Statistiques de scan spatiales

Moustapha Sarr

Master 2 ISN
Université de Lille

Plan

Introduction

Au cours des dernières décennies, les avancées technologiques dans les capteurs et le stockage des données ont conduit à l'émergence de données mesurées en temps quasi-continu. Par exemple, la concentration de particules fines dans l'air extérieur est mesurée quotidiennement par les autorités de surveillance de la qualité de l'air. Dans le domaine de l'épidémiologie, les autorités sanitaires surveillent en temps réel l'évolution de l'incidence d'une maladie afin d'identifier rapidement le début d'une épidémie. Le développement de ce type de **données fonctionnelles univariées et multivariées** a conduit à la popularisation de l'analyse des données fonctionnelles (Ramsay et Silverman, 1997).

Dans les domaines où les données sont fondamentalement de nature spatiale (par exemple, la démographie, les sciences de l'environnement...), la massification des données fonctionnelles a conduit à la définition **des données spatio-fonctionnelles** (Delicado et al., 2010). Ces données sont caractérisées par l'observation d'une ou plusieurs courbes dans chaque localisation spatiale. Par exemple, la pollution de l'air est mesurée par des capteurs répartis sur une zone géographique, chaque capteur enregistrant de manière quasi-continue la concentration d'un ou plusieurs polluants, ce qui conduit à l'observation d'une ou plusieurs courbes.
Nous pouvons dire que **les données fonctionnelles** sont des données où chaque **observation est une fonction** plutôt qu'une valeur unique. La philosophie fondamentale de l'analyse des données fonctionnelles est de considérer les fonctions observées comme des entités uniques, plutôt que comme de simples séries d'observations individuelles.

Introduction

L'émergence de ce type de données a naturellement conduit au développement de méthodes statistiques telles que :

- Les méthodes de clustering (Giraldo et al., 2012 ; Jiang et Serban, 2012 ; Romano et al., 2017 ; Vandewalle et al., 2021).

Ces méthodes visent à partitionner les données spatio-fonctionnelles selon un critère de dissimilarité tout en tenant compte de la dépendance spatiale. Cependant, ces approches ne permettent pas :

- D'identifier des clusters spatiaux
- De tester la significativité statistique des clusters.

Dans ce contexte, **les statistiques de scan spatial** semblent bien adaptées pour répondre à ces questions. Cependant, la nature particulière des données spatio-fonctionnelles ne permet pas d'utiliser les modèles classiques.

Si nous utilisons une **statistique de scan spatial univariée** en moyennant sur la période de temps, cela entraîne une perte d'information importante.

Introduction

Le problème peut également être abordé en utilisant une statistique de scan spatial multivariée (Kulldorff et al., 2007 ; Cucala et al., 2019) en considérant chaque instant de mesure comme une variable.
Cett approchoche peut conduire à :

- Un problème de données manquantes, si les temps de mesure diffèrent entre les unités spatiales.
- Un problème de grande dimension, si le nombre de temps de mesure est élevé.
- Un problème de multicolinéarité, en raison de la dépendance temporelle entre les mesures effectuées à différents instants.

  Récemment, plusieurs auteurs ont proposé des statistiques de scan spatial paramétriques et non paramétriques pour les données fonctionnelles univariées et multivariées (Frévent et al., 2021a,b ; Smida et al., 2022). Ces approches permettent de surmonter les limites des modèles classiques univariés et multivariés dans l'analyse des données spatio-fonctionnelles.

Principe générale

Soit $\{s_1, \ldots, s_n\}$ un ensemble de n emplacements spatiaux disjoints dans un domaine d'observation $S \subseteq \mathbb{R}^2$.

Les données observables à chaque emplacement s_i sont données par :

$$\left\{ \left(X_i^{(1)}(t), X_i^{(2)}(t), \ldots, X_i^{(p)}(t) \right)^T, t \in \mathcal{T} \right\}$$

où :

- $\mathcal{T}$ est un intervalle de $\mathbb{R}$.
- $X_i^{(j)}(\cdot)$ représente l'observation d'un processus stochastique réel $\{X^{(j)}(t), t \in \mathcal{T}\}$ avec $j = 1, \ldots, p$.

Le paramètre p représente la dimension des données fonctionnelles :

- Si $p = 1$, on parle de **données fonctionnelles univariées**.
- Si $p \geq 2$, on parle de **données fonctionnelles multivariées**.

En pratique, chaque processus $X^{(j)}$ est mesuré à des instants discrets de $\mathcal{T}$, avec ou sans erreur. Ces instants peuvent :

- Varier en nombre et en valeur selon l'emplacement spatial.
- Différer d'un processus à l'autre.

