TD 2

Exercice 1 (Rappels de probabilités)
Soit $\mathbf{X} = (X_1, \dots, X_d)^\top$ un vecteur aléatoire en dimension d. On suppose que les deux premiers moments existent : $\mathbb{E}(\mathbf{X}) \in \mathbb{R}^d$ et $\mathbb{V}(\mathbf{X}) \in \mathbb{R}^{d\times d}$. Soit $\mathbf{a} \in \mathbb{R}^d$ et $A \in \mathbb{R}^{m\times d}$ des éléments déterministes (constantes). Simplifier au maximum les quantités suivantes :

1. $\mathbb{E}(\mathbf{a}^\top\mathbf{X})$ et $\mathbb{V}(\mathbf{a}^\top\mathbf{X})$.
2. $\mathbb{E}(A\mathbf{X})$ et $\mathbb{V}(A\mathbf{X})$.
3. Montrez que $\mathbb{V}(\mathbf{X})$ est toujours une matrice semi-définie positive.

Exercice 2 (PCA : cadre général)
L'énoncé de la PCA ci-dessus est limité : il est défini en fonction des observations (et de n). Le but de cet exercice est de le généraliser en remplaçant la moyenne empirique par la moyenne d'une variable aléatoire. Ainsi, la projection obtenue dépendra directement de la **distribution** de $\mathbf{X}$ et non pas des données. On considère un vecteur aléatoire $\mathbf{X}$ en dimension d. On définit la meilleure projection sur un sous-espace de dimension $k < d$ par :

$$\min_{\substack{E \\ \dim(E)=k}} \quad \mathbb{E}\left(\|\mathbf{X} - \mathrm{proj}_E(\mathbf{X})\|^2\right)$$

1. Montrez que pour $k = 1$, la solution est donnée par la droite dirigée par le vecteur propre associée à la plus grande valeur propre de $\mathbb{V}(\mathbf{X})$ noté par $\mathbf{p}_1$.
2. Comme dans l'étude, en supposant que l'on peut construire une base orthogonale $(\mathbf{p}_1, \dots, \mathbf{p}_k)$ de E de proche en proche, déterminez la solution pour $k > 1$.
3. Les axes de la base $(\mathbf{p}_1, \dots, \mathbf{p}_d)$ sont appelés des composantes principales. On pose $\mathbf{P}$ la matrice (de taille d x d) dont les colonnes sont les $\mathbf{p}_i$. Soit E un sous-espace de dimension k et $\mathbf{x} \in \mathbb{R}^d$. Montrez que les coordonnées de $\mathrm{proj}_E(\mathbf{x})$ dans la base $(\mathbf{p}_1, \dots, \mathbf{p}_k)$ sont données par $\mathbf{P}[:, :k]^\top\mathbf{x}$.
4. Comment peut-on estimer $\mathbf{P}$ en pratique à partir de n observations vectorielles $\mathbf{x}_1, \dots, \mathbf{x}_n$?
5. On note comme d'habitude X la matrice dont les lignes sont les observations $\mathbf{x}_1, \dots, \mathbf{x}_n$. Montrez que la matrice dont les lignes sont les projections $\mathrm{proj}_E(\mathbf{x}_i)$ est donnée par $X\mathbf{P}[:, :k]$.

Définition : PCA à $k \leq d$ composantes principales

Soit $\mathbf{P}$ la matrice des vecteurs propres de $\mathbb{V}(\mathbf{X})$. On définit la PCA à k composantes de $\mathbf{X}$ appliquée à un vecteur $\mathbf{x} \in \mathbb{R}^d$ par : $PCA_k(\mathbf{x}) = \mathbf{P}[:, :k]^\top\mathbf{x}$. Pour $k = d$, le problème d'optimisation a une solution triviale (projection = identité) qui n'est pas intéressante. On définit donc la PCA à d composante par : $PCA_d(\mathbf{x}) = \mathbf{P}^\top\mathbf{x}$ qui correspond à un changement de base orthonormale.

Exercice 3 (PCA et variance retenue)
On se place dans le cadre ci-dessus. Le but de cet exercice est de vous montrer un avantage de la PCA autre que la réduction de dimension : la maximisation de la variance retenue. On note le vecteur aléatoire en fonction de ses composantes : $\mathbf{X} = (\mathrm{X}_1, \dots, X_d)^\top$. La variance totale (théorique) est définie par $\mathrm{V}_{\text{totale}}(\mathbf{X}) \stackrel{\text{def}}{=} \sum_{i=1}^d \mathbb{V}(X_i) \in \mathbb{R}_+$. On définit le vecteur aléatoire $\mathbf{Y} = \mathrm{PCA}_d(\mathbf{X})$ de dimension d donné par la transformation PCA de l'exercice 2 avec $k = d$.

