Décomposition PARAFAC pour données longitudinales

Lucas Sort, Fabien Girka, Laurent Le Brusquet, Arthur Tenenhaus

Décomposition PARAFAC pour données longitudinales.

Lucas Sort [1], Fabien Girka [1], Laurent Le Brusquet [1] & Arthur Tenenhaus [1]

[1] *Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, Gif-sur-Yvette, 91190, France, lucas.sort@l2s.centralesupelec.fr*

Résumé. Face à des données tensorielles, la décomposition PARAFAC permet de comprendre la structure des données. Lorsque l'un (ou plusieurs) des axes du tenseur est une grandeur continue (comme le temps ou une grandeur spatiale) et que les données sont échantillonnées de manière irrégulière, PARAFAC ne s'applique plus directement. Cet article propose une extension de PARAFAC pour des données dépendant de 3 axes *individu* × *grandeur* × *temps*. Chaque individu est ainsi décrit par un ensemble de variables longitudinales avec pour chacune d'elle un échantillonnage qui lui est propre.

La démarche proposée permet une reconstruction des facteurs canoniques associés à l'axe *grandeur*, des fonctions canoniques associées à l'axe longitudinal (*temps*), et du vecteur de scores associé à chaque individu. Elle est fondée sur (i) la reformulation du critère optimisé classiquement dans PARAFAC pour qu'il ne dépende plus que des fonctions de covariance du processus temporel, (ii) une estimation par agrégation et lissage de ces fonctions de covariance à partir d'observations bruitées et échantillonnées irrégulièrement, et (iii) un calcul du vecteur de scores associé à chaque individu reposant sur une modélisation gaussienne de ce vecteur.

La démarche est testée sur des données simulées. Il est montré que dans le contexte d'un échantillonnage irrégulier et très parcimonieux, il est possible de conserver une qualité d'estimation des facteurs et des fonctions canoniques en observant un plus grand nombre d'individus.

Mots-clés. PARAFAC, données longitudinales, données tensorielles, échantillonnage irrégulier.

Abstract. When faced with tensor data, the PARAFAC decomposition allows an understanding of the structure of the data. When one (or more) of the tensor axes is a continuous quantity (such as time or a spatial quantity) with irregularly sampled data, PARAFAC is not directly applicable. This paper proposes an extension of PARAFAC for data depending on 3 axes : *individual* × *magnitude* × *time*. A set of longitudinal measurements with a variable-and-subject-specific sampling describes each individual. The proposed approach allows a reconstruction of the canonical factors associated with the *magnitude* axis, the canonical functions associated with the longitudinal axis (*time*), and the score vector associated with each individual. It is based on (i) the reformulation of the criterion classically optimized in PARAFAC so that it depends only on the covariance functions of the temporal process, (ii) an estimation by aggregation and smoothing of these covariance functions from noisy and irregularly sampled observations, and (iii) a computation of the score vector associated

to each individual based on Gaussian modeling of this vector. We validate the approach on simulated data. We show that, in the context of irregular and very parsimonious sampling, it is possible to keep the quality of estimation of the factors and the canonical functions by observing a larger number of individuals.

Keywords. PARAFAC, longitudinal data, tensor data, irregular sampling.

1 Introduction

Plusieurs méthodes de décomposition tensorielle existent. Dans un cadre non supervisé, les plus classiques sont PARAFAC (Harshman et al., 1970; Carroll and Chang, 1970) et Tucker (Tucker, 1966). Dans un contexte où l'une des dimensions du tenseur correspond à une dimension *individu*, ces méthodes extraient les facteurs associés à chaque dimension ce qui permet de nombreuses utilisations (i) réduction de la dimension (ii) interprétation des facteurs liés aux variables observées permettant ainsi d'appréhender les variables à l'origine des plus grandes variations entre individus (iii) calcul d'un vecteur de scores associé à chaque individu ce qui constitue une étape préalable aux approches ayant par exemple pour objectif la visualisation des données ou le clustering.

