Synthèse des Scores ACM à partir de Données Catégorielles Fonctionnelles

Moustapha Sarr

31/03/2025

1 Introduction

On cherche à extraire des scores ACM à partir d'un processus catégoriel X_t défini sur $[0, T]$ avec des valeurs dans un ensemble fini $S = \{s_1, \ldots, s_K\}$.

2 Étapes principales

2.1 Transformation en variables binaires

Pour chaque catégorie $x \in S$, on définit la variable indicatrice binaire associée:

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

Cela permet de convertir les données catégorielles en une représentation numérique exploitable pour l'analyse en composante principale (ACP).

2.2 Calcul des probabilités

- Probabilité marginale : $p^x(t) = P(X_t = x)$.
- Probabilité conjointe : $p^{x,y}(t, s) = P(X_t = x, X_s = y)$.

2.3 Hypothèses

- **Hypothèse** H_1 : Le processus X est continu en probabilité

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

  Ce qui signifie que de petits changements dans le temps entraînent rarement des changements brusques d'état.

- **Hypothèse** H_2 : Pour chaque instant $t \in [0, T]$ (à l'exception possible d'un nombre fini de temps discrets), chaque état a une probabilité strictement positive d'apparaître :

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

2.4 L'opérateur d'espérance conditionnelle

Nous défénissons l'opérateur d'espérance conditionnelle associé à X_t, par :

$$E_t : L^2(W) \to L(X_t), \quad z \in L^2(W), \quad z \mapsto E_t(z) = \sum_{x \in S} \mathbb{E}(z \mid X_t = x) \mathbf{1}_x^t$$

- $L^2(W)$: L'espace des variables aléatoires avec une variance finie.
- $L(X_t)$: L'espace linéaire engendré par les indicateurs 1_t^x

Définition du coefficient $\eta^2(z; X_t)$

$$\eta^2(z; X_t) = \frac{Var(E_t(z))}{Var(z)}$$

- C'est un coefficient de corrélation entre la variable aléatoire z qu'on cherche à construire et la variable catégorielle X_t.
- $Var(E_t(z))$: Variance expliquée.
- $Var(z)$: Variance totale.

Si on a plusieurs instants $t_1, t_2, \ldots, t_p$, alors on cherche un z qui est globalement bien expliquée par toutes ces variables catégorielles. On fait :

$$\text{maximiser} \sum_{i=1}^{p} \eta^2(z; X_t)$$

C'est à dire trouver une variable aléatoire z qui résume au mieux l'information partagée par plusieurs variables catégorielles. Cette variable devient la première composante principale.

Extension au cas fonctionnel

Dans le cas fonctionnel, on ne se contente plus de quelques instants, considère tous les instants $t \in [0, T]$. Alors on maximise :

$$\int_0^T \eta^2(z; X_t) dt \tag{1}$$

C'est à dire trouver une variable aléatoire z qui est au maximum corrélée avec l'ensemble du processus X_t sur toute la durée d'observation.

2.5 Problème aux valeurs propres

Sous les hypothèses H_1 et H_2, z qui maximise (1) est associé à la plus grande valeur propre du problème aux valeurs propres stochastique :

$$\int_0^T E_t(z) dt = \lambda z \tag{2}$$

. L'opérateur $Q = \int_0^T$ est positif, hermitien et compact. Cela signifie que :

- Ses valeurs propres sont bien définies et positives.

- L'ensemble de ses vecteurs propres $\{z_i\}_i \geq 1$ est comptable.
- On peut ordonner ses valeurs propres par importance :

  $$\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$$

  .

  Les variables $\{z_i\}_i \geq 1$ sont les composantes principales. Elles sont centrées et non corrélés.

2.6 Résolution de l'équation d'auto-valeur (2)

L'idée est d'exprimer z sous une forme alternative en introduisant une fonction ψ_t telle que:

$$\psi_t = \frac{1}{\lambda} E_t(z), \forall t \in [0, T]$$

Expression de z en fonction de ψ_t
En intégrant la définition de ψ_t, on retrouve :

$$z = \int_0^T \psi_t dt. \tag{3}$$

2.7 Nouvelle formulation sous forme d'un problème aux valeurs propres

$$\int_0^T K(t, s)\psi_s ds = \lambda\psi_t, \forall t \in [0, T] \tag{4}$$

avec $K(t, s) = E_t E_s$.
Normalisation pour rendre les solutions uniques

$$\int_0^T Var(\psi_t)dt = \int_0^T \mathbb{E}(\psi_t^2)dt = 1$$

2.8 Relation entre la variance de z et λ

A partir de la contrainte précédente et de l'équation (3), on obtient

$$Var(z) = \mathbb{E}(z^2) = \lambda$$

. La variance de la première composante principale est égale à la plus grande valeur propre λ.

