Sur l'utlisation de la matrice de Gram pour l'analyse en composantes principales fonctionnelles multivariées

Steven Golovkine[1] & Edward Gunning[1] & Andrew J. Simpkin[3] & Norma Bargary[1]

[1] *University of Limerick, Ireland, name.surname@ul.ie*
[3] *University of Galway, Ireland, andrew.simpkin@nuigalway.ie*

Résumé. En utilisant la dualité entre les espaces des lignes et des colonnes de la matrice de données, nous proposons d'estimer les éléments propres d'un ensemble de données fonctionnelles multivariées en utilisant sa matrice de Gram. Ces réalisations aléatoires peuvent éventuellement être multidimensionnelles, par exemple des surfaces. Nous développons les formules de passage entre les éléments propres de l'opérateur de covariance et ceux de la matrice de Gram. Cette relation permet de choisir la méthode la plus adaptée au problème. Nous donnons des recommandations quant à l'utilisation de l'opérateur de covariance ou de la matrice de Gram pour l'estimation des éléments propres de la matrice de données basées sur un calcul de complexité des algorithmes et de simulations.

Mots-clés. Données fonctionnelles multivariées, Analyse en composantes principales fonctionnelles multivariées, Matrice de Gram, Réduction de dimension

Abstract. Using the duality relation between the row and column spaces of a data matrix, we propose to estimate the eigenvalues and eigenfunctions of a set of multivariate functional data by using its Gram matrix between the curves. The random functional data realizations could be multidimensional, such as surfaces. We develop a formula to convert the eigenvalues and eigenfunctions of the covariance operator to the eigenvalues and eigenvectors of the Gram matrix. This relationship allows for choosing the most suitable method for the problem. We provide recommendations regarding the use of the covariance operator or the Gram matrix for estimating the eigencomponents of the data matrix based on a derivation of algorithmic complexity and simulations.

Keywords. Dimension reduction, Gram matrix, Multivariate functional data, Multivariate functional principal components analysis,

1 Introduction

Un intérêt croissant est porté sur la modélisation de données enregistrées sous forme de mesures discrètes à travers le temps. L'analyse de données fonctionnelles (ADF) considère ces données comme des réalisations d'un processus stochastique, éventuellement enregistré avec erreur à des instants aléatoires. Avec la démocratisation des capteurs, de plus en plus

de données peuvent être modélisées comme des données fonctionnelles. La dimension des données récoltées augmente rapidement, que ce soit en entrée, comme les imageries par résonance magnétique fonctionnelles par exemple, ou bien en sortie, comme en recherche sur le mouvement humain. L'analyse en composante principale fonctionnelle est donc un outil intéressant permettant de réduire la dimension de ces données et ainsi permettre leur analyse.

Supposons un échantillon de N courbes provenant d'un même processus aléatoire. Nous proposons d'utiliser la dualité entre les lignes et les colonnes d'une matrice de données pour calculer les éléments propres de cet échantillon de N courbes, permettant de réduire la dimension des données. Cette dualité a été beaucoup étudié dans le cadre de données représentés par des vecteurs de $\mathbb{R}^N$ (e.g. Escofier (1979), Härdle et Simar (2003), Pagès (2004)). Cependant, depuis l'article de Ramsay (1982), cette relation n'a pratiquement pas été étudié dans le cadre de données fonctionnelles. Benko, Härdle et Kneip (2009) utilisent cette dualité pour calculer les éléments propres d'un ensemble de courbes univariés et comparer les courbes moyennes de deux groupes.

2 Modèle

La structure de nos données consiste en des trajectoires indépendantes d'un processus stochastique $X = \left(X^{(1)}, \ldots, X^{(P)}\right)^\top, P \geq 1$. Pour chaque $1 \leq p \leq P$, soit $\mathcal{T}_p = [0,1]^{d_p}, d_p \geq 1$. Chaque coordonnée $X^{(p)} : \mathcal{T}_p \to \mathbb{R}$ est supposée appartenir à $\mathcal{L}^2(\mathcal{T}_p)$ muni du produit scalaire usuel, noté $\langle \cdot, \cdot \rangle_2$. Ainsi, X est un processus stochastique indexé par $\mathbf{t} = (t_1, \ldots, t_P)$ appartenant à $\mathcal{T} \coloneqq \mathcal{T}_1 \times \cdots \times \mathcal{T}_P$ et prenant ses valeurs dans $\mathcal{H} \coloneqq \mathcal{L}^2(\mathcal{T}_1) \times \cdots \times \mathcal{L}^2(\mathcal{T}_P)$. Considérons la fonction $\langle\!\langle \cdot, \cdot \rangle\!\rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$,