Principe générale

On note ces observations sous la forme :

$$\{X_{i,1}^{(j)}, \ldots, X_{i,m_{i,j}}^{(j)}\}$$

où $X_{i,k}^{(j)}$ sont les mesures discrètes du $j^{\text{ème}}$ processus $X^{(j)}$ à l'emplacement spatial s_i, effectuées en $m_{i,j}$ instants de $\mathcal{T}$.

L'objectif est de transformer ces mesures discrètes en une fonction $X_i^{(j)}(t_{i,k}), \quad k = 1, \ldots m_{i,j}$, qui peut être évaluée pour n'importe quelle valeur $t_{i,k}$, pas seulement aux instants mesurés.

Traitement des données quantitatives fonctionnelles :

Puisque les observations sont longitudinales et discrètes, il est nécessaire de reconstruire la courbe originale de chaque processus $X^{(j)}$ pour tous les emplacements s_i.

Une méthode consiste à approximer $X^{(j)}$ dans un espace fonctionnel généré par une base de fonctions $\varphi_k^{(j)}(t)$.

Principe générale

$$X_i^{(j)}(t) = \sum_{k=1}^{K_j} a_{ik}^{(j)} \varphi_k^{(j)}(t), \quad j = 1, \ldots, p; \quad i = 1, \ldots, n$$

où :

- K_j est le nombre de fonctions de base utilisées pour approximer $X^{(j)}$.
- $a_{ik}^{(j)}$ sont les coefficients de base (matrice de taille $n \times K_j$).
- La base de fonctions $\varphi_k^{(j)}(t)$ est choisie en fonction de la structure des données.

Deux approches existent en fonction de la présence ou non d'erreurs de mesure :

1. Interpolation : Si les mesures sont exactes et sans erreur, on cherche une fonction qui passe exactement par les points mesurés.

$$X_{i,r}^{(j)} = X_i^{(j)}(t_{i,r}), \quad r = 1, \ldots, m_{i,j}$$

2. Lissage (smoothing) par les moindres carrés (ordinaires ou généralisés) : Si les mesures contiennent du bruit ou des erreurs, on ne veut pas nécessairement que la fonction passe par tous les points.

$$X_{i,r}^{(j)} = X_i^{(j)}(t_{i,r}) + e_{i,r}^{(j)}, \quad r = 1, \ldots, m_{i,j}$$

où $e_{i,r}^{(j)}$ représente l'erreur de mesure associée.

Choix de la base de fonctions

Le choix de la base dépend de la structure des données fonctionnelles :

- **Base de B-splines** : Adaptée aux données non périodiques.
- **Base de Fourier** : Utile pour les données périodiques.
- **Base d'ondelettes** : Appropriée pour les données avec discontinuités ou changements de comportement.

Référence : Ramsay et Silverman (2005)

Représentation des fonctions par des bases de fonctions

Définition : Un système de bases de fonctions est un ensemble de fonctions indépendantes $\{\phi_k\}$ permettant d'approximer toute fonction $x(t)$ comme une combinaison linéaire :

$$x(t) = \sum_{k=1}^{K} c_k \phi_k(t)$$

où c_k sont les coefficients de projection.

Exemples de bases de fonctions

Série de puissances (polynômes) :

$$1, t, t^2, t^3, ..., t^k, ...$$

Utilisée pour construire des séries de puissances..

Série de Fourier :

$$1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t), ..., \sin(k\omega t), \cos(k\omega t), ...$$

Idéale pour les fonctions périodiques.

Représentation matricielle

Forme matricielle :

$$x = \Phi c$$

où :

- Φ est la matrice des bases $\phi_k(t)$.
- c est le vecteur des coefficients.

Cela transforme un problème en dimension infinie en un problème de dimension finie.

Choix du nombre de bases K

Impact de K :

- **Si** $K = n$: Interpolation exacte ($x(t_j) = y_j$), n est le nombre d'observations.
- **Si** $K < n$: Lissage des données (réduction du bruit).

Bon choix de K :

- Maximiser les degrés de liberté.
- Réduire la complexité des calculs.
- Faciliter l'interprétation des coefficients c_k.

Le système de base de Fourier pour les données périodiques

L'une des expansions en base les plus connues est la série de Fourier :

$$\hat{x}(t) = c_0 + c_1 \sin(\omega t) + c_2 \cos(\omega t) + c_3 \sin(2\omega t) + c_4 \cos(2\omega t) + \ldots$$

Cette base est définie par les fonctions :

$$\phi_0(t) = 1$$

$$\phi_{2r-1}(t) = \sin(r\omega t)$$

$$\phi_{2r}(t) = \cos(r\omega t)$$

Elle est périodique, et le paramètre ω détermine la période $\frac{2\pi}{\omega}$.

Base B-spline pour les fonctions splines

Une **fonction spline** est une fonction définie par morceaux, composée de polynômes sur des sous-intervalles appelés **nœuds** (*knots*).
Mais comment construire une telle fonction ?

