Propriétés et interprétation de la covariance relationnelle en ACP

Properties and interpretation of relational covariance on PCA

Sullivan Hidot *, Jean-Yves Lafaye et Christophe Saint-Jean
L3I – Université de La Rochelle, UPRES EA 1216, avenue Michel Crépeau, 17042 La Rochelle cedex 1,
{shidot,jylafaye,csaintje}@univ-lr.fr

Manuscrit reçu le 7 octobre 2005

Résumé et mots clés

Cet article s'intéresse à l'étude des propriétés de l'Analyse en Composantes Principales Relationnelle (ACPR) qui analyse un vecteur aléatoire conditionnellement à la réalisation d'un paramètre induisant une relation binaire sur l'espace probabilisé de référence. Nous détaillons les propriétés de la covariance et de l'espérance relationnelles qui sont à la base de cette technique d'analyse connue mais finalement peu étudiée. L'article présente quelques illustrations des propriétés que nous mettons en évidence, et qui éclairent les interprétations en ACPR.

Covariance relationnelle, ACP, Laplacien de graphe, Analyse de Données Spatio-temporelles.

Abstract and key words

This paper is dedicated to the study of the main properties of the so called 'Relational Principal Components Analysis' (RPCA), that achieves the analysis of a random vector, with respect to the prior knowledge of one binary relationship upon the underlying probabilistic space. We detail the relational covariance and expectation properties that are the grounds of this technique, which whilst not being novel, remains scarcely studied. The paper presents with didactic examples for the properties we previously addressed and throw some light on interpretations in RPCA.

Relational covariance, PCA, Graph Laplacian, Spatio-temporal Data Analysis.

Introduction

Notre travail a été motivé par une recherche appliquée à l'étude de trajectoires dans des contextes variés d'analyse du comportement : mouvement dansé, biologie marine, médecine, analyse de scènes... D'une manière générale, les données expérimentales sont exploitées à des fins exploratoires par des techniques de segmentation et de reconnaissance de formes, mais des questions d'archivage peuvent conduire à s'intéresser à la compression et l'indexation des données. Les méthodes factorielles sont adaptées pour traiter ces questions et l'on peut citer des exemples d'utilisation de l'ACP classique pour l'étude de phénomènes temporels dans des domaines applicatifs divers tels que la biologie moléculaire [BWOS96], l'animation faciale [KMMT01] ou encore la reconnaissance de mouvements [WHT02].

Dans le cas de données numériques, l'ACP standard se place dans le contexte d'un espace probabilisé Ω, et optimise des critères d'inertie pour rendre compte des corrélations entre

* Auteur de référence : shidot@univ-lr.fr, tél : +33-5-4645-8324, fax : +33-54646-8242

variables et des distances mutuelles entre individus dans des espaces de dimension réduite. L'analyste dispose souvent d'informations complémentaires non probabilistes, sur la structure de l'espace Ω. Ces informations sont apportées par des variables exogènes quantitatives ou non, par exemple de natures temporelle, spatiale, sociologique,... Dans notre exemple traitant de trajectoires, la suite des dates de capture peut induire une structure d'ordre sur les observations. Cette information exogène n'est pas directement prise en compte par l'ACP. Il est toujours possible de surcharger les plans factoriels d'une ACP en représentant par exemple les trajectoires des individus pour introduire l'ordre temporel dans l'analyse. Certaines approches utilisent le temps comme variable instrumentale (ACPVI) [PL97, Sab87] mais on en trouve relativement peu d'exemples d'applications dans la littérature.