Prenons l'exemple de grandeurs médicales acquises pour différents individus. Sauf dans le cas d'une cohorte spécialement constituée pour un essai clinique avec un protocole rigoureux de collecte des données, il est fréquent que les instants d'acquisition varient d'un individu à l'autre et également d'une grandeur à l'autre. Il est aussi possible que certaines observations soient manquantes et l'échantillonnage apparait alors irrégulier. Ces données, dites *longitudinales*, correspondent à l'acquisition de grandeurs évoluant au cours du temps, c'est-à-dire de fonctions. De nombreuses méthodes d'analyse fonctionnelle existent (Ramsay and Silverman, 2006; Shang, 2014), permettant ainsi de dépasser le contexte des données structurées en tableaux ou tenseurs.

Ce papier présente une version fonctionnelle de PARAFAC, permettant ainsi de traiter le cas de grandeurs acquises à des instants différents d'un individu à l'autre, et d'une grandeur à l'autre. Cette version fonctionnelle repose principalement sur une reformulation de PARAFAC ne dépendant plus des données, mais uniquement des fonctions de covariance (section 2). Ces dernières sont estimées par une technique d'agrégation et de lissage (section 3). La section 4 présente la résolution du problème reformulé. La reformulation à l'aide des fonctions de covariance ne dépendant plus des scores associés à chaque individu, un calcul spécifique leur est dédié (section 5). Enfin, la démarche est testée sur des données simulées (section 6) avec différents niveaux d'irrégularité pour l'échantillonnage.

2 Description des données et modèle sous-jacent

Dans un souci de simplicité, la méthode est décrite dans le contexte d'un tenseur d'ordre 3 où seul un des trois axes est échantillonné de manière irrégulière. C'est le cas où, par exemple, pour n individus (axe 1), K grandeurs (axe 2) ont été observées au cours du temps (axe 3). Les instants d'observation peuvent être propres à chaque individu et à chaque grandeur. La méthode s'étend facilement à des fonctions définies sur des ensembles de dimensions quelquonques (par exemple des fonctions spatio-temporelles acquises à des positions et des instants variables d'un individus à un autre).

Notons i l'indice correspondant à l'individu $i \in \{1 \dots n\}$ et k l'indice correspondant aux grandeurs observées $k \in \{1 \dots K\}$. Pour un individu et une grandeur donnés, notons $t_{i,k}$ le vecteur dont les éléments sont les $q_{i,k}$ instants d'observation. Les éléments $t_{i,k}$ sont supposés appartenir à l'intervalle temporel $[0, T]$. Les 3 axes liés aux indices i, k et aux instants d'observation seront désignés par les trois modes *individu*, *grandeur* et *temps*.

Soit $Y_{i,k}$ le vecteur de longueur $q_{i,k}$ correspondant aux valeurs acquises pour la grandeur k de l'individu i. On suppose que $Y_{i,k}$ est l'observation bruitée d'une trajectoire temporelle (fonction) $t \mapsto X_{i,k}(t)$.

Le bruit d'observation est supposé additif :

$$Y_{i,k,j} = X_{i,k}\left(t_{i,k,j}\right) + \varepsilon_{i,k,j}, \qquad 1 \leq j \leq q_{i,k}$$

avec les $\varepsilon_{i,k,j}$ IID de loi $\mathcal{N}(0, \sigma_k^2)$.

A l'instar de la décomposition PARAFAC, on choisit de :

— modéliser les trajectoires $t \mapsto X_{i,k}(t)$ par une décomposition de rang R, $R \geq 1$:

$$X_{i,k}(t) \quad \approx \sum_{r=1}^{R} z_i^r f_k^r \phi^r(t)$$

où $z_i = (z_i^1, \dots, z_i^R) \in \mathbb{R}^{1 \times R}$ est le vecteur de scores de l'individu i, $\phi^1, \dots, \phi^R$ les fonctions canoniques $\mathbb{R} \to \mathbb{R}$ liées à l'axe longitudinal, et $f^1, \dots, f^R \in \mathbb{R}^{K \times 1}$ les facteurs canoniques liés aux K grandeurs.

