) and \( V \in \mathbb{R}^{d \times d} \) is a matrix with orthonormal columns \( \{\vartheta_1, \ldots, \vartheta_d\} \). For all \( U \in \mathbb{R}^{d \times p} \) matrix with orthonormal columns define \( B = V^T U \) so that, as \( V \in \mathbb{R}^{d \times d} \) is an orthogonal matrix,

\[
VB = V V^T U = U \quad \text{and} \quad U^T \Sigma_n U = B^T V^T V D_n V^T V B = B^T D_n B.
\]

Therefore,

\[
\text{Trace}(U^T \Sigma_n U) = \text{Trace}(B^T D_n B) = \sum_{i=1}^d \lambda_i \sum_{j=1}^p b_{i,j}^2.
\]

(1)

On the other hand,

\[
B^T B = U^T V V^T U = U^T U = I_p,
\]

so that the columns of \( B \) are orthonormal and

\[
\sum_{i=1}^d \sum_{j=1}^p b_{i,j}^2 = p.
\]

Hence, introducing for all \( 1 \leq i \leq d \), \( \alpha_i = \sum_{j=1}^p b_{i,j}^2 \), by (1),

\[
\text{Trace}(U^T \Sigma_n U) = \sum_{i=1}^d \alpha_i \lambda_i,
\]

with, for all \( 1 \leq i \leq d \), \( \alpha_i \in [0, 1] \) and \( \sum_{i=1}^d \alpha_i = p \). As \( \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_d \),

\[
\text{Trace}(U^T \Sigma_n U) \leq \sum_{i=1}^p \lambda_i.
\]

Indeed, the function \( f_d : (\alpha_1, \ldots, \alpha_d) \mapsto \sum_{i=1}^d \alpha_i \lambda_i \) is maximized under the constraints \( \alpha_i \in [0, 1] \) and \( \sum_{i=1}^d \alpha_i = p \) by \((\alpha_i^*)_{1 \leq i \leq d} \) such that \( \alpha_1^* = \ldots = \alpha_d^* = 1 \). Assume that \( (\alpha_1, \ldots, \alpha_d) \) is such that there exists \( 1 \leq j_0 \leq p \) such that \( \alpha_{j_0} < 1 \). Then, \( \sum_{j=p+1}^d \alpha_j \geq 1 - \alpha_{j_0} \) and we may write, as

\[
f_d : (\alpha_1, \ldots, \alpha_d) \leq \sum_{i=1, i \neq j_0}^p \alpha_i \lambda_i + \lambda_{j_0} + \sum_{i=p+1}^d \hat{\alpha}_i \lambda_i,
\]

where \((\hat{\alpha}_i)_{p+1 \leq i \leq d}\) are in \([0, 1]\) and such that \( \sum_{i=1, i \neq j_0}^p \alpha_i + 1 + \sum_{i=p+1}^d \hat{\alpha}_i = p \).

As the columns of \( U_n \) are \( \{\vartheta_1, \ldots, \vartheta_p\} \), for all \( 1 \leq i \leq d \) and \( 1 \leq j \leq p \), \( b_{i,j} = \langle \vartheta_i; \vartheta_j \rangle = \delta_{i,j} \). Therefore, for all \( 1 \leq i \leq d \), \( \sum_{j=1}^p b_{i,j}^2 = 1 \) and

\[
\text{Trace}(U_n^T \Sigma_n U_n) = \sum_{i=1}^p \lambda_i,
\]

which completes the proof. \( \square \)
4. For any dimension $1 \leq p \leq d$, let $\mathcal{F}_d^p$ be the set of all vector subspaces of $\mathbb{R}^d$ with dimension $p$. Consider the linear span $V_d$ defined as

$$V_p \in \arg\min_{V \in \mathcal{F}_d^p} \sum_{i=1}^n \|X_i - \pi_V(X_i)\|^2,$$

where $\pi_V$ is the orthogonal projection onto the linear span $V$. Prove that $V_1 = \text{span}\{v_1\}$ where

$$v_1 \in \arg\max_{v \in \mathbb{R}^d : \|v\| = 1} \sum_{i=1}^n (X_i, v)^2.$$

Solution.