$$\langle\!\langle f, g \rangle\!\rangle \coloneqq \sum_{p=1}^{P} \langle f^{(p)}, g^{(p)} \rangle_2, \quad f, g \in \mathcal{H}.$$

Happ et Greven (2018) montrent que $\mathcal{H}$ est un espace de Hilbert pour le produit scalaire $\langle\!\langle \cdot, \cdot \rangle\!\rangle$. Notons $|\!|\!| \cdot |\!|\!|$, la norme induite par $\langle\!\langle \cdot, \cdot \rangle\!\rangle$. Soit $\mu : \mathcal{T} \to \mathcal{H}$ la fonction moyenne du processus X, $\mu(\mathbf{t}) \coloneqq \mathbb{E}(X(\mathbf{t})), \mathbf{t} \in \mathcal{T}$. Soit C la fonction de covariance défini, pour $\mathbf{s}, \mathbf{t} \in \mathcal{T}$, par

$$C(\mathbf{s}, \mathbf{t}) \coloneqq \mathbb{E}\left(\{X(\mathbf{s}) - \mu(\mathbf{s})\}\{X(\mathbf{t}) - \mu(\mathbf{t})\}^\top\right), \quad \mathbf{s}, \mathbf{t} \in \mathcal{T}.$$

Plus précisement, pour $1 \leq p, q \leq P$, l'entrée (p, q) de la matrice $C(\mathbf{s}, \mathbf{t})$ est la fonction de covariance entre les composantes p et q du processus X:

$$C_{p,q}(s_p, t_q) \coloneqq \mathbb{E}\left(\{X^{(p)}(s_p) - \mu^{(p)}(s_p)\}\{X^{(q)}(t_q) - \mu^{(q)}(t_q)\}\right), \quad s_p \in \mathcal{T}_p, t_q \in \mathcal{T}_q.$$

Soit $\Gamma : \mathcal{H} \to \mathcal{H}$, l'opérateur de covariance de X, défini comme l'opérateur intégral de noyau C. C'est-à dire que, pour $f \in \mathcal{H}$ et $\mathbf{t} \in \mathcal{T}$, la composante p de $\Gamma f(\mathbf{t})$ est donnée

par

$$(\Gamma f)^{(p)}(t_p) \coloneqq \langle\langle C_{p,\cdot}(t_p, \cdot), f(\cdot)\rangle\rangle = \langle\langle C_{\cdot,p}(\cdot, t_p), f(\cdot)\rangle\rangle, \quad t_p \in \mathcal{T}_p.$$

Notons $\mathcal{X} = \{X_1, \ldots, X_n, \ldots, X_N\}$ un échantillon aléatoire du processus X avec des trajectoires continues. Nous supposons que les courbes sont observées sans erreur. L'ensemble de données peut être vu comme une matrice avec N lignes et P colonnes où chaque entrée est une courbe, eventuellement multidimensionelle. Chaque ligne de cette matrice représente une observation; alors que chaque colonne représente une variable fonctionnelle. À l'intersection de la ligne n et de la colonne p, nous avons donc $X_n^{(p)}$, ce qui correspond à la composante p de la courbe de l'individu n.