On utilise un **système de base de fonctions B-spline** noté $\phi_k(t)$, qui possède les propriétés suivantes :

- Chaque fonction de base $\phi_k(t)$ est elle-même une spline définie par un **ordre** m et une **séquence de nœuds** τ.
- Les **combinaisons linéaires** de ces bases restent des splines, car les sommes et différences de splines sont aussi des splines.
- Toute spline définie par m et τ peut être exprimée comme combinaison linéaire de ces bases.

Définition mathématique d'une B-spline

La notation classique d'une **fonction B-spline** est :

$$B_k(t, \tau)$$

où :

- k est l'indice du plus grand nœud situé à gauche ou immédiatement à gauche de t,
- τ représente la séquence des nœuds.

Une **fonction spline** $S(t)$ s'écrit alors comme une combinaison linéaire de fonctions B-splines :

$$S(t) = \sum_{k=1}^{m+L-1} c_k B_k(t, \tau)$$

où :

- c_k sont les **coefficients** de la combinaison linéaire,
- m est l'**ordre** de la spline,
- L est le **nombre total de nœuds** (points de rupture internes).

Clusters spatiaux de données fonctionnelles

Notons :

$$\left\{ \left(X_i^{(1)}(t), X_i^{(2)}(t), \ldots, X_i^{(p)}(t) \right)^T, t \in \mathcal{T} \right\}$$

un processus stochastique vectoriel à p dimensions, dont chaque composante correspond aux processus individuels $X^{(j)}$ pour $j = 1, \ldots, p$.

Selon Dai et al. (2020), une observation fonctionnelle peut être déviante par :

- **Son ampleur (magnitude)** : L'observation suit la même forme que les autres, mais avec une échelle différente (amplitude plus grande ou plus petite).
- **Sa forme (shape)** : L'observation suit une forme différente de celle de la majorité des données.

Types de clusters spatiaux de données fonctionnelles (Frévent et al., 2021a)

- **Cluster spatial de magnitude** : Les courbes fonctionnelles dans cette zone ont la **même forme** que les autres, mais avec une **amplitude différente** (ex. valeurs plus élevées ou plus faibles).
- **Cluster spatial de forme** : Les courbes fonctionnelles dans cette zone **n'ont pas la même forme** que la majorité des données.
- Clusters mixtes (magnitude + forme)

Définition mathématique d'un cluster spatial fonctionnel

Un cluster fonctionnel est alors une zone spatiale où les courbes diffèrent de celles du reste des localisations étudiées.

En suivant *Frévent et al.* (2021a, b), on peut formellement définir un cluster spatial w à l'aide d'un décalage fonctionnel non nul $\Delta(t)$ appliqué à la moyenne des données fonctionnelles :

$$\mathbb{E}\big[X_i(t) \mid s_i \in z\big] = \mathbb{E}\big[X_i(t) \mid s_i \notin z\big] + \Delta(t), \quad i = 1, \ldots, n$$

où $\Delta(t)$ est un vecteur fonctionnel :

$$\Delta(t) = \big(\Delta_1(t), \Delta_2(t), \ldots, \Delta_p(t)\big)^T$$

La forme du décalage fonctionnel $\Delta(t)$ permet de distinguer différents types de clusters :

- **Cluster de magnitude** : si $\Delta(t)$ affecte uniquement l'amplitude.
- **Cluster de forme** : si $\Delta(t)$ modifie la forme de la fonction.
- **Cluster mixte** : si $\Delta(t)$ combine les deux effets.

Définition de la statistique de scan spatial fonctionnelle

- $\mathbb{E}\left[X_i(t) \mid s_i \in z\right]$: est l'espérance (la moyenne) des courbes dans la zone définissant le cluster z.
- $\mathbb{E}\left[X_i(t) \mid s_i \notin z\right]$: est l'espérance des courbes en dehors du cluster.

Comme dans les cadres univarié et multivarié, l'indice de concentration paramétrique vise à comparer les fonctions moyennes entre les clusters potentiels et l'extérieur.

Le processus X est supposé prendre ses valeurs dans l'espace $\mathcal{L}^2(\mathcal{T}, \mathbb{R})$, qui est l'espace des fonctions réelles de carré intégrable sur $\mathcal{T}$.

Travaux antérieurs :

- *Cuevas et al.* (2004)
- *Górecki et Smaga* (2015)

ont adapté la statistique F classique de l'ANOVA aux processus dans $\mathcal{L}^2$.

L'espace $\mathcal{L}^2(T, \mathbb{R}^p)$

Définition :
L'espace $\mathcal{L}^2(T, \mathbb{R}^p)$ est l'espace de Hilbert des fonctions vectorielles $X(t)$ définies sur un intervalle $\mathcal{T}$, à valeurs dans $\mathbb{R}^p$, et qui sont de carré intégrables.