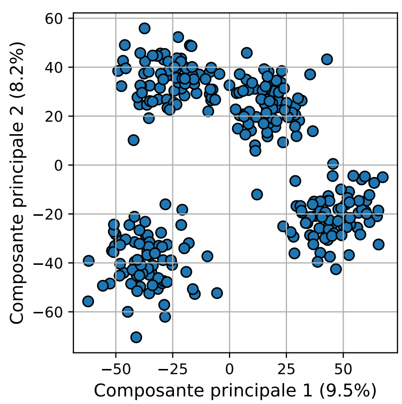
1. Calculez la variance $\mathbb{V}(\mathbf{Y})$. Que pouvez-vous en déduire concernant les composantes de $\mathbf{Y}$?
2. Comparez les variances totales $V_{\text{totale}}(\mathbf{X})$ et $V_{\text{totale}}(\mathbf{Y})$.
   On rappelle que pour toutes matrices A, B : trace(AB) = trace(BA).
3. On suppose à présent $1 \leq k \leq d$. Déterminez le pourcentage de variance totale retenue dans la projection en fonction des valeurs propres de $\mathbb{V}(\mathbf{X})$.

Exercice 4 (PCA comme maximisation de la variance)
Soit $\mathbf{X}$ un vecteur aléatoire dans $\mathbb{R}^d$ tel que $\mathbb{E}(\mathbf{X}) = 0$. Soit $\mathbf{q} \in \mathbb{R}^d_*$ avec $\|\mathbf{q}\| = 1$. On pose $Z \stackrel{\text{def}}{=} \mathbf{q}^\top \mathbf{X}$. On rappelle que la projection orthogonale sur Vect($\mathbf{q}$) est donnée par : $\mathbf{x} \mapsto \text{proj}_{\mathbf{q}}(\mathbf{x}) = \langle \mathbf{q}, \mathbf{x} \rangle \mathbf{q}$.

1. Déterminez $\mathbb{V}(Z)$.
2. Résoudre le problème d'optimisation : $\max_{\substack{\mathbf{q} \in \mathbb{R}^d_* \\ \|\mathbf{q}\|=1}} \mathbb{V}(Z)$
3. Comparez avec la transformation PCA_1.
4. Comment peut-on généraliser cette définition pour une PCA à k composantes ?

Exercice 5 (PCA comme décorrélation linéaire)
Soit $\mathbf{X}$ un vecteur aléatoire dans $\mathbb{R}^d$ tel que $\mathbb{E}(\mathbf{X}) = 0$. Trouver une matrice A telle que les composantes de $\mathbf{Y} \stackrel{\text{def}}{=} A\mathbf{X}$ ne soient pas corrélées. Commenter ?

Exercice 6 (Analyse de données et Machine learning (examen 2024))
On observe des données $\text{X} \in \mathbb{R}^{300 \times 100}$ dont les variables sont supposées centrées. On a également un vecteur de labels $\mathbf{y} \in \mathbb{R}^{300}$ contenant un label parmi $\{0, 1, 2, 3\}$. On souhaite développer un modèle de prédiction des labels.



1. Expliquez les étapes du calcul de la PCA.
2. On applique une PCA sur X et obtient la figure ci-dessus. Que représentent les pourcentages mentionnés dans les axes principaux ? Donnez leur formule.
3. Comment peut-on interpréter cette PCA ?
4. On utilise à présent une PCA pour réduire la dimension de 100 à 10 afin d'appliquer un modèle de Machine learning (une fonction de prédiction $f : \mathbf{x} \in \mathbb{R}^{10} \mapsto \mathbf{y} \in \{0, 1, 2\}$). La PCA + f sont obtenus en utilisant les données d'entrainement X et $\mathbf{y}$. On suppose que l'on a de nouvelles données $\mathbf{x}_{n+1}, \ldots \mathbf{x}_{n+5}$ non-vues lors de l'entrainement. On souhaite à présenter effectuer la prédiction des labels de ces nouvelles données. Comment doit-on procéder ?