Notons $\{\lambda_k\}_{k\geq 1}$, l'ensemble des valeurs propres, telle que $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ et $\Phi = \{\phi_k\}_{k\geq 1}$, l'ensemble des fonctions propres de l'opérateur de covariance Γ. L'ensemble Φ contient un nombre infini d'éléments et forme une base orthonormale complète de $\mathcal{H}$:

$$\Gamma\phi_k = \lambda_k\phi_k, \quad k \geq 1.$$

Utilisant le théoreme de Karhunen-Loève multivarié (voir, e.g., Happ et Greven, 2018), nous obtenons la décomposition suivante pour chaque observation:

$$X_n(\mathbf{t}) = \mu(\mathbf{t}) + \sum_{k=1}^{\infty} \mathfrak{c}_{nk}\phi_k(\mathbf{t}), \quad \mathbf{t} \in \mathcal{T},$$

où les coefficients $\mathfrak{c}_{nk} = \langle\langle X_n - \mu, \phi_k\rangle\rangle$ sont les projections des courbes centrées sur les fonctions propres. Les coefficients $\mathfrak{c}_{nk}$ ont les propriétés suivantes: $\mathbb{E}(\mathfrak{c}_{nk}) = 0$, $\mathbb{E}(\mathfrak{c}_{nk}^2) = \lambda_k$ et $\mathbb{E}(\mathfrak{c}_{nk}\mathfrak{c}_{nr}) = 0$ for $k \neq r$.

3 MFPCA

Happ et Greven (2018) proposent d'estimer les éléments propres de l'opérateur de covariance Γ en diagonalisant chaque opérateur univarié $\Gamma^{(p)}$, puis en combinant les projections des courbes univariées $X_n^{(p)}$ sur les fonctions propres univariées pour construire les fonctions propres multivariées.

Nous proposons d'utiliser la dualité entre les espaces définis par les lignes et les colonnes de la matrice de données pour estimer les éléments propres de l'opérateur de covariance. Considérons la matrice des produits scalaires M avec des entrées

$$M_{ij} = \langle\langle X_i - \mu, X_j - \mu\rangle\rangle, \quad i, j = 1, \ldots, N.$$

Notons $l_k, 1 \leq k \leq N$, l'ensemble des valeurs propres, telles que $l_1 \geq \cdots \geq l_N \geq 0$, et $v_k, 1 \leq k \leq N$, l'ensemble des vecteurs propres de la matrice M. La relation entre les valeurs propres non nulles de l'opérateur de covariance Γ et les valeurs propres de M est donnée par

$$\lambda_k = \frac{l_k}{N}, \quad k = 1, 2, \ldots, N.$$

De plus, la relation entre les fonctions propres multivariées de l'opérateur de covariance Γ et les vecteurs propres (orthonormés) de M est donnée par

$$\phi_k(\mathbf{t}) = \sum_{n=1}^{N} v_{nk} \left\{ X_n(\mathbf{t}) - \mu(\mathbf{t}) \right\}, \quad \mathbf{t} \in \mathcal{T}, \quad k = 1, 2, \ldots, N,$$

où v_{nk} est la n-ème entrée du vecteur v_k. Les scores, définis comme le produit scalaire entre les courbes et les fonctions propres, sont ensuite calculés par

$$\mathfrak{c}_{nk} = \sqrt{l_k} v_{nk}, \quad n = 1, 2, \ldots, N, \quad k = 1, 2, \ldots, N.$$

3.1 Complexité computationnelle

Nous décrivons la complexité algorithmique pour le calcul des éléments propres d'un ensemble de données fonctionnelles multivariées en utilisant l'opérateur de covariance et la matrice des produits scalaires. Soit N courbes de P composantes, nous supposons que toutes les observations de la composante p sont échantillonnées sur une grille régulière de M_p points. Pour $a \in \mathbb{N}$, notons $M^a = \sum_p M_p^a$. Notons K le nombre d'éléments propres à estimer. Notons K_p le nombre d'éléments propres pour la composante p, et $K = \sum_p K_p$. Nous supposons que les courbes sont parfaitement observées, et donc, aucune étape de lissage n'est nécessaire.