Write $V_1 = \text{span}\{v_1\}$ for $v_1 \in \mathbb{R}^d$ such that $\|v_1\| = 1$. Then,

$$\sum_{i=1}^n \|X_i - \pi_{V_1}(X_i)\|^2 = \sum_{i=1}^n \|X_i - (X_i; v_1) v_1\|^2 = \sum_{i=1}^n \left(\|X_i\|^2 - 2\langle X_i; (X_i; v_1) v_1 \rangle + \|\langle X_i; v_1 \rangle v_1\|^2\right),$$

Consequently, $V_1$ is a solution if and only if $v_1$ is solution to:

$$v_1 \in \arg\max_{v \in \mathbb{R}^d : \|v\| = 1} \sum_{i=1}^n (X_i, v)^2.$$

5. For all $2 \leq p \leq d$, following the same steps, prove that a solution to the optimization problem is given by $V_p = \text{span}\{v_1, \ldots, v_p\}$ where

$$v_1 \in \arg\max_{v \in \mathbb{R}^d : \|v\| = 1} \sum_{i=1}^n (X_i, v)^2 \quad \text{and for all } 2 \leq k \leq p, \quad v_k \in \arg\max_{v \in \mathbb{R}^d : \|v\| = 1, v \perp v_1, \ldots, v_{k-1}} \sum_{i=1}^n (X_i, v)^2. \quad (2)$$

Solution.

Write $V_p = \text{span}\{v_1, \ldots, v_p\}$ where $\{v_1, \ldots, v_p\}$ is an orthonormal family. Then,

$$\sum_{i=1}^n \|X_i - \pi_{V_p}(X_i)\|^2 = \sum_{i=1}^n \|X_i - \sum_{k=1}^p \langle X_i; v_k \rangle v_k\|^2 = \sum_{i=1}^n \left(\|X_i\|^2 - \sum_{k=1}^p \langle X_i; v_k \rangle^2\right).$$

$(v_1, \ldots, v_p)$ is therefore solution to

$$v = (v_1, \ldots, v_p) \in \arg\max_{v \in \mathbb{R}^d : \|v\| = 1} \sum_{i=1}^n \sum_{k=1}^p (X_i; v_k)^2.$$

The additive form of the function to be maximized allows to build the orthonormal basis of $V_p$ sequentially as claimed.

6. Prove that the vectors $\{v_1, \ldots, v_k\}$ defined by (2) can be chosen as the orthonormal eigenvectors associated with the $k$ largest eigenvalues of the empirical covariance matrix $\Sigma_n$.

Solution.

Note that for all $v \in \mathbb{R}^d$ such that $\|v\| = 1$,

$$\frac{1}{n} \sum_{i=1}^n (X_i, v)^2 = \frac{1}{n} \sum_{i=1}^n (v^T X_i)(X_i^T v) = v^T \Sigma_n v.$$
As \((\vartheta_1)_{1 \leq i \leq d}\) are the orthonormal eigenvectors associated with the eigenvalues \(\lambda_1 \geq \ldots \geq \lambda_d \geq 0\) of \(\Sigma_n\). Then,
\[
\frac{1}{n} \sum_{i=1}^{n} \langle X_i, \vartheta \rangle^2 = v^T \left( \sum_{i=1}^{d} \lambda_i \vartheta_i \vartheta_i^T \right) v = \sum_{i=1}^{d} \lambda_i \langle v, \vartheta_i \rangle^2 \leq \lambda_1 \sum_{i=1}^{d} \langle v, \vartheta_i \rangle^2
\]
and, as \((\vartheta_i)_{1 \leq i \leq d}\) is an orthonormal basis of \(\mathbb{R}^d\), \(\sum_{i=1}^{d} \langle v, \vartheta_i \rangle^2 = \|v\|^2 = 1\). Therefore,
\[
\frac{1}{n} \sum_{i=1}^{n} \langle X_i, \vartheta \rangle^2 \leq \lambda_1.
\]

On the other hand, for all \(2 \leq i \leq d\), \(\langle \vartheta_1, \vartheta_i \rangle = 0\) and \(\langle \vartheta_1, \vartheta_1 \rangle = 1\) so that \(\sum_{i=1}^{d} \lambda_i \langle \vartheta_1, \vartheta_i \rangle^2 = \lambda_1\) which proves that \(\vartheta_1\) is solution to (2).