Cela signifie que l'intégrale de leur norme euclidienne au carré est finie :

$$\int_{\mathcal{T}} \|X(t)\|^2 dt < \infty.$$

Cette condition garantit que les fonctions considérées ne "divergent pas trop" et qu'elles admettent des moments d'ordre deux (moyenne et variance bien définies).

Produit scalaire :

$$\langle X, Y \rangle = \int_T X(t)^\top Y(t) dt.$$

Hypothèses du test

En considérant deux échantillons indépendants de trajectoires issus de deux processus $\mathcal{L}^2$, X_z et X_{z^c}, appartenant respectivement aux groupes z et z^c, le test compare les fonctions moyennes μ_z et μ_{z^c}, où :

$$\mu_z(t) = \mathbb{E}[X_z(t)], \quad \mu_{z^c}(t) = \mathbb{E}[X_{z^c}(t)].$$

Hypothèse nulle (H_0) : Absence de cluster

$$H_0 : \forall z \in Z, \quad \mu_z = \mu_{z^c} = \mu_S$$

où :

- μ_z : fonction moyenne dans le cluster potentiel z,
- μ_{z^c} : fonction moyenne en dehors de z,
- μ_S : fonction moyenne sur l'ensemble S.

Hypothèse alternative (H_1) : Existence d'un cluster

$$H_1^z : \mu_z \neq \mu_{z^c}$$

Test ANOVA fonctionnel univarié

Dans ce contexte, ils ont proposé la **statistique de scan spatial fonctionnel paramétrique (PFSS)** :

$$\Lambda_{\text{PFSS}} = \max_{z \in Z} F_n^z$$

où

$$F_n^z = \frac{|z| \, \|\bar{X}_z - \bar{X}\|_2^2 + |z^c| \, \|\bar{X}_{z^c} - \bar{X}\|_2^2}{\frac{1}{n-2} \sum_{j, s_j \in z} \|\bar{X}_j - \bar{X}_z\|_2^2 + \sum_{j, s_j \in z^c} \|\bar{X}_j - \bar{X}_{z^c}\|_2^2}$$

où :

- $\bar{X}_z(t) = \frac{1}{|z|} \sum_{i, s_i \in z} X_i(t)$ est l'estimateur empirique de la fonction moyenne dans z,
- $\bar{X}_{z^c(t)} = \frac{1}{|z|} \sum_{i, s_i \in z^c} X_i(t)$ est l'estimateur empirique de la fonction moyenne dans z^c,
- $\bar{X}(t) = \frac{1}{n} \sum_{i=1}^{n} X_i(t)$ est la fonction moyenne sur S,
- $\|x\|_2^2 = \int_T x^2(t) dt$ est la norme $\mathcal{L}^2$.

Rappel de l'analyse de la Variance (ANOVA) à un facteur

L'ANOVA est une méthode statistique permettant d'étudier la modification de la moyenne μ d'une variable quantitative Y sous l'influence d'un ou plusieurs facteurs qualitatifs.

Cas d'une ANOVA à un facteur

Si la moyenne μ est influencée par un seul facteur A, on parle d'ANOVA à un facteur ("one-way ANOVA").

- Le facteur A est une variable qualitative avec un nombre fini de modalités (K).
- Y suit une loi normale $\mathcal{N}(\mu_i, \sigma^2)$ pour chaque sous-population définie par A.

Hypothèse nulle

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_I$$

Hypothèse alternative

$$H_1 : \exists\ i_0 \neq j_0 \text{ tel que } \mu_{i_0} \neq \mu_{j_0}$$

(il existe au moins deux moyennes différentes)

Rappel de l'analyse de la Variance (ANOVA) à un facteur

Pour chaque population i, on dispose d'un échantillon de n_i observations :

$$x_{i,1}, x_{i,2}, \ldots, x_{i,n_i}$$

La statistique de test est définie par :

$$F = \frac{\text{SCF}}{k-1} \div \frac{\text{SCR}}{n-k} = \frac{S_F^2}{S_R^2}$$

où :

$$\text{SCF} = \sum_{i=1}^{k} n_i (\bar{X}_i(t) - \bar{X}(t))^2,$$

$$\text{SCR} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij}(t) - \bar{X}_i(t))^2.$$

sont respectivement, les variances théorique et résiduelle.
De plus, on définit :

$$\bar{X}(t) = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{ij}(t),$$

$$\bar{X}_i(t) = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{ij}(t).$$

Test ANOVA fonctionnel univarié

Correspondance entre les deux cas

Concept	**ANOVA Fonctionnelle Classique**	**ANOVA Fonctionnelle Spatiale**
Nombre de groupes k	k prédéfinis	$k = 2$: le cluster z et le reste z^c
Groupes	$X_{i,j}(t)$ est associé à un groupe g_i	$X_j(t)$ est soit dans z, soit dans z^c
Hypothèse nulle H_0	$\mu_1 = \mu_2 = \cdots = \mu_k$	$\mu_z = \mu_{z^c}$
Hypothèse alternative H_1	$\mu_1 \neq \mu_2 \neq \ldots \neq \mu_k$	Il existe au moins un cluster z où $\mu_z \neq \mu_{z^c}$
Test statistique	F-test classique pour comparer les groupes	Scan spatial basé sur ANOVA fonctionnelle