La complexité de l'estimation des éléments propres par décomposition de l'opérateur de covariance, utilisant la méthode de Happ et Greven (2018), est donnée par :

$$\mathcal{O}\left(NM^2 + M^3 + N\sum_{p=1}^{P} M_p K_p + NK^2 + K^3 + K\sum_{p=1}^{P} M_p K_p + NK^2 \right).$$

La complexité de l'estimation des éléments propres par décomposition de la matrice des produits scalaires est donnée par :

$$\mathcal{O}\left(N^2 M^1 + N^3 + KNP + KN \right).$$

Généralement, le nombre de composantes à estimer K est très petit par rapport au nombre de courbes N et au nombre de points d'échantillonnage M^1. Les complexités peuvent donc être réduites à $\mathcal{O}(NM^2 + M^3)$ et $\mathcal{O}(N^2 M^1 + N^3)$. Ainsi, si le nombre d'observations est grand, il est préférable d'estimer les éléments propres en utilisant l'opérateur de covariance, alors que si le nombre de points d'échantillonnage est grand, il est plus intéressant d'utiliser la matrice de Gram. Nous avons également remarqué cela lors des simulations.

4 Simulation

Nous comparons les méthodologies sur deux exemples. Pour le premier exemple, nous reprenons le processus de Happ et Greven (2018) pour des données fonctionnelles multivariées définies sur des domaines unidimensionnels. Pour cela, nous utilisons une base

de Wiener de 10 fonctions sur $[0, 10]$ et nous coupons cet intervalle en P parties pour générer la base de fonctions multivariées. Les courbes sont ensuite générées en utilisant le théorème de Karhunen-Loève multivarié avec les coefficients simulés suivant une loi normale de moyenne 0 et de variance λ_k avec une décroissance exponentielle de λ_k lorsque k augmente. Nous considérons $P = 2, 10$ et 20 pour la simulation. Pour le deuxième exemple, nous considérons des données fonctionnelles univariées ($P = 1$) mais multidimensionnelles ($\mathcal{T} = [0, 1]^2$). Les éléments de Φ sont construits avec un produit tensoriel des cinq premières fonctions de la base de fonctions de Fourier. Les surfaces sont générées en utilisant le théorème de Karhunen-Loève avec les coefficients simulés suivant une loi normale de moyenne 0 et de variance λ_k avec une décroissance exponentielle de λ_k lorsque k augmente. Pour les deux exemples, les valeurs testées pour N et M sont $25, 50, 75, 100$. Chaque courbe est observée sur une grille de M points équidistants et chaque surface est observée sur une grille de $M \times M$ points. Les simulations sont lancées 500 fois sur un MacBook Pro M1 (2021).

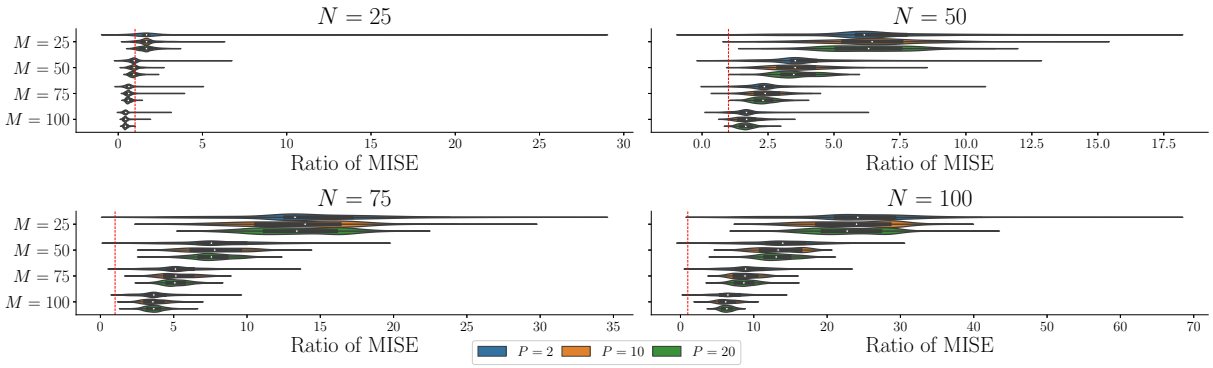
Nous avons estimé les $K = 5$ premières composantes principales pour les deux exemples. Pour le premier exemple, nous avons comparé les décompositions de l'opérateur de covariance et de la matrice de Gram. Pour le deuxième exemple, nous avons comparé la décomposition de la matrice de Gram à la décomposition par l'algorithme FCP-TPA (Allen, 2013). Soient deux ensembles de courbes $\mathcal{X}$ et $\mathcal{Y}$. Nous avons considéré la qualité de reconstruction des courbes en calculant l'erreur de reconstruction :

$$\text{MISE}(\mathcal{X}, \mathcal{Y}) = \frac{1}{N} \sum_{i=1}^{N} \sum_{p=1}^{P} \int_{\mathcal{T}_p} \left\{ X_i^{(p)}(t) - Y_i^{(p)}(t) \right\}^2 \mathrm{d}t.$$