Le temps n'est qu'une dimension structurante parmi d'autres et la prise en compte de relations topologiques ou géographiques par exemple entre les observations a donné lieu à de nombreux développements. Les analyses locales fondées sur les variogrammes et les corrélogrammes [Mat65, Kri51] établissent les prémisses des méthodes factorielles dédiées à l'analyse de données géologiques. On verra ensuite apparaître de façon plus ou moins concomitante plusieurs déclinaisons d'une même méthode factorielle sous des appellations variées mettant en avant différents points de vue. « L'ACP pondérée » [Fol82] insiste sur l'introduction d'une pondération des couples d'observations parallèlement à la pondération classique induite par la mesure de probabilité sur Ω. « L'ACP locale » [BL84] exploite les structures de proximité mutuelle entre observations de façon conforme aux principes établis par Matheron. « L'ACP lissée » [BE90] s'intéresse au filtrage de type « moyenne mobile » ou « différences premières » des variables analysées. Dans [RSV90], « l'ACP sous contrainte » permet d'analyser des structures moléculaires tenant compte de lois physiques d'interaction entre particules.

Nous retenons de toutes ces déclinaisons (pour lesquelles un exposé théorique est donnée dans [LMP00]) qu'elles ont en commun d'appliquer la démarche classique de l'ACP, en introduisant dans le modèle initial une information de type relationnel, à savoir une relation binaire entre les individus. Cette relation, selon qu'elle traduit un ordre, une équivalence, une distance, etc, donne lieu à un développement spécifique et à une dénomination particulière. Des approches fédératrices existent, telles que l'ACPVI déjà citée, ou encore l'ACP à noyaux [SSM96] qui proposent un formalisme très général capable de représenter une majeure partie des techniques factorielles citées plus haut ainsi que d'autres méthodes : analyse des correspondances, canonique, discriminante...

Notre article présente un formalisme qui permet d'intégrer n'importe quel type de relation binaire à l'ACP classique. En conséquence, nous utiliserons désormais le terme « ACP relationnelle » – que nous empruntons à [JN02] – et qui a le mérite de mettre l'accent sur le caractère relationnel de l'information intégrée au modèle sans rien supposer d'autre sur la forme de cette relation. Notre présentation décline les différentes approches évoquées plus haut.

Nos principales contributions sont les suivantes. La relation étudiée est directement introduite dans le modèle probabiliste initial sous forme de variable aléatoire. Nous nous plaçons dans un contexte continu car les formules explicitant les critères optimisés sont alors plus simples et plus naturelles. Nous proposons ensuite un ensemble de réécritures des paramètres usuels manipulés par l'ACP (inertie, variance, covariance, corrélation, moyenne). Chaque formulation permet de mettre en évidence et d'appréhender une interprétation particulière de la quantité concernée. Ces interprétations utilisent successivement des considérations probabilistes ou géométriques. Il ne s'agit pas d'explorer les formalismes, mais bien d'expliciter et de justifier l'ensemble des conclusions susceptibles d'être obtenues via l'ACP relationnelle.

Dans une première section, nous rappelons brièvement le principe et les objectifs de l'ACP standard dans le cadre de l'analyse d'un vecteur de p variables aléatoires réelles de $\mathcal{L}^2(\Omega)$. La seconde section présente l'ACP relationnelle (ACPR) comme une généralisation de l'ACP standard, lorsque l'on dispose d'une information relationnelle exogène supportée par le graphe d'une relation « g » quelconque définie sur $\Omega \times \Omega$. L'ACPR est fondée sur la notion de « covariance relationnelle » attachée à la relation g. Cette section précise les notations, fait l'inventaire des différentes expressions de la covariance relationnelle et explicite leur sens, leurs relations mutuelles et leurs propriétés. Dans l'objectif de préciser et justifier les modalités d'interprétation pratique des résultats de l'ACPR, une formulation de la covariance relationnelle particulièrement intéressante est présentée en section 3. Elle fait référence à la notion que nous introduisons sous le nom d'« espérance relationnelle ». La section 4 évoque les questions d'échantillonnage et de mise en œuvre effective de l'ACPR. Elle introduit la section 5 qui présente deux exemples didactiques destinés à mettre clairement en évidence l'interprétation de la covariance et de l'espérance relationnelles dans deux contextes simples d'analyse de données temporelles.