Statistique de scan spatial distribution-free

Ensuite, Frévent et al. (2021a) ont proposé de combiner la **statistique de scan spatial distribution-free** pour les données univariées introduite par Cucala (2014) et la statistique max de Lin et al. (2021).
Ils ont supposé que, pour chaque instant t et pour tout $i \in \{1, \ldots, I\}$, la variance de $X_i(t)$ est donnée par $\sigma^2(t)$. Ils ont ensuite défini l**a statistique de scan spatial fonctionnel distribution-free (DFFSS)** comme :

$$\Lambda_{\text{DFFSS}} = \max_{z \in Z} I^z,$$

où

$$I^z = \sup_{t \in T} \frac{\left|\bar{X}_z(t) - \bar{X}_{z^c}(t)\right|}{\sqrt{\widehat{\text{Var}}\left(\bar{X}_z(t) - \bar{X}_{z^c}(t)\right)}},$$

avec

$$\widehat{\text{Var}}\left(\bar{X}_z(t) - \bar{X}_{z^c}(t)\right) = \hat{\sigma}^2(t)\left(\frac{1}{|z|} + \frac{1}{|z^c|}\right),$$

et

$$\hat{\sigma}^2(t) = \frac{1}{n-2}\left[\sum_{i;s_i \in z}\left(X_i(t) - \bar{X}_z(t)\right)^2 + \sum_{i;s_i \in z^c}\left(X_i(t) - \bar{X}_{z^c}(t)\right)^2\right].$$

Rappel sur la comparaison des moyennes de deux échantillons indépendants à

Variance connue

On suppose que la variance théorique commune aux deux populations est connue. Sous ces hypothèses, on a :

$$\bar{X}_z - \bar{X}_{z^c} \sim \mathcal{N}\left(\mu_z - \mu_{z^c}, \hat{\sigma}^2\left(\frac{1}{|z|} + \frac{1}{|z^c|}\right)\right)$$

Ainsi, sous H_0, on a :

$$\bar{X}_z - \bar{X}_{z^c} \sim \mathcal{N}\left(0, \hat{\sigma}^2\left(\frac{1}{|z|} + \frac{1}{|z^c|}\right)\right)$$

ce qui s'écrit encore :

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}\sqrt{\frac{1}{|z|} + \frac{1}{|z^c|}}} \sim \mathcal{N}(0,1)$$

Approche

Dans le cadre non paramétrique, Smida et al. (2022) ont proposé de considérer les hypothèses suivantes : En notant P_z et P_{z^c} les mesures de probabilité de X dans z et dans z^c, respectivement, l'hypothèse nulle est définie comme :

$$H_0 : \forall z \in Z, \quad P_z = P_{z^c}$$

tandis que l'hypothèse alternative associée à un cluster potentiel z est donnée par :

$$H_1^z : P_z \text{ et } P_{z^c} \text{ diffèrent par un décalage } \tau \neq 0.$$

Ils ont ensuite défini la **statistique de scan spatial fonctionnelle non paramétrique (NPFSS)** pour les données fonctionnelles univariées comme :

$$\Lambda_{\text{NPFSS}} = \max_{z \in Z} \| T_z \|$$

où

$$T_z = \frac{1}{\sqrt{|z||z^c|n}} \sum_{i;s_i \in z} \sum_{j;s_j \in z^c} \frac{X_j - X_i}{\|X_j - X_i\|}.$$

Statistique de scan fonctionnelle basée sur les rangs (URBFSS)

De plus, une approche basée sur les rangs est également proposée, à savoir **l'URBFSS**, qui vise à tester l'hypothèse nulle :

$$H_0 : \forall z \in Z, \forall t \in \mathcal{T}, F_{z,t} = F_{z^c,t}$$

contre une hypothèse alternative $H_1^{(z)}$ associée à un cluster potentiel z :

$$H_1^{(z)} : \exists t \in T \text{ tel que } F_{z,t}(x) = F_{z^c,t}(x - \Delta_t), \quad \Delta_t \neq 0$$

où $F_{z,t}$ et $F_{z^c,t}$ sont les fonctions de répartition cumulées de $X(t)$ dans z et en dehors de z, respectivement.
Dans ce contexte, la statistique de scan spatial fonctionnelle basée sur les rangs univariés (URBFSS) est définie comme :

$$\Lambda_{\text{URBFSS}} = \max_{z \in z} T^{(z)}$$

où

$$T^{(z)} = \sup_{t \in T} \left| Z_t^{(z)} - \frac{|z|(n+1)}{2} \right| \Big/ \sqrt{|z||z^c| \frac{(n+1)}{12}}$$

avec

$$Z_t^{(z)} = \sum_{i, s_i \in z} R_i(t)$$

où $R_i(t)$ est le rang de $X_i(t)$ parmi $\{X_1(t), \ldots, X_n(t)\}$, en utilisant le rang moyen en cas d'observations ex aequo.