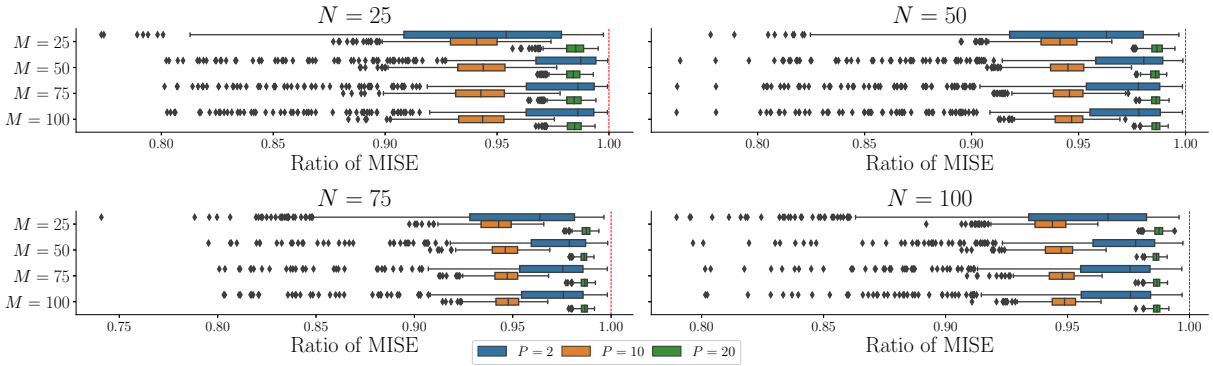
Pour comparer les méthodologies, nous avons calculé le ratio $\text{MISE}(\mathcal{X}, \widehat{\mathcal{X}})/\text{MISE}(\mathcal{X}, \widetilde{\mathcal{X}})$ où $\mathcal{X}$ est l'ensemble des observations, $\widehat{\mathcal{X}}$ est la reconstruction des observations en utilisant la matrice de Gram et $\widetilde{\mathcal{X}}$ est la reconstruction des observations en utilisant l'opérateur de covariance. Les résultats pour le premier exemple sont donnés dans la Figure 1 et ceux pour le deuxième exemple sont donnés dans la Figure 2.

La décomposition de l'opérateur de covariance semble plus rapide pour des données multivariées mais unidimensionnelles (Figure 1a) dans la majorité des cas. La diagonalisation de la matrice de Gram est plus rapide lorsque $M \gg N$, et le nombre de composantes P ne semble pas avoir d'influence particulière sur le temps de calcul (ce qui est cohérent avec la complexité computationnelle). Pour un nombre de composantes fixé (K fixé), l'erreur de reconstruction est légèrement plus faible en utilisant la décomposition de la matrice de Gram (Figure 1b) pour tous les cas considérés.

Nous remarquons que pour des données multidimensionnelles, la décomposition de la matrice de Gram est plus rapide que l'algorithme FCP-TPA sauf pour $N = 100$ (Figure 2a). Ce résultat est dû au nombre de points d'échantillonnage pour chaque observation ($M \times M$). Pour un nombre de composantes fixes (K fixé), l'erreur de reconstruction est bien plus faible avec la matrice de Gram (Figure 2b). Ce résultat peut s'expliquer par le



(a) Temps de calcul



(b) Erreur de reconstruction

Figure 1: Résultats de la première simulation.

fait que l'algorithme FCP-TPA est itératif avec une étape d'optimisation, tandis que la décomposition de la matrice de Gram a une solution analytique.