1. ACP standard

On dispose d'un espace probabilisé $(\Omega, \mathcal{T}, P)$, où Ω est un ensemble constitué d'éléments appelés *individus*, $\mathcal{T}$ et P désignent respectivement une tribu et une mesure de probabilité sur Ω. Soit $\mathbb{X} = (X_1, \ldots, X_p)$ un vecteur aléatoire constitué de p variables réelles centrées, de carré intégrable et définies sur $(\Omega, \mathcal{T}, P)$. Pour deux variables X_i et X_j, le produit scalaire usuel sur $\mathcal{L}^2(\Omega)$ est donné par la relation :

$$< X_i, X_j >= \int_\Omega X_i(\omega) X_j(\omega) dP(\omega) \qquad (1)$$

On suppose en outre $\mathbb{R}^p$ muni d'une métrique $M = (m_{ij})_{i,j=1}^p$. L'ACP [Jol86] est une méthode factorielle dont le but est de déterminer des nouvelles variables pertinentes permettant d'expliquer les corrélations éventuelles entre les X_i. Ces nouvelles variables, appelées *composantes principales* (CP), sont obtenues par combinaisons linéaires des X_i pondérées par les coefficients des vecteurs propres de VM où V est la matrice de covariance entre les variables :

$$V_{ij} = cov(X_i, X_j) =< X_i, X_j > \tag{2}$$

Si $v = (v_1, ..., v_p)$ est un vecteur propre M-normé de VM, la CP associée est donnée par :

$$\forall \omega \in \Omega, \qquad C(\omega) = \sum_{i,j=1}^{p} v_j m_{ij} X_i(\omega) \tag{3}$$

Comme les vecteurs propres de VM forment une base M-orthonormée de l'espace des individus, les CP associées sont non corrélées, engendrant ainsi une base sur l'espace des variables. Soient $\mathcal{C}$ la matrice des CP de $\mathbb{X}$, U la matrice des vecteurs propres M-normés de VM, et $(\lambda_i)_{i=1}^p$ les valeurs propres associées.
La projection M-orthogonale d'un individu ω sur le j-ème vecteur propre est égale à la mesure de la j-ème CP sur ω.
La coordonnée de la variable X_i sur la CP réduite $C_j/\sqrt{\lambda_j}$ est égale au produit de la i-ème coordonnée du j-ème vecteur propre par la racine carrée de la valeur propre associée.

2. ACP Relationnelle

L'espace Ω est sans structure autre que probabiliste. Dans la réalité, on dispose souvent d'informations complémentaires sur les liaisons entre les individus telles que proximité spatiale, temporelle, sociale, ... Une relation binaire est une forme générale pour représenter ce genre d'information. En considérant des liens éventuels entre les individus, l'ACPR s'impose naturellement pour rendre compte de cette information.
L'ACPR est basée sur la construction d'un graphe que l'on suppose dépendre d'un paramètre θ qui exprime les liens entre les individus. Ce graphe est induit par la variable aléatoire g_θ définie sur $\Omega \times \Omega$ par la relation :

$$g_\theta(\omega,\omega') = \begin{cases} 1 & \text{si } \omega \text{ est connecté à } \omega' \\ 0 & \text{sinon} \end{cases} \tag{4}$$

Définition 1 (Covariance globale). *La covariance globale entre deux variables X et Y se calcule en utilisant la formule quadratique :*

$$Cov(X,Y) = \frac{1}{2} \iint_{\Omega\times\Omega} (X(\omega) - X(\omega'))(Y(\omega) - Y(\omega'))\, dP_\omega dP_{\omega'} \tag{5}$$

En calculant cette covariance via g_θ et en introduisant le facteur de normalisation K_θ qui représente la connectivité du graphe, (5) s'exprime conformément à la définition suivante :