— estimer les scores, les facteurs et les fonctions canoniques en cherchant à minimiser le risque :

$$\min_{\substack{z_1 \dots z_n \\ f^1 \dots f^R \\ \phi^1 \dots \phi^R}} \quad \sum_{i=1}^{n} \sum_{k=1}^{K} \left\| X_{i,k} - \sum_{r=1}^{R} z_i^r f_k^r \phi^r \right\|^2 \tag{1}$$

Les trajectoires $X_{i,k}$ n'étant accessibles que partiellement (acquisition bruitée en quelques points seulement), la résolution directe de (1) conduirait à un problème mal posé. Plusieurs solutions ont été proposées : (i) reconstruire au préalable la trajectoire $X_{i,k}$, (ii) imposer des contraintes de régularité aux fonctions ϕ^r (Choi et al., 2018) , ou (iii) reformuler le problème

en raisonnant sur la fonction de covariance du processus X. Cet article s'inscrit dans la démarche (iii).

Notations. Soient $\mathbf{Z} \in \mathbb{R}^{n\times R}$ la matrice des R scores et $\mathbf{F} \in \mathbb{R}^{K\times R}$ la matrice des R facteurs $f^1, \ldots, f^R$ associés à l'axe *grandeur*. Dans le même esprit, on définit la fonction multidimentionnelle :

$$\begin{array}{rccl} \phi: & [0,T] & \to & \mathbb{R}^{1\times R} \\ & t & \mapsto & \left(\phi^1(t), \ldots, \phi^R(t)\right) \end{array}$$

regroupant les R fonctions canoniques. S'inspirant de la notation usuelle $[\![.,.,.]\!]$ pour PARAFAC, on définit : $[\![\mathbf{Z}, \mathbf{F}, \phi]\!]$ telle que $[\![\mathbf{Z}, \mathbf{F}, \phi]\!]_{i,k}(t) = \sum_{r=1}^{R} z_i^r f_k^r \phi^r(t)$ avec $i \in \{1, \ldots, n\}$, $k \in \{1, \ldots, K\}$ et $t \in [0, T]$.

Estimation théorique. Les trajectoires temporelles $t \mapsto X_{i,k}(t)$ peuvent être considérées comme la réalisation d'un processus aléatoire sous-jacent :

$$\begin{array}{rccl} X: & [0,T] & \to & \mathbb{R}^{K\times 1} \\ & t & \mapsto & (X_1(t), \ldots, X_K(t)) \,. \end{array}$$

Supposons connue la loi jointe des K processus X_k. Le risque défini dans (1) s'interprète comme un estimateur du risque quadratique

$$R_Q = \mathbb{E}\left\{\|X - [\![Z, \mathbf{F}, \phi]\!]\|^2\right\}$$

où le vecteur des scores Z est maintenant un vecteur aléatoire de $\mathbb{R}^{1\times R}$.

La dissymétrie entre le rôle de Z (vecteur aléatoire dépendant du processus X) et celui de $\mathbf{F}$ et ϕ (facteurs et fonctions canoniques dont les valeurs dépendent de la loi de X mais pas directement de X) conduit à une dissymétrie dans la minimisation de R_Q. Soit $\Psi : X \mapsto Z$. La décomposition optimale s'obtient en minimisant R_Q par rapport aux facteurs canoniques et fonctions canoniques $(\mathbf{F}, \phi)$ et par rapport à la fonction Ψ :

$$\left(\hat{\Psi}, \hat{\mathbf{F}}, \hat{\phi}\right) = \underset{\Psi, \mathbf{F}, \phi}{\operatorname{argmin}} \mathbb{E}\left\{\|X - [\![\Psi(X), \mathbf{F}, \phi]\!]\|^2\right\} \tag{2}$$

Le problème (2) permet de séparer l'expression de $\hat{\Psi}$ de celle de $(\hat{\mathbf{F}}, \hat{\phi})$. En effet :

$$\begin{aligned} \min_{\Psi, \mathbf{F}, \phi} \mathbb{E}\left\{\|X - [\![\Psi(X), \mathbf{F}, \phi]\!]\|^2\right\} &= \min_{\mathbf{F}, \phi} \mathbb{E}\left\{\min_{Z} \|X - [\![Z, \mathbf{F}, \phi]\!]\|^2\right\} \\ &= \min_{\mathbf{F}, \phi} \mathbb{E}\left\{\min_{Z} \sum_{k=1}^{K} \int_{[0,T]} |X_k(t) - \langle Z, f_k \circ \phi(t)\rangle|^2 \,\mathrm{d}t\right\} \end{aligned} \tag{3}$$

où $f_k = \left(f_k^1, \ldots, f_k^R\right) \in \mathbb{R}^{1\times R}$ et $\circ$ désigne le produit d'Hadamard.