Assume now that \(v \in \mathbb{R}^d\) is such that \(\|v\| = 1\) and for all \(1 \leq j \leq k - 1\), \(\langle v, \vartheta_j \rangle = 0\) and write
\[
\frac{1}{n} \sum_{i=1}^{n} \langle X_i, \vartheta \rangle^2 = \sum_{i=1}^{d} \lambda_i \langle v, \vartheta_i \rangle^2 \leq \lambda_k \sum_{i=1}^{d} \langle v, \vartheta_i \rangle^2 \leq \lambda_k,
\]

since, as \((\vartheta_i)_{1 \leq i \leq d}\) is an orthonormal basis of \(\mathbb{R}^d\), \(\sum_{i=1}^{d} \langle v, \vartheta_i \rangle^2 = \|v\|^2 = 1\). On the other hand, for all \(1 \leq i \leq d\), \(i \neq k\), \(\langle \vartheta_k, \vartheta_i \rangle = 0\) and \(\langle \vartheta_k, \vartheta_k \rangle = 1\) so that \(\sum_{i=1}^{d} \lambda_i \langle \vartheta_k, \vartheta_i \rangle^2 = \lambda_k\) which proves that \(\vartheta_k\) is solution to (2).

Therefore, \(V_p = \text{span}\{\vartheta_1, \ldots, \vartheta_p\}\) is a solution to (2) and, as \((\vartheta_i)_{1 \leq i \leq p}\) is an orthonormal family, the projection matrix onto \(V_p\) is given by \(U_p U_p^T\) where \(U_p\) is a \(\mathbb{R}^{p \times p}\) matrix with columns \(\{\vartheta_1, \ldots, \vartheta_p\}\).

7. The orthonormal eigenvectors associated with the eigenvalues of \(\Sigma_n\) allow to define the principal components as follows. Then, as \(V_d = \text{span}\{\vartheta_1, \ldots, \vartheta_d\}\), for all \(1 \leq i \leq n\),
\[
\pi_{V_d}(X_i) = \sum_{k=1}^{d} \langle X_i, \vartheta_k \rangle \vartheta_k = \sum_{k=1}^{d} \langle X_i^T \vartheta_k \rangle \vartheta_k = \sum_{k=1}^{d} c_k(i) \vartheta_k,
\]
where for all \(1 \leq k \leq d\), the \(k\)-th principal component is defined as \(c_k = X \vartheta_k\). Prove that \((c_1, \ldots, c_d)\) are orthogonal vectors.

**Solution.**

The \(k\)-th principal component is the vector whose components are the coordinates of each \(X_i\), \(1 \leq i \leq n\), relative to the basis \(\{\vartheta_1, \ldots, \vartheta_d\}\) of \(V_d\). For all \(1 \leq i \neq j \leq d\),
\[
\langle c_i, c_j \rangle = \vartheta_i^T X^T X \vartheta_j = \vartheta_i^T (n \Sigma_n) \vartheta_j = n \lambda_i \vartheta_i^T \vartheta_j = 0,
\]
as \(\{\vartheta_1, \ldots, \vartheta_d\}\) is an orthonormal family.

**Application to RKHS**

Let \((X_i)_{1 \leq i \leq n}\) be \(n\) observations in a general space \(\mathcal{X}\) and \(k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}\) a positive kernel. \(\mathcal{W}\) denotes the Reproducing Kernel Hilbert Space associated with \(k\) and for all \(x \in \mathcal{X}\), \(\phi(x)\) denotes the function \(\phi(x) : y \to k(x, y)\). The aim is now to perform a PCA on \((\phi(X_1), \ldots, \phi(X_n))\). It is assumed that
\[
\sum_{i=1}^{n} \phi(X_i) = 0.
\]

Define
\[
K = (k(X_i, X_j))_{1 \leq i, j \leq n}.
\]

1. Prove that
\[
f_1 = \arg\max_{f \in \mathcal{W} : \|f\|_{\mathcal{W}} = 1} \sum_{i=1}^{n} \langle \phi(X_i), f \rangle_{\mathcal{W}}^2
\]
may be written
\[
f_1 = \sum_{i=1}^{n} \alpha_1(i) \phi(X_i), \quad \text{where} \quad \alpha_1 = \arg\max_{\alpha \in \mathbb{R}^n : \alpha^T K \alpha = 1} \alpha^T K^2 \alpha.
\]
Solution.