Définition 2 (Covariance relationnelle). *La covariance relationnelle entre deux variables X et Y relativement au graphe induit par g_θ se calcule grâce à la formule :*

$$Cov_\theta(X,Y) = \frac{1}{2K_\theta} \iint_{\Omega\times\Omega} (X(\omega) - X(\omega'))(Y(\omega) - Y(\omega'))\, g_\theta(\omega,\omega')\, dP_\omega dP_{\omega'} \tag{6}$$

avec

$$K_\theta = \iint_{\Omega\times\Omega} g_\theta(\omega,\omega')\, dP_\omega dP_{\omega'} \tag{7}$$

Si g_θ est constante égale à 1 alors le graphe est complet et le résultat est la covariance globale. Sinon, seuls interviennent dans la formule les couples d'individus en relation par g_θ.
La covariance relationnelle est analogue à la notion de semi-variogramme croisé, très utilisée en géostatistique [Mat65] pour analyser un processus aléatoire de variables régionalisées. Le variogramme est utilisé dans le krigeage [Cre93], méthode d'interpolation spatiale qui estime la valeur d'un phénomène naturel en des sites non observés. Notre approche est différente car elle n'utilise pas le variogramme comme un outil mais comme un objet d'étude apportant de l'information sur l'interaction et la dispersion des phénomènes.
En développant (6), on obtient :

$$Cov_\theta(X,Y) = \frac{1}{K_\theta} \iint_{\Omega\times\Omega} (X(\omega) - X(\omega'))Y(\omega) \left(\frac{g_\theta(\omega,\omega') + g_\theta(\omega',\omega)}{2} \right) dP_\omega dP_{\omega'} \tag{8}$$

L'application G_θ qui à (ω,ω') associe $(g_\theta(\omega,\omega') + g_\theta(\omega',\omega))/2$ est symétrique et on a donc :

$$Cov_\theta(X,Y) = \frac{1}{K_\theta} \iint_{\Omega\times\Omega} (X(\omega) - X(\omega'))Y(\omega) G_\theta(\omega,\omega')\, dP_\omega dP_{\omega'} \tag{9}$$

En permutant les indices ω et ω', on obtient :

$$Cov_\theta(X,Y) = \frac{1}{K_\theta} \iint_{\Omega\times\Omega} (X(\omega') - X(\omega))Y(\omega') G_\theta(\omega,\omega')\, dP_\omega dP_{\omega'} \tag{10}$$

En additionnant (9) et (10), on en déduit :

$$Cov_\theta(X,Y) = \frac{1}{2K_\theta} \iint_{\Omega\times\Omega} (X(\omega) - X(\omega'))(Y(\omega) - Y(\omega'))\, G_\theta(\omega,\omega')\, dP_\omega dP_{\omega'} \tag{11}$$

La covariance relationnelle utilise implicitement une version symétrisée du graphe initial g_θ. En fait, l'ACPR porte sur des classes d'équivalence de graphes où deux graphes, notés G et H, sont équivalents si les symétrisés coïncident :

$$G \sim H \Longleftrightarrow (G + G^t = H + H^t) \tag{12}$$

Le graphe valué symétrique G_θ définit plus haut peut prendre trois valeurs : 0, 1 ou 1/2. Le cas où $G_\theta(i,j) = 0$ signifie que i et j ne sont pas connectés pour g_θ. Celui où $G_\theta(i,j) = 1$ signifie que i et j sont connectés par deux arcs opposés. Enfin, $G_\theta(i,j) = 1/2$ signifie que i et j sont connectés par un arc unique. Dans la suite, on pourra donc supposer sans perte de généralité que le graphe est symétrique. De même, la présence d'arêtes réflexives n'a pas d'influence sur l'analyse.
La covariance relationnelle se définit également par l'opérateur du Laplacien du graphe. En effet, soient A_θ et D_θ les opérateurs de $\mathcal{L}^2(\Omega)$ dans $\mathcal{L}^2(\Omega)$:

$$A_\theta(X) = \int_\Omega X(\omega) G_\theta(\omega,.) dP_\omega \tag{13}$$

$$D_\theta(X) = \int_\Omega X(\omega) G_\theta(\omega,.) deg(\omega) \delta_{.\omega} dP_\omega \tag{14}$$