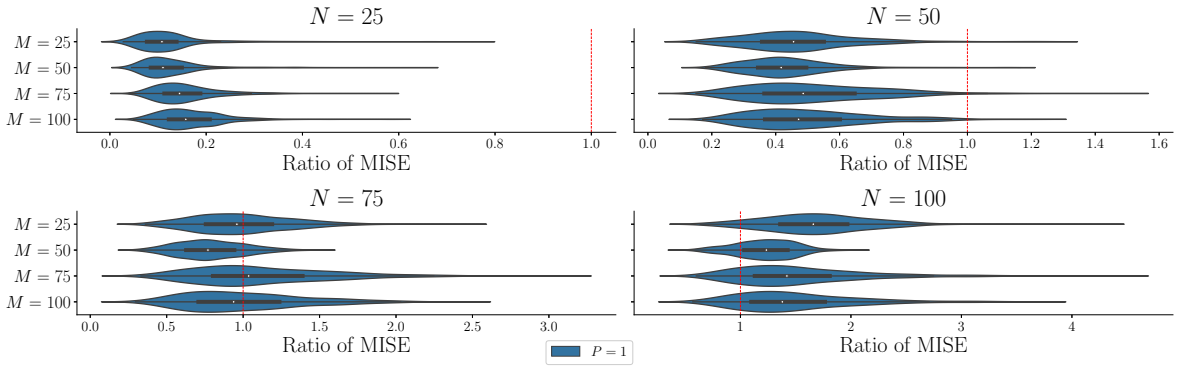
En conclusion, il semble intéressant d'utiliser la matrice de Gram lorsque les données sont multidimensionnelles (surfaces), peu importe le nombre d'observations, de points d'échantillonnage et de composantes. Lorsqu'il n'y a que des données unidimensionnelles, la décomposition de l'opérateur de covariance semble plus adaptée.

Bibliographie

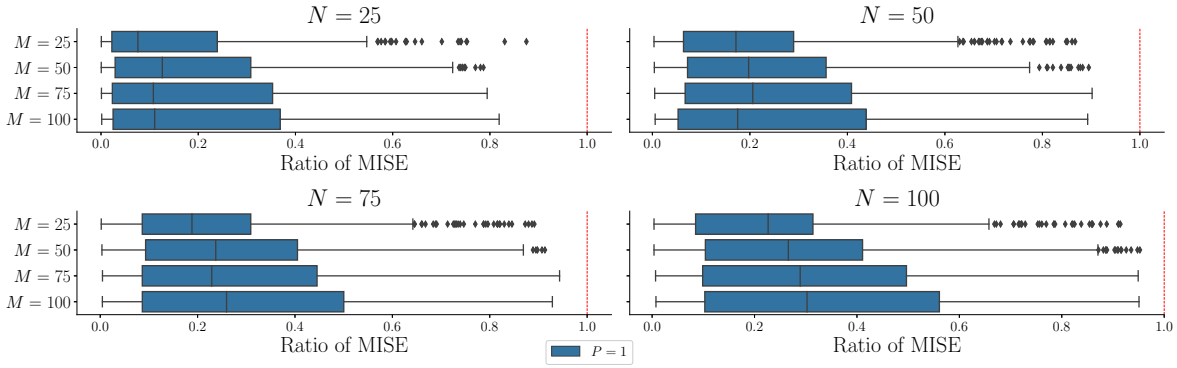
Allen, G. I. (2013). Multi-way functional principal components analysis. In 2013, *5th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*.

Benko, M., Härdle, W. and Alois Kneip (2009). Common functional principal components. *The Annals of Statistics*, 37(1) 1-34.

Escofier, B. (1979). Traitement simultané de variables qualitatives et quantitatives en



(a) Temps de calcul



(b) Erreur de reconstruction

Figure 2: Résultats de la deuxième simulation.

analyse factorielle. *Cahiers de l'analyse des données*, 4(2):137–146, 1979.

Happ, C. and Greven, S. (2018). Multivariate Functional Principal Component Analysis for Data Observed on Different (Dimensional) Domains. *Journal of the American Statistical Association*, 113 649-659.

Härdle, W. and Simar, L. (2003). *Applied Multivariate Statistical Analysis*. Springer, Berlin.

Pagès, J. (2004). Analyse factorielle de données mixtes. *Revue de Statistique Appliquée* 52, 93–111.

Ramsay, J. O. (1982). When the data are functions. *Psychometrika*, 47(4):379–396.