Exercice 7 (Preuve de la SVD)
Soit $X \in \mathbb{R}^{n\times d}$ une matrice de rang r. On souhaite démontrer l'existence de la SVD de X. Formellement, on cherche à montrer que :

$$\exists \mathbf{U} \in \mathbb{R}^{n\times n}, \mathbf{V} \in \mathbb{R}^{d\times d}, \boldsymbol{\Sigma} \in \mathbb{R}^{n\times d} \text{ tels que } X = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$$

Avec $\mathbf{U} = \begin{pmatrix}\mathbf{u}_1 & \dots & \mathbf{u}_n\end{pmatrix}$, $\mathbf{V} = \begin{pmatrix}\mathbf{v}_1 & \dots & \mathbf{v}_d\end{pmatrix}$ deux matrices orthogonales et $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0)$ avec $\sigma_i > 0$.

1. Montrez que $X^\top X$ est une matrice de rang r.
2. Montrez que $X^\top X$ est symétrique semi-définie positive.
3. En déduire une base orthonormale $(\mathbf{v}_1, \dots, \mathbf{v}_d)$ de $\mathbb{R}^d$.
4. Définir une famille orthonormale $(\mathbf{u}_1, \dots, \mathbf{u}_r)$ de $\mathbb{R}^n$ et $\sigma_1, \dots, \sigma_r > 0$ tels que $X\mathbf{v}_i = \sigma_i \mathbf{u}_i$ pour $i = 1, \dots, r$.
5. Comment peut-on compléter les familles $(\mathbf{u}_1, \dots, \mathbf{u}_r)$ et $(\mathbf{v}_1, \dots, \mathbf{v}_r)$ pour obtenir une base orthonormale de $\mathbb{R}^n$ et $\mathbb{R}^d$?
6. En déduire la SVD de X et la SVD réduite de X vue en cours.

Exercice 8 (PCA via une SVD)
On se place dans le même cadre d'habitude : $\mathbf{X}$ est un vecteur aléaoire centré et $X \in \mathbb{R}^{n\times d}$ est la matrice des observations. Jusqu'à présent, nous avons vu que la PCA en pratique passe par la diagonalisation de la variance empirique $\hat{\Sigma} = \frac{1}{n}X^\top X$. Le but de cet exercice est de montrer que ceci n'est pas nécessaire.

1. Écrire la décomposition en valeurs singulières (SVD) réduite de X et en déduire $\hat{\Sigma}$.
2. On souhaite appliquer une PCA à k composantes. Quelle est la matrice $\mathbf{P}[:, :k]$ et le pourcentage de variance retenue en fonction des éléments de la SVD ?
3. Comment peut-on désormais facilement obtenir les projections données par $X\mathbf{P}[:, :k]$?
4. La complexité de la diagonalisation d'une matrice symétrique de taille $d \times d$ est $O(d^3)$. La complexité d'une SVD d'une matrice $(n \times d)$ est $O(nd\min(n, d))$. Quelle méthode faut-il choisir pour calculer la PCA ?

Exercice 9 (Normes matricielles)
Soit $\mathbf{A} \in \mathbb{R}^{n\times d}$. On peut généraliser la norme Euclidienne (ℓ_2) aux matrices en prenant par exemple la norme de Frobenius définie par $\|\mathbf{A}\|_F^2 = \sum_{i,j} A_{ij}^2$. On peut également voir la matrice comme un opérateur linéaire de $\mathbb{R}^d$ dans $\mathbb{R}^n$ et définir la norme d'opérateur (aussi appelée norme spectrale, ou norme 2) de $\mathbf{A}$ par $\|\mathbf{A}\|_{op} = \max_{\|\mathbf{x}\|\neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}$. Le but de cet exercise est de montrer que ces normes sont facilement calculables à partir de la SVD de $\mathbf{A}$. On note les valeurs singulières de $\mathbf{A}$ par $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_{\min(n,d)}$.

1. Montrez que $\|\mathbf{A}\|_F^2 = \mathrm{trace}(\mathbf{A}^\top\mathbf{A})$.
2. En déduire que $\|\mathbf{A}\|_F^2 = \sum_{i=1}^{\min(n,d)} \sigma_i^2$.
3. Montrez que $\|\mathbf{A}\|_{op} = \sigma_1$.

On peut généraliser ces normes en définissant la Schatten-norme d'ordre p par $\|\mathbf{A}\|_{Sp} = \left(\sum_{i=1}^{\min(n,d)} |\sigma_i|^p\right)^{1/p}$. En particulier, avec $p = 1$ on obtient la norme nucléaire de $\mathbf{A}$ définie par $\|\mathbf{A}\|_* = \sum_{i=1}^{\min(n,d)} |\sigma_i|$ qui est souvent utilisée en optimisation, en particulier pour les systèmes de recommendation.