où δ est le symbole de Kronecker et $deg(\omega)$ est une application de Ω dans $[0,1]$ appelée *degré* de l'individu ω :

$$deg(\omega) = \int_\Omega G_\theta(\omega,\omega') dP_{\omega'} \tag{15}$$

Définition 3 (Laplacien du graphe). *Le Laplacien du graphe Δ_θ est défini comme la différence entre les opérateurs D_θ et A_θ [Moh91] :*

$$\Delta_\theta = D_\theta - A_\theta \tag{16}$$

Il s'ensuit la définition de la matrice de covariance relationnelle :

Définition 3 (Matrice de covariance relationnelle). *La matrice de covariance relationnelle V_θ s'écrit :*

$$V_\theta(i,j) = \frac{1}{K_\theta} < X_i, \Delta_\theta(X_j) > = \frac{1}{K_\theta} < X_i, X_j >_{\Delta_\theta} \quad 1 \le i,j \le p \tag{17}$$

Par linéarité des opérateurs A_θ et D_θ et d'après le théorème de Fubini, $< .,. >_{\Delta_\theta}$ est une forme bilinéaire symétrique sur $\mathcal{L}^2(\Omega) \times \mathcal{L}^2(\Omega)$.
De plus, $\mathcal{K}er(\Delta_\theta)$ est l'espace engendré par la variable constante unitaire. Il est donc inutile de centrer préalablement les variables pour déterminer la matrice de covariance si l'on utilise Δ_θ. On en déduit que la covariance relationnelle est invariante par translation des individus.
L'ACPR est un cas général de l'ACP standard où l'on introduit un graphe de relation de voisinages connectant un certain nombre d'individus. Si cette connexion est totale (*i.e.* G_θ graphe complet) alors on retrouve l'ACP classique.

3. Espérance relationnelle

Nous allons définir dans cette section la notion d'espérance, conditionnellement à un graphe de connexité afin de donner une interprétation complémentaire de la covariance relationnelle.
Comme G_θ induit un graphe symétrique, on peut utiliser (10) et écrire :

$$Cov_\theta(X,Y) = \frac{1}{K_\theta} \iint_{\Omega\times\Omega} (X(\omega)Y(\omega) - X(\omega)Y(\omega')) G_\theta(\omega,\omega') dP_\omega dP_{\omega'} \tag{18}$$

qui devient par factorisation :

$$Cov_\theta(X,Y) = \frac{1}{K_\theta} \int_\Omega X(\omega) \left(Y(\omega) - \frac{1}{deg(\omega)} \int_\Omega Y(\omega') G_\theta(\omega,\omega') dP_{\omega'} \right) deg(\omega) dP_\omega \tag{19}$$

Définition 5 (Espérance relationnelle). *Posons pour tout $\omega \in \Omega$:*

$$E(Y|G_\theta)(\omega) = \frac{1}{deg(\omega)} \int_\Omega Y(\omega') G_\theta(\omega,\omega') dP_{\omega'} \tag{20}$$

$E(Y|G_\theta)(\omega)$ est appelée espérance relationnelle de la variable Y mesurée sur le voisinage de l'individu ω selon le graphe.

L'espérance relationnelle $E(Y|G_\theta)(\omega)$ est une fonction du paramètre aléatoire θ de $\mathcal{L}^2(\Omega)$. En conséquence, celle-ci peut également être vue comme une variable aléatoire de $\mathcal{L}^2(\Omega)$.
Avec ces notations :

$$Cov_\theta(X,Y) = \frac{1}{K_\theta} \int_\Omega X(\omega) \left(Y(\omega) - E(Y|G_\theta)(\omega) \right) deg(\omega) dP_\omega \tag{21}$$

En posant pour tout $t \in \Omega$,

$$q(t) = \frac{p_t . deg(t)}{K_\theta} \tag{22}$$

On définit alors une densité de probabilité $Q = \{q_t\}_{t\in\Omega}$ et on obtient la formule suivante :