Afin d'alléger l'écriture, on définit les opérateurs $\bullet$ et $\blacksquare$:

— Soient $U : [0,T] \to \mathbb{R}^{q\times D}$ et $V : [0,T] \to \mathbb{R}^{D\times p}$. L'opérateur $\blacksquare$ réalise la somme sur les indices du second axe et l'intégrale selon l'axe *temps*. On définit $(U \blacksquare V) \in \mathbb{R}^{q\times p}$ par

$$(U \blacksquare V)_{u,v} = \sum_{d=1}^{D} \int_{[0,T]} U_{u,d}(t) V_{d,v}(t) \mathrm{d}t$$

— Soient $U : [0,T] \to \mathbb{R}^{1\times R}$ et $M \in \mathbb{R}^{K\times R}$. L'opérateur $\bullet$ étend le produit d'Hadamard. On définit $(M \bullet U) : [0,T] \to \mathbb{R}^{K\times R}$ par

$$(M \bullet U)_{k,r} : t \mapsto M_{k,r} U_r(t)$$

La minimisation par rapport à Z dans l'équation (3) conduit à :

$$\begin{aligned}\hat{\Psi} : X \mapsto \hat{Z} &= \left(\sum_{k=1}^{K} \int_{[0,T]} (f_k \circ \phi(t)) X_k(t) \mathrm{d}t\right)^{\top} \left(\sum_{k=1}^{K} \int_{[0,T]} (f_k \circ \phi(t))(f_k \circ \phi(t))^{\top} \mathrm{d}t\right)^{-1} \\ &= \left((\mathbf{F} \bullet \phi)^{\top} \blacksquare X\right)^{\top} \left((\mathbf{F} \bullet \phi)^{\top} \blacksquare (\mathbf{F} \bullet \phi)\right)^{-1}\end{aligned}$$

En injectant cette expression dans (3), il vient :

$$\begin{aligned}\left(\hat{\mathbf{F}}, \hat{\phi}\right) &= \arg\min_{\mathbf{F},\phi} \mathbb{E}\left\{\sum_{k=1}^{K} \int_{[0,T]} (X_k(t) - \langle \hat{\Psi}(X), f_k \circ \phi(t)\rangle)^2 \mathrm{d}t\right\} \\ &= \arg\min_{\mathbf{F},\phi} \mathbb{E}\left\{-\left((\mathbf{F} \bullet \phi) \blacksquare X\right)^{\top} \left((\mathbf{F} \bullet \phi) \blacksquare (F \bullet \phi)\right)^{-1} \left((\mathbf{F} \bullet \phi) \blacksquare X\right)\right\} \\ &= \arg\min_{\mathbf{F},\phi} -\sum_{k=1}^{K}\sum_{k'=1}^{K} \int_{[0,T]}\int_{[0,T]} \mathbb{E}\left\{X_k(t) X_{k'}(t')\right\} \Bigg((f_k \circ \phi(t))^{\top} \\ &\qquad \left((\mathbf{F} \bullet \phi)^{\top} \blacksquare (\mathbf{F} \bullet \phi)\right)^{-1} \left((f_k \circ \phi(t')) \mathrm{d}t' \mathrm{d}t\right) \Bigg) \end{aligned} \tag{4}$$

La loi jointe des K processus n'intervient que via les termes $C_{k,k'}(t,t') = \mathbb{E}\{X_k(t) X_{k'}(t')\}$. Par abus de langage, ces termes sont désignés dans l'article par *fonctions de covariance* malgré le fait que les $X_k(t)$ ne soient pas centrés. Ces fonctions de covariance étant inconnues, elle sont estimées à partir des observations.