Approche dans le cas mutivarié

Frévent et al. (2021b) ont proposé **une statistique de scan paramétrique pour les données fonctionnelles multivariées,** basée sur **le test de trace Lawley-Hotelling de la MANOVA fonctionnelle** (Górecki et Smaga, 2017).
En résumé, ils supposent que X prend ses valeurs dans l'espace $L^2(T;\mathbb{R}^p)$ des fonctions vectorielles à p dimensions, intégrables au carré sur T.
Ils définissent alors la **statistique de scan spatial fonctionnelle multivariée paramétrique (MPFSS)** comme :

$$\Lambda_{\text{MPFSS}} = \max_{z \in Z} \text{Trace}(H_z E_z^{-1})$$

où

$$H_z = |z| \int_T (\bar{X}_z(t) - \bar{X}(t))(\bar{X}_z(t) - \bar{X}(t))^\top dt + |z^c| \int_T (\bar{X}_{z^c}(t) - \bar{X}(t))(\bar{X}_{z^c}(t) - \bar{X}(t))^\top dt$$

et

$$E_z = \sum_{j;s_j \in z} \int_T (X_j(t) - \bar{X}_z(t))(X_j(t) - \bar{X}_z(t))^\top dt + \sum_{j;s_j \in z^c} \int_T (X_j(t) - \bar{X}_{z^c}(t))(X_j(t) - \bar{X}_{z^c}(t))^\top dt.$$

Enfin, les moyennes sont définies comme suit :

$$\bar{X}_g(t) = \frac{1}{|g|} \sum_{i;s_i \in g} X_i(t), \quad \bar{X}(t) = \frac{1}{n} \sum_{i=1}^{n} X_i(t).$$

Approche dans le cas multivarié

Frévent et al. (2021b) ont également proposé une statistique de scan spatial sans distribution pour les données fonctionnelles multivariées afin de tester H_0 contre l'hypothèse alternative composite associée à un cluster potentiel z.
En résumé, ils supposent que pour chaque instant t, on a :

$$\mathbb{V}\text{ar}(X_i(t)) = \Sigma(t,t), \quad \forall i \in \{1,\dots,n\}$$

où Σ est une fonction de matrice de covariance $p \times p$.
Ils définissent alors la **statistique de scan spatial fonctionnelle multivariée sans distribution (MDFFSS)** comme :

$$\Lambda_{\text{MDFFSS}} = \max_{z \in Z} T^z_{n,\max}$$

avec

$$T^z_{n,\max} = \sup_{t \in T} T_n(t)^z,$$

où $T_n(t)$ est la statistique ponctuelle définie par le test de Hotelling T^2-test statistic (Qiu et al., 2021) :

$$T_n(t)^z = \frac{|z||z^c|}{n}(\bar{X}_z(t) - \bar{X}_{z^c}(t))^\top \hat{\Sigma}(t,t)^{-1}(\bar{X}_z(t) - \bar{X}_{z^c}(t)).$$

Approche dans le cas multivarié

La matrice de covariance empirique est définie comme suit :

$$\hat{\Sigma}(s,t) = \frac{1}{n-2}\left[\sum_{i;s_i \in z}(X_i(s) - \bar{X}_z(s))(X_i(t) - \bar{X}_z(t))^\top + \sum_{i;s_i \in z^c}(X_i(s) - \bar{X}_{z^c}(s))(X_i(t) - \bar{X}_{z^c}(t))^\top\right]. \quad (1)$$

Choix de la Méthode Appropriée

Le choix entre une approche paramétrique et une approche non paramétrique dépend des données et du contexte.

- **Frévent et al. (2021a,b)** ont étudié cette question à travers des simulations.
- Les méthodes **ponctuelles** (DFFSS, URBFSS, MDFFSS, MRBFSS) sont les plus performantes.

Performances selon le type de données :

- **Cas gaussien** : DFFSS et MDFFSS sont les plus adaptées.
- **Cas non gaussien** : URBFSS et MRBFSS sont préférables (mieux adaptées aux valeurs extrêmes).

Avantages et inconvénients :

- **Méthodes paramétriques** : Moins de faux positifs, mais peuvent ne pas détecter certains clusters.
- **Méthodes non paramétriques** : Détectent plus de clusters mais augmentent le risque de faux positifs.