Propriété 1

$$Cov_\theta(X,Y) = \frac{1}{K_\theta} < X, Y >_{\Delta_\theta} = < X, Y - E(Y|G_\theta) >_Q \tag{23}$$

où $< .,. >_Q$ désigne le produit scalaire pour Q :

$$\begin{aligned} < X,Y >_Q &= \int_\Omega X(\omega)Y(\omega)dq(\omega) \\ &= \int_\Omega X(\omega)Y(\omega)deg(\omega)dP_\omega \end{aligned} \tag{24}$$

La covariance relationnelle entre X et Y est égale au produit scalaire (pour la métrique Q) de X avec l'écart de Y à son espérance relationnelle.
Par linéarité du produit scalaire, (23) implique que dans le cas où la variable X est Q-centrée, la variance de X pour Q se décompose comme suit

$$Var_Q(X) = Var_\theta(X) + < X,E(X|G_\theta) >_Q \tag{25}$$

La variance totale d'une variable s'exprime comme la somme de sa variance relationnelle et de sa variabilité globale (appelée aussi *autocovariance spatiale*). Le terme $< X,E(X|G_\theta) >_Q$ est le produit scalaire entre la variable X et sa moyenne locale dans le graphe et permet de caractériser la dispersion locale de X. On ne peut considérer ce terme comme une variance car il peut arriver que ce terme soit négatif. Thioulouse *et al.* [TCC02] utilisent cette notion dans le cas discret pour analyser la structure locale et globale de dispersions d'oiseaux sur des zones géographiques données. Par cette formule appliquée à des données adaptées, il est possible de mesurer l'indépendance spatiale en utilisant l'indice de Geary [Gea54] et de Moran [Mor48]. L'indice de Moran (ou *autocorrélation spatiale*) s'apparente à un coefficient de corrélation et se définit comme le quotient entre variabilité globale et variance totale. L'indice de Geary, complémentaire, est le rapport entre variance locale et variance totale.

Propriété 2. *En notant par $\Delta_\mathbf{1}$ le Laplacien du graphe complet, on obtient la décomposition suivante :*

$$Var(X) = < X,X >_{\Delta_\mathbf{1}} = < X,X >_{\Delta_\theta} + < X,X >_{\Delta_\mathbf{1}-\Delta_\theta} \tag{26}$$

La variance totale s'exprime en fonction de la covariance relationnelle à laquelle s'ajoute le terme résiduel $< X,X >_{\Delta_\mathbf{1}-\Delta_\theta}$. Celui-ci s'interprète comme une covariance via le graphe complémentaire de G_θ. Une formule analogue en géostatistique donne sous une hypothèse de stationnarité une décomposition de la variance totale en fonction du semi-variogramme et du covariogramme [Cre93, p. 67]. L'opérateur $\Delta_\mathbf{1} - \Delta_\theta$ induit un produit scalaire puisqu'il possède les mêmes propriétés que Δ_θ. On en déduit l'égalité de « type » Pythagore :

$$Var(X) = K_\theta Var_\theta(X) + (1 - K_\theta)Var_{\bar\theta}(X) \tag{27}$$

où $Var_{\bar\theta}(X)$ est la variance de X relative au graphe complémentaire de G_θ :

$$Var_{\bar\theta}(X) = \frac{1}{1 - K_\theta} < X,X >_{\Delta_\mathbf{1}-\Delta_\theta} \tag{28}$$

$E(X|G_\theta)$ est une variable aléatoire définie sur Ω. Elle associe à tout individu la moyenne de la variable X calculée sur ses voisins selon G_θ. Nous nous plaçons bien dans le cadre d'une espérance conditionnelle dont chaque valeur dépend de la réalisation de la variable aléatoire G_θ définie sur $\Omega \times \Omega$. Nous pouvons formuler les trois propriétés suivantes :