3 Estimation des fonctions de covariance

Les durées séparant deux observations successives d'une trajectoire $X_{i,k}$ étant variables, l'estimation des fonctions de covariance devra être robuste à un échantillonnage irrégulier, parcimonieux, et à la présence d'observations bruitées. Les techniques d'agrégation et de lissage s'inscrivent dans ce contexte. Nous avons utilisé la technique décrite dans les articles (Yao et al., 2005; Yang et al., 2011).

Les fonctions de covariance $C_{k,k'}$ sont estimées pour tout couple $(k,k') \in \{1,\ldots,K\}^2$. Cette estimation consiste :

— dans un premier temps à associer à chaque couple $(Y_{i,k,j}, Y_{i,k',j'})$ observé pour un même individu i la covariance ponctuelle, calculée aux instants $(t_{i,k,j}, t_{i,k',j'})$:

$$\tilde{C}_{k,k'}[i,j,j'] = Y_{i,k,j} Y_{i,k',j'}$$

— puis à calculer $\hat{C}_{k,k'}$ à partir des $\tilde{C}_{k,k'}[i,j,j']$ pour tout couple (t,t') dont on souhaite estimer $C_{k,k'}(t,t')$. Ce calcul agrège l'ensemble des $\tilde{C}_{k,k'}[i,j,j']$ en les pondérant par un noyau de lissage κ : plus le couple $(t_{i,k,j}, t_{i,k',j'})$ est proche de (t,t') et plus son poids est important.
La régularité des estimations $\hat{C}_{k,k'}$ est accrue en réalisant conjointement avec le lissage précédent une régression polynomiale locale (ici de degré 1) :

$$\begin{aligned}\left(\hat{C}_{k,k'}(t,t'),\, .\,,\, .\right) = \underset{\beta_0,\beta_1,\beta_1'}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{j=1}^{q_{i,k}}\sum_{j'=1}^{q_{i,k'}} \kappa\left(\frac{t-t_{i,k,j}}{b_k}\right)\kappa\left(\frac{t'-t_{i,k',j'}}{b_{k'}}\right) \\ \times\left|\tilde{C}_{k,k'}[i,j,j'] - (\beta_0+\beta_1(t-t_{i,k,j})+\beta_1'(t'-t_{i,k',j'}))\right|^2\end{aligned} \tag{5}$$

Les scalaires b_k permettent de faire varier le degré de lissage, pour chacune des K composantes du processus X. En pratique, on les choisit en fonction de la régularité présupposée de chaque composante et du nombre total de points observés pour cette composante $(\sum_{i=1}^{n} q_{i,k})$.

Remarque. Lorsque $k = k'$, les $\tilde{C}_{k,k}[i,j,j]$ sont écartés de la somme dans (5) car les termes $Y_{i,k,j}^2$ sont liés à la variance des observations et non à la variance du processus X_k (c'est-à-dire à $C_{k,k}(t,t) + \sigma_k^2$ et non à $C_{k,k}(t,t)$).

Estimation des σ^2.

Soit $c_k(t) = C_{k,k}(t,t) + \sigma_k^2$ la variance du processus X_k bruité. Les fonctions de variance c_k peuvent être estimées avec la même technique que pour l'estimation de $C_{k,k'}(t,t')$:

$$(\hat{c}_k(t),\, .\,) = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{N}\sum_{j=1}^{q_{i,k}} \kappa\left(\frac{t-t_{i,k,j}}{b_k}\right) \times \left|Y_{i,k,j}^2 - (\beta_0+\beta_1(t-t_{i,k,j}))\right|^2$$

On en déduit l'estimation de la variance du bruit d'observation :

$$\hat{\sigma}_k^2 = \frac{1}{T}\int_0^T \left(\hat{c}_k(t) - \hat{C}_{k,k}(t,t)\right) \mathrm{d}t$$

4 Estimation des fonctions et facteurs canoniques

L'estimation $(\hat{\mathbf{F}}, \hat{\phi})$ de $(\mathbf{F}, \phi)$ s'obtient en résolvant le problème d'optimisation (4) avec $\hat{C}_{k,k'}(t,t')$ à la place $C_{k,k'}(t,t')$. Le problème (4) n'a pas de solution analytique. L'équivalence entre la formulation initiale du problème (1) et la reformulation à l'aide des fonctions de

covariance (4) a été exploitée : le problème (1), résolu pour un nombre croissant de trajectoires générées selon une loi ayant la fonction de covariance $C_{k,k'}(t,t')$ conduit à des estimations $(\hat{\mathbf{F}},\hat{\phi})$ dont la limite est la solution de (4). Soient donc M trajectoires $\tilde{X}_{m,k}(t)$ générées selon une loi ayant pour fonctions de covariance $\hat{C}_{k,k'}(t,t')$.