2.9 Définition des fonctions d'encodage optimal

$$\psi_t = \sum_{x \in S} a^x(t) 1_t^x$$

Les coefficients $\{a^x\}_{x \in S}$ sont des fonctions déterministes, appelées fonctions d'encodage optimal, car elle permettent de transformer un état x en une valeur numérique optimisé.

En remplaçant ψ_t dans l'équation (4), on obtient

$$\int_0^T \sum_{y\in S} p^{x,y}(t,s)a^y(s)ds = \lambda a_x(t)p_x(t), \quad \forall t \in [0,T], \forall x \in S$$

Nouvelle contarainte de normalisation

$$\int_0^T \sum_{x\in S}[a^x(t)]^2 p^x(t)dt = 1$$

Cette équation impose que les fonctions d'encodage optimal $a^x(t)$ aient une variance totale normalisée à 1.

2.10 Expression des composantes principales z_i en fonction des $a^x(t)$

Finalement, on peut exprimer les composantes principales z_i comme :

$$z_i = \int_0^T \sum_{x\in S} a_i^x(t) 1_t^x dt, \quad \forall i \geq 1$$

Chaque composante principale z_i est obtenue en intégrant dans le temps les fonctions d'encodage $a_i^x(t)$.

Expansion des indicatrices 1_t^x en termes des composantes principales

$$1_t^x = \sum_{i\geq 1} z_i a_i^x(t)\frac{1}{p^x(t)}, \quad \forall x \in S$$

. C'est l'analogue de la décomposition de **Karhunen-Loève**, mais appliqués à des données catégorielles.

Décomposition de la probabilité jointe avec le théorème de Mercer

$$p^{x,y}(t,s) = p^x(t)p^y(s)\sum_{i\geq 1}\lambda_i a_i^x(t)a_i^y(t)$$

Probabilité marginale $p^x(t)$
Si on pose $x = y$ et $t = s$, on obtient

$$p^x(t) = \left(\sum_{i\geq 1}[a_i^x(t)]^2\right)^{-1}$$

Cela relie directement les probabilités marginales aux fonctions propres.

2.11 Réduction de dimension

L'idée est d'approximer X en utilisant un nombre limité de composantes principales :

$$1_t^x = \sum_{i\geq 1}^{q} z_i a_i^x(t)\frac{1}{p^x(t)}, \quad \forall x \in S$$

Au lieu d'utiliser toutes les composantes principales, on peut se limiter aux q premières pour un bonne approximation.

Adaptation de l'espace fonctionnel aux données catégorielles

Dans le cas classique des données fonctionnelles réelles ou vectorielles, on travaille dans l'espace de Hilbert :

$$L^2(T, \mathbb{R}^p) = \left\{ X : T \to \mathbb{R}^p \;\middle|\; \int_T \|X(t)\|^2 dt < \infty \right\}.$$

Cependant, dans le cas **catégoriel**, les trajectoires prennent leurs valeurs dans un ensemble fini :

$$S = \{s_1, s_2, \ldots, s_K\}, \quad \text{avec } X(t) \in S.$$

Représentation par encodage binaire

Pour chaque état $x \in S$, on définit une fonction indicatrice :

$$\mathbf{1}_t^x = \begin{cases} 1 & \text{si } X_t = x \\ 0 & \text{sinon} \end{cases}$$

On peut alors représenter chaque trajectoire $X(t)$ par le vecteur :

$$\tilde{X}(t) = (\mathbf{1}_t^{s_1}, \ldots, \mathbf{1}_t^{s_K}) \in \{0, 1\}^K.$$

Encodage optimal et espace L^2 réel

À l'aide d'un *encodage optimal* $\{a_x^i(t)\}_{x \in S}$, on projette la trajectoire dans l'espace réel :

$$x_i(t) = \sum_{x \in S} a_x^i(t) \cdot \mathbf{1}_t^x \quad \in L^2(T, \mathbb{R}).$$

Les fonctions $x_i(t)$ représentent des composantes principales, à partir desquelles on peut travailler dans $L^2(T, \mathbb{R})$.

Conclusion : Grâce à cette projection, les données fonctionnelles catégorielles sont étudiées dans un espace fonctionnel réel, permettant l'application d'outils classiques tels que l'ACP fonctionnelle, le clustering ou la régression.