Propriété 3. *L'opérateur $E(.|G_\theta)$ est auto-adjoint pour $< .,. >_Q$:*

$$< X,E(Y|G_\theta) >_Q = < E(X|G_\theta),Y >_Q \tag{29}$$

Propriété 4. *La moyenne de l'espérance relationnelle d'une variable est égale à son espérance totale :*

$$E(E(X|G_\theta)) = E(X) \tag{30}$$

Propriété 5. *En notant par $E^{(n)}(X|G_\theta)$ l'espérance relationnelle de X itérée n fois :*

$$E^{(n)}(X|G_\theta) = \underbrace{E(...E(E}_{n \text{ fois}}(X|G_\theta)|G_\theta)...|G_\theta) \tag{31}$$

Alors, dans le cas d'une densité uniforme et pour un graphe dont la fermeture transitive de G_θ est un graphe complet, on peut montrer la relation suivante :

$$\lim_{n\to\infty} E^{(n)}(X|G_\theta) = E(X) \tag{32}$$

De manière générale, $E(.|G_\theta)$ ne peut s'interpréter comme une espérance ordinaire par défaut d'idempotence. Ainsi, $E(.|G_\theta)$ n'est en général pas un projecteur :

$$E(E(X|G_\theta)|G_\theta) \neq E(X|G_\theta) \tag{33}$$

Enfin, $E(.|G_\theta)$ ne vérifie pas la propriété d'orthogonalité d'une espérance conditionnelle ordinaire, et l'on a en général :

$$< E(X|G_\theta),X - E(X|G_\theta) > \neq 0 \tag{34}$$

Par contre, si G_θ est une relation d'équivalence (*i.e.* : partition), chaque individu est connecté pour le graphe à l'ensemble des éléments de sa classe, alors l'espérance relationnelle coïncide exactement avec l'espérance conditionnelle et la covariance (resp. variance) relationnelle s'interprète comme la covariance (resp. variance) intra-classe.

4. Discrétisation

Pour mettre en œuvre les techniques précédentes, il convient d'obtenir un échantillon des variables analysées. On passe ainsi

à une formulation discrète des questions évoquées plus haut dans un cadre continu. L'ensemble Ω est fini (*e.g.* : de taille n), le vecteur aléatoire $\mathbb{X}$ de $\mathbb{R}^p$ est constitué de p variables aléatoires définies sur $\mathcal{L}^2(\{1,...,n\})$ identifiable à $\mathbb{R}^n$. La métrique sur $\mathbb{R}^n$ est représentée par la matrice diagonale D des poids des individus et la métrique sur $\mathbb{R}^p$ reste associée à la matrice M. L'expérience aléatoire conduit à la fois à l'obtention des n réalisations du vecteur aléatoire $\mathbb{X}$ et à celle du graphe de la relation G_θ.

Les opérateurs A_θ et D_θ, qui conduisent à l'expression du Laplacien D_θ ont des formes matricielles symétriques $(n \times n)$ et conservent les propriétés générales précédemment étudiées.

Ces applications sont définies en (13) (14) et s'écrivent matriciellement :

- A_θ est la matrice de connexité de G_θ exprimant la relation binaire induite par la variable G_θ :

$$A_\theta(\omega,\omega') = p_\omega p_{\omega'} G_\theta(\omega,\omega') \quad 1 \leq \omega,\omega' \leq n \tag{35}$$

- D_θ est la matrice diagonale constituée de la somme des éléments de chaque ligne de A_θ :

$$D_\theta(\omega,\omega) = \sum_{\omega'=1}^{n} A_\theta(\omega,\omega') = p_\omega.deg(\omega) \tag{36}$$

où $deg(\omega)$ est la somme discrétisée des poids des individus connectés à ω pour G_θ :

$$deg(\omega) = \sum_{\omega'=1}^{n} p_{\omega'} G_\theta(\omega,\omega') \quad 1 \leq \omega \leq n \tag{37}$$

La covariance relationnelle s'exprime elle aussi matriciellement [Leb69] :

$$V_\theta = \frac{1}{K_\theta}\mathbb{X}^t(D_\theta - A_\theta)\mathbb{X} = \frac{1}{K_\theta}\mathbb{X}^t\Delta_\theta\mathbb{X} \tag{38}$$

Δ_θ est une matrice symétrique semi-définie positive d'ordre n et de rang $n-1$ si le graphe est connexe [Bol98]. En général, si le graphe a k composantes connexes, alors le rang du Laplacien du graphe sera $n-k$.