De la relation

$$\lim_{M\to\infty}\frac{1}{M}\underbrace{\sum_{m=1}^{M}\sum_{k=1}^{K}\left\|\tilde{X}_{m,k}-\sum_{r=1}^{R}z_m^r f_k^r\phi^r\right\|^2}_{\tilde{R}_M(z,F,\phi)}=\mathbb{E}\left\{\|X-[\![Z,\mathbf{F},\phi]\!]\|^2\right\}$$

on conclut que $\lim_{M\to\infty}(\mathbf{F}_M,\phi_M)=\left(\hat{\mathbf{F}},\hat{\phi}\right)$ où $(.,\mathbf{F}_M,\phi_M)=\arg\min_{\mathbf{Z},\mathbf{F},\phi}\tilde{R}_M(\mathbf{Z},\mathbf{F},\phi)$.

La fonction objectif $\tilde{R}_M$ dépendant de trajectoires temporelles simulées, elle peut être approchée aussi finement que souhaité par un échantillonnage régulier de l'intervalle $[0,T]$. La minimisation de $\tilde{R}_M$ est obtenue en appliquant l'algorithme PARAFAC à ces trajectoires échantillonnées régulièrement.

5 Calcul des vecteurs de scores

Le vecteur de scores associé à l'individu i ne peut être obtenu directement en raison de la connaissance très partielle des trajectoires $X_{i,k}$. La démarche développée dans (Zhou et al., 2008) a été utilisée. Elle consiste à :

— décomposer les trajectoires

$$X_{i,k}=\sum_{r=1}^{R}\xi_{i,r}f_k^r\phi^r+\rho_{i,k} \tag{6}$$

où $\rho_{i,k}:[0,T]\to\mathbb{R}$ représente l'erreur de modélisation. On définit ξ_i comme le vecteur minimisant la norme de l'erreur de modèle

$$\begin{aligned}\xi_i &= \arg\min_{\beta\in\mathbb{R}^R}\sum_{k=1}^{K}\int_{[0,T]}\left|X_{i,k}(t)-\beta^\top(f_k\circ\phi(t))\right|^2\mathrm{d}t\\ &= \left((\mathbf{F}\bullet\phi)^\top\blacksquare(\mathbf{F}\bullet\phi)\right)^{-1}\left((\mathbf{F}\bullet\phi)^\top\blacksquare(X_{i,1},\ldots,X_{i,K})\right)\end{aligned}$$

— modéliser $\xi_i\in\mathbb{R}^R$ par un vecteur gaussien, et estimer z_i par l'espérance conditionnelle de ξ_i par rapport aux données observées pour l'individu i : $Y_i=\left\{Y_{i,k,1\ldots q_{i,k}}\right\}_{1\le k\le K}$.
Par linéarité du modèle (6), Y_i est un vecteur gaussien. Soit Σ_{Y_i} sa matrice de covariance et $\Sigma_{\xi_i Y_i}$ la matrice de covariance entre ξ_i et Y_i.
On a (Rasmussen et al., 2006) : $\xi_i \mid Y_i\sim\mathcal{N}(\mu_i,\Sigma_i)$ avec $\mu_i=\Sigma_{\xi_i Y_i}\Sigma_{Y_i}^{-1}Y_i$.

Expression de Σ_{Y_i} et de $\Sigma_{\xi_i Y_i}$.