Any solution to the optimization problem lies in the vectorial subspace $V = \text{span}\{\phi(X_1), \ldots, \phi(X_n)\}$. Let $f = \sum_{i=1}^n \alpha(i)\phi(X_i)$ be such that $\|f\|_V = 1$. Then,

$$\|f\|_V^2 = \sum_{i,j=1}^n \alpha(i)\alpha(j)\langle \phi(X_i), \phi(X_j) \rangle_V = \alpha^T \mathbf{K} \alpha .$$

On the other hand, $\langle \phi(X_i), f \rangle_V = f(X_i) = [\mathbf{K}\alpha](i)$ so that,

$$\sum_{i=1}^n \langle \phi(X_i), f \rangle_V^2 = \sum_{i=1}^n f^2(X_i) = \sum_{i=1}^n (\mathbf{K}\alpha)(i)^2 = (\mathbf{K}\alpha_1)^T \mathbf{K}\alpha_1 = \alpha^T \mathbf{K} \alpha .$$

2. Prove that $\alpha_1 = \lambda_1^{-1/2}b_1$ where $b_1$ is the unit eigenvector associated with the largest eigenvalue $\lambda_1$ of $\mathbf{K}$.

Solution.

Let $\lambda_1 \geq \ldots \geq \lambda_n \geq 0$ be the eigenvalues of $\mathbf{K}$ associated with the orthonormal basis of eigenvectors $(b_1, \ldots, b_n)$. For any $\alpha \in \mathbb{R}^n$ such that $\alpha^T \mathbf{K}\alpha = 1$,

$$\alpha^T \mathbf{K}^2 \alpha = \alpha^T \left( \sum_{i=1}^n \lambda_i b_i b_i^T \right) \alpha = \sum_{i=1}^n \lambda_i^2 \langle a, b_i \rangle^2 \leq \lambda_1 \sum_{i=1}^n \lambda_i^2(a, b_i)^2 = \lambda_1 ,$$

as $\alpha^T \mathbf{K}\alpha = \sum_{i=1}^n \lambda_i(a, b_i)^2 = 1$. On the other hand,

$$\left( \lambda_1^{-1/2}b_1 \right)^T \mathbf{K}^2 \left( \lambda_1^{-1/2}b_1 \right) = \lambda_1^{-1} \sum_{i=1}^n \lambda_i^2(b_1, b_i)^2 = \lambda_1 .$$

Following the same steps, $f_j$ may be written $f_j = \sum_{i=1}^n \alpha_j(i)\phi(x_i)$ with $\alpha_j = \lambda_j^{-1/2}b_j$.

3. Write $H_d = \text{span}\{f_1, \ldots, f_d\}$. Prove that, for all $1 \leq i \leq n$,

$$\pi_{H_d}(\phi(X_i)) = \sum_{j=1}^d \lambda_j \alpha_j(i) f_j .$$

Solution.

Note first that the $(f_1, \ldots, f_d)$ is an orthonormal family. Therefore,

$$\pi_{H_d}(\phi(X_i)) = \sum_{j=1}^d \langle \phi(X_i), f_j \rangle_V f_j = \sum_{j=1}^d \langle \phi(X_i), \sum_{\ell=1}^d \alpha_j(\ell)\phi(X_\ell) \rangle_V f_j = \sum_{j=1}^d [\mathbf{K}\alpha_j](i) f_j .$$

Therefore,$$
\pi_{H_d}(\phi(x_i)) = \sum_{j=1}^d \lambda_j^{-1/2}[\mathbf{K}b_j](i) f_j = \sum_{j=1}^d \lambda_j^{1/2}b_j(i) f_j = \sum_{j=1}^d \lambda_j \alpha_j(i) f_j .$$