Recommandations :

- **Minimiser les faux positifs** (coût élevé d'analyse locale) $\Rightarrow$ Méthode **paramétrique**.
- **Détecter tous les clusters (même avec de faux positifs)** $\Rightarrow$ Méthode **non paramétrique**.

Conclusion : Le choix dépend de la **distribution des données** et du compromis entre **taux de vrais positifs** et **taux de faux positifs**.

Introduction

En pratique, les données fonctionnelles sont souvent observées et enregistrées de **manière discrète** sous la forme d'un ensemble de n **paires** (t_j, y_j), où y_j représente un instantané de la fonction au temps t_j, potentiellement altéré par une **erreur de mesure**.
La plupart des travaux consacrés aux **données fonctionnelles** considèrent les données comme des **trajectoires d'un processus stochastique à valeurs réelles**, $X = \{X^{(j)}(t), t \in \mathcal{T}\}$, où $X^{(j)}(t) \in \mathbb{R}^p$, avec $p \geq 1$, et où $\mathcal{T}$ est un intervalle de $\mathbb{R}$.

Cet article présente **l'analyse des données fonctionnelles catégorielles** comme une **extension de l'analyse des correspondances multiples aux données fonctionnelles**, ainsi que son implémentation dans le **package R cfda.**

L'article **cfda** (Categorical Functional Data Analysis) s'intéresse au cas où **le modèle stochastique sous-jacent générant les données** est un **processus stochastique en temps continu**, $X = \{X^{(j)}(t), t \in \mathcal{T}\}$, où, pour tout $t \in \mathcal{T}$, $X^{(j)}(t)$ est une **variable aléatoire catégorielle** plutôt qu'une variable à valeurs réelles.

Définition

- **Variables catégorielles** : Contrairement aux processus fonctionnels classiques qui prennent des **valeurs réelles**, ici les observations prennent des **valeurs discrètes** (ex. : "rouge", "vert", "bleu" ou "malade", "guéri", etc.)
- **Processus stochastique en temps continu** : Un processus où l'évolution des états se fait **à des instants aléatoires** dans un intervalle continu de temps.
- **Processus de saut** : Un processus qui évolue par transitions successives entre différents états, à des temps de saut aléatoires.
- **État absorbant** : Un état à partir duquel le processus ne **peut plus évoluer**. Par exemple, dans une maladie, "décès" est un état absorbant.

L'article propose deux façons de représenter ces données :

- **Sans ordre :** Les états sont indépendants (ex. "bleu", "rouge", "vert").
- **Avec ordre :** Les états suivent une progression logique (ex. "léger", "modéré", "sévère").

Modélisation

Soit $(\Omega, \mathcal{A}, P)$ un espace probabilisé, et $\mathcal{S} = \{s_1, \ldots, s_K\}$ un ensemble fini de K états, avec $K \geq 2$.
Le processus X est alors une famille de variables aléatoires catégorielles indexées par $\mathcal{T}$:

$$X = \{X_t; X_t : \Omega \to \mathcal{S}, \quad t \in \mathcal{T}\}.$$

Ainsi, pour un certain $\omega \in \Omega$, une trajectoire du processus X, notée $X(\omega)$, est une séquence d'états $s_{ij} = s_{ij}(\omega)$ et d'instants de transition $t_i = t_i(\omega)$ entre ces états :

$$\{(t_0, s_{i_0}), (t_1, s_{i_1}), (t_2, s_{i_2}), \ldots\}$$

où $0 = t_0 < t_1 < t_2 < \ldots$ sont les temps de saut dans $\mathcal{T}$, et où $s_{ij} \in S$ avec $ij \in \{1, \ldots, K\}$ pour tout $j \geq 0$.
Interprétation de la trajectoire
Cette trajectoire peut être interprétée de la manière suivante :

- À l'instant $t_0 = 0$, l'état est s_{i_0}.
- À l'instant t_1 $(t_1 > t_0)$, un saut se produit, et le processus passe de s_{i_0} à s_{i_1}.
- À l'instant t_2 $(t_2 > t_1)$, un autre saut a lieu, et le processus passe de s_{i_1} à s_{i_2}, etc.

Si $\mathcal{T}$ correspond à un intervalle de temps $[0, \mathcal{T}]$ pour un certain $\mathcal{T} > 0$, alors l'observation du processus s'arrête lorsque le temps $\mathcal{T}$ est atteint ou lorsqu'un état absorbant est atteint.

Exemple illustratif : Évolution d'un patient dans un hôpital

Imaginons un patient suivi dans un hôpital, où son état clinique évolue selon un processus stochastique.

Définition des éléments du modèle :

- **Espace d'états** :

$$S = \{\text{Sain}, \text{Malade}, \text{Hospitalisé}, \text{Guéri}, \text{Décédé}\}$$

  Chaque état correspond à une situation médicale du patient.