Elles s'écrivent en fonction de $\hat{C}_{k,k'}(t,t')$ et de $\hat{\sigma}_k$:

$$\begin{cases} \mathbb{E}(Y_{i,k,j}Y_{i,k',j'}) & = \hat{C}_{k,k'}(t_{i,k,j},t_{i,k',j'}) + \hat{\sigma}_k^2\delta_{(k,j)=(k',j')} \\ \Sigma_{\xi_i Y_{i,k,j}} & = \mathbb{E}\left\{\xi_i X_{i,k}(t_{i,k,j})\right\} \\ & = \left((\mathbf{F}\bullet\phi)^\top \blacksquare (\mathbf{F}\bullet\phi)\right)^{-1}\left((\mathbf{F}\bullet\phi)^\top \blacksquare C_{k,t_{i,k,j}}\right) \end{cases}$$

$$\text{où } C_{k,t} : \begin{array}{ccc} [0,T] & \to & \mathbb{R}^{K\times 1} \\ t' & \mapsto & (C_{k,1}(t,t'),\ldots,C_{k,K}(t,t')) \end{array}$$

6 Test sur données simulées

Les données simulées permettent d'avoir une référence et ainsi d'évaluer directement la qualité de reconstruction des scores et des vecteurs et fonctions canoniques. Elles permettent également de générer des scénarii de plus ou moins grande complexité. n trajectoires ont été générées selon un modèle PARAFAC avec $K = 10$ et $R = 4$. Les scores et les facteurs ont été générés selon des lois IID alors que les fonctions canoniques ont été générées de manière à obtenir des trajectoires $X_{i,k}$ régulières (voir Figure 1).

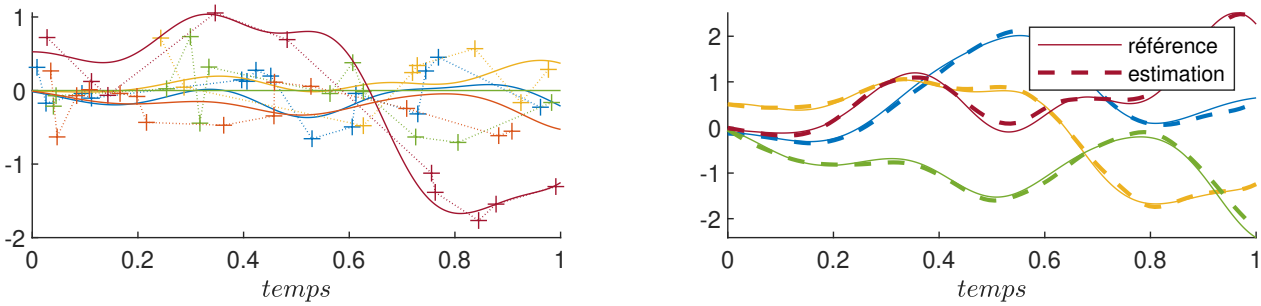

FIGURE 1 – Gauche : 5 trajectoires $X_{i,k}$ ($k = 1$, $i = 1\ldots 5$) et observations $Y_{i,k}$. Droite : ϕ_r ($r = 1\ldots R$) et leurs estimations ($[q_{\min}, q_{\max}] = [5, 15]$ et $n = 200$).

Deux échantillonnages différents ont été mis en œuvre :

1. Echantillonnage avec $q_{i,k}$ variable d'une grandeur k à une autre et d'un individu i à un autre. Cet échantillonnage permet d'évaluer le comportement de l'approche proposée en fonction de la parcimonie ($q_{i,k}$ faible) et du nombre d'individus (n).
2. Echantillonnage avec $q_{i,k} = q$ (constante) avec toutefois un échantilonnage irrégulier. Cet échantillonnage se rencontre par exemple lors de la constitution d'une cohorte médicale au cours de laquelle certains participants ont leurs examens décalés par rapport aux dates de référence.
   Les données collectées pouvant être regroupées au sein d'un tenseur de taille $n \times K \times q$, il est tentant d'utiliser directement PARAFAC, bien que les données ne soient pas synchrones.

Echantillonnage 1.
Les observations ont été obtenues en échantillonnant uniformément l'intervalle temporel $[0, T]$ avec un nombre d'instants tiré aléatoirement selon une loi uniforme entre $q_{\min}$ et $q_{\max}$ (avec $q_{\max} = 3q_{\min}$).

Le tableau 1 montre que la qualité des estimations décroit lorsque l'échantillonnage devient très parcimonieux. Il est cependant possible de récupérer une bonne qualité d'estimation en augmentant le nombre d'individus.

n	$[q_{\min}, q_{\max}]$	γ_ϕ	γ_z
50	[25,75]	0.957	0.991
50	[10,30]	0.939	0.969
50	[5,15]	0.833	0.799
100	[5,15]	0.984	0.920
200	[5,15]	0.994	0.984

TABLE 1 – Méthode proposée, Echantillonnage 1. Qualité d'estimation des fonctions canoniques (γ_ϕ) et des scores (γ_z). γ_z et γ_ϕ sont les moyennes des R coefficients de corrélation entre référence et estimés.

Echantillonnage 2.
L'échantillonnage selon une loi uniforme de l'intervalle $[0, T]$ a été conservé, mais avec un nombre d'instants constant (noté q).

Le tableau 2 montre que la méthode proposée permet de mieux reconstruire les fonctions canoniques et les scores. Ces résultats montrent l'intérêt d'utiliser une approche prenant spécifiquement en compte le caractère irrégulier de l'échantillonnage plutôt que PARAFAC.

		Méthode proposée		PARAFAC	
n	q	γ_ϕ	γ_z	γ_ϕ	γ_z
100	20	0.995	0.990	[0.665 , 0.979]	[0.605 , 0.980]
50	20	0.988	0.987	[0.640 , 0.978]	[0.573 , 0.980]

TABLE 2 – Echantillonnage 2. Comparaison des résultats donnés par PARAFAC et par la méthode proposée sur des données échantillonnées irrégulièrement mais à nombre d'échantillons invariant d'une grandeur à l'autre et d'un individu à l'autre. Les résultats obtenus avec PARAFAC présentent une forte variance ; c'est pourquoi les intervalles de confiance à 50% ont été donnés.

7 Conclusion

La démarche proposée permet une décomposition similaire à celle de PARAFAC, mais en présence de données échantillonnées irrégulièrement. Les résultats montrent que lorsque

la méthode est appliquée pour un nombre suffisamment grand d'individus, elle permet une reconstruction quasi parfaite des scores et des fonctions canoniques, c'est-à-dire sans perte par rapport aux résultats qu'aurait donnés PARAFAC appliqué à des données complètes. Les résultats montrent également le risque qu'il y aurait à utiliser PARAFAC sur des observations non synchrones : la qualité d'estimation est alors en deça de celle qu'il est possible d'atteindre avec la méthode proposée.

A noter que le contexte des données échantillonnées irrégulièrement permet d'englober celui des données manquantes ; une donnée manquante correspondant à un instant exclus de la liste des instants d'acquisition (pour un individu et pour une ou plusieurs grandeurs).

Seul le cas où l'échantillonnage ne concerne qu'un axe est ici présenté. La technique d'estimation de fonctions de covariance par agrégation et lissage étant applicable en dimension plus élevée, la démarche proposée se généralise à un nombre supérieur d'axes, avec potentiellement plusieurs axes échantillonnés irrégulièrement. Des données spatio-temporelles échantillonnées irrégulièrement temporellement et/ou spatialement pourraient ainsi être traitées.

Bibliographie

Carroll, J. D. and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "eckart-young" decomposition. *Psychometrika*, 35(3) :283–319.

Choi, J. Y. et al. (2018). Functional parallel factor analysis for functions of one-and two-dimensional arguments. *psychometrika*, 83 :1–20.

Harshman, R. A. et al. (1970). Foundations of the parafac procedure : Models and conditions for an" explanatory" multimodal factor analysis.

Ramsay, J. O. and Silverman (2006). *Functional Data Analysis*. Springer New York.

Rasmussen, C. E. et al. (2006). *Gaussian processes for machine learning*, volume 1. Springer.

Shang, H. (2014). A survey of functional principal component analysis. *AStA Adv Stat Anal*, 98 :121–142.

Tucker, L. R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3) :279–311.

Yang, W. et al. (2011). Functional singular component analysis. *Journal of the Royal Statistical Society Series B : Statistical Methodology*, 73(3) :303–324.

Yao, F. et al. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, 100(470) :577–590.

Zhou, L. et al. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika*, 95(3) :601–619.