- **Processus stochastique** :

$$X_t, \quad t \in \mathcal{T}$$

  À chaque instant t, le patient est dans l'un des états de S.

- **Temps de saut** : Le passage d'un état à un autre se produit à des instants aléatoires $t_1, t_2, \ldots$.

Exemple de trajectoire du processus

Une trajectoire possible du processus pour un patient pourrait être :

$$\{(t_0 = 0, \text{Sain}), (t_1 = 5, \text{Malade}), (t_2 = 10, \text{Hospitalisé}), (t_3 = 20, \text{Guéri})\}$$

Interprétation :

- $t_0 = 0$: Le patient est en bonne santé.
- $t_1 = 5$: Il tombe malade après 5 jours.
- $t_2 = 10$: Il est hospitalisé après 10 jours.
- $t_3 = 20$: Il guérit après 20 jours.

Le processus s'arrête à $\mathcal{T} = 20$, car l'état "Guéri" est atteint et plus de transitions ne sont observées.

Modélisation des Données Fonctionnelles Catégorielles

Problématique :

- Extension de l'Analyse des Correspondances Multiples (ACM) aux données fonctionnelles.
- On considère un **processus temporel** X_t prenant des valeurs discrètes dans un ensemble d'états $S = \{s_1, ..., s_K\}$.
- **Objectif :** Extraire des **composantes principales** qui expliquent la variabilité de X_t dans le temps.

Représentation des Données

Transformation en indicateurs binaires :

- Pour chaque état x et chaque instant t, on définit une variable indicatrice :

$$1_x^t = \begin{cases} 1, & \text{si } X_t = x, \\ 0, & \text{sinon.} \end{cases}$$

Nous définissons les probabilités suivantes :

- $p^x(t) = P(X_t = x)$: la probabilité que le processus X_t soit dans l'état x à l'instant t.
- $p^{x,y}(t,s) = P(X_t = x, X_s = y)$: la probabilité conjointe que $X_t = x$ et $X_s = y$.

Les hypothèses générales considérées dans ce cadre sont :
Hypothèse 1 (H1) : Le processus X est continu en probabilité :

$$\lim_{h \to 0} P(X_{t+h} \neq X_t) = 0.$$

Hypothèse 2 (H2) : Pour tout instant $t \in [0, T]$ (sauf éventuellement un ensemble fini d'instants discrets), chaque état a une probabilité strictement positive d'apparaître :

$$p^x(t) \neq 0, \quad \forall x \in S, \quad \forall t \in [0, T].$$

Extraction des Composantes Principales

Trois formulations équivalentes du problème des valeurs propres :

- **Équation intégrale stochastique (5) :** Maximisation de la corrélation avec le processus X_t.
- **Problème d'évaluation conditionnelle (8) :** Relation entre X_t et les valeurs propres.
- **Équation aux valeurs propres discrétisée (12) :** Dépend des probabilités de transition $P(X_t = x, X_s = y)$ et des fonctions d'encodage optimal $a^x(t)$.

Résultats :

- **Composantes principales** z_i : projections optimales du processus X_t.
- **Fonctions d'encodage optimal** $a_i^x(t)$: transformation des données catégorielles en espace continu.

Approximations et Réduction de Dimension

Réduction de dimension :

- Comme en ACP, on garde les q premières composantes principales.
- Permet :
  - **Visualisation** des trajectoires dans un espace réduit.
  - **Clustering** et régression sur les données encodées.

Approximation des fonctions optimales :

- Les fonctions $a_i^x(t)$ ne sont pas calculables explicitement.
- Approche par **décomposition en base fonctionnelle** (ex : bases de Fourier, splines).
- **Avantages :**
  - Réduction du **coût de calcul**.
  - Meilleur contrôle de la **précision**.

Litterature

L'auteur mentionne que les **travaux existants** sur les **données fonctionnelles catégorielles** se concentrent principalement sur l'extension des techniques factorielles, notamment :

- L'analyse canonique
- L'analyse des correspondances multiples (ACM)

es méthodes, initialement conçues pour **des données catégorielles classiques**, ont été adaptées aux **données fonctionnelles**, et cette adaptation est appelée "**analyse harmonique**" par les auteurs de ces travaux.

Nous pouvons citer :

- Deville, J.C.; Saporta, G. Correspondence Analysis with an Extension towards Nominal Time Series. J. Econom. 1983, 22, 169–189.
- Deville, J.C. Analyse de Données Chronologiques Qualitatives : Comment Analyser des Calendriers? Ann. De L'INSEE 1982, 45, 45–104.
- Boumaza, R. Contribution à l'Étude Descriptive d'une Fonction Aléatoire Qualitative. Ph.D. Thesis, Université Paul Sabatier, Toulouse, France, 1980.
- Saporta, G. Méthodes Exploratoires d'Analyse de Données Temporelles; Université Pierre et Marie Curie: Paris, France, 1981.