5. Exemples types

Dans cette section, on supposera que $\Omega = [a,b]$ est un intervalle fermé et borné de $\mathbb{R}$ et que la densité f des variables est uniforme :

$$f \equiv \frac{1}{b-a} \tag{39}$$

On considère le graphe symétrique suivant :

$$G_\theta = \{(\omega,\omega') \in \Omega \times \Omega \ / \ |\omega - \omega'| \leq \theta\} \tag{40}$$

Ainsi, l'opérateur espérance relationnelle $E(.|G_\theta)$ s'interprète comme un filtre par une fonction porte et la covariance relationnelle met en jeu le résidu $X - E(X|G_\theta)$ qui ne conserve que les hautes fréquences.

Le paramètre θ contrôle l'étendue du graphe de connexité : plus θ est grand et plus le graphe connectera d'individus. En se restreignant à ces cas particuliers, nous proposons d'illustrer les notions définies dans cet article par quelques exemples et propriétés simples de la covariance et de l'espérance relationnelles.

5.1. Variance relationnelle : aspects spatial et temporel

Pour une série temporelle donnée, la covariance (ou la variance) relationnelle tient compte de l'ordre des individus sur lesquels la série est mesurée. La figure 1 présente en (a) une série ainsi qu'une permutation en (b). Les écarts-types globaux des deux séries sont tous deux égals à 89.59. Avec une valeur du paramètre $\theta = 5$, l'écart-type relationnel de la série permutée est très supérieure à celle de la série initiale (82.75 au lieu 18.14). Comme on le verra dans la section qui suit, les parties linéaires de la série n'influencent pas la variance relationnelle mais le comportement erratique de la série permutée implique une variance relationnelle importante.

Pour le deuxième exemple, on considère deux variables telles que la disposition des individus sur le plan soit en « dents de scie » (fig. 2). On pratique deux ACP sur ces variables : une relationnelle ($\theta = 5$) et une standard. Le premier vecteur propre de chaque ACP est montré figure 2. On remarque que la direction

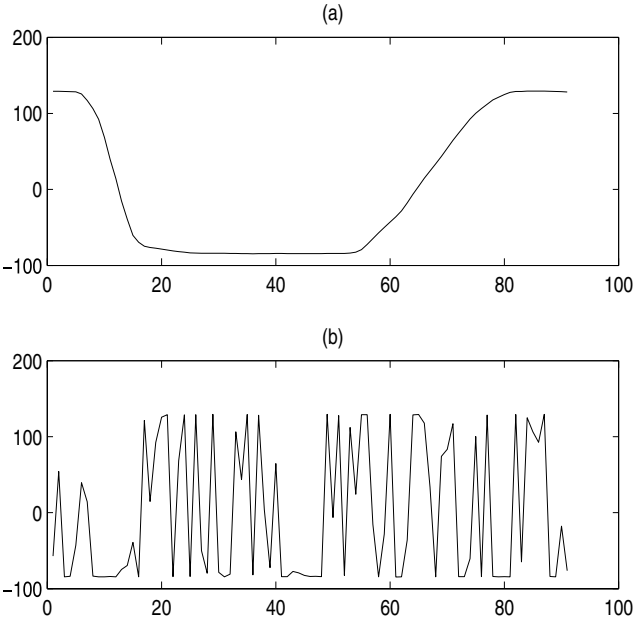


Figure 1. (a) Série temporelle X, $\sigma(X) = 89.59$ et $\sigma_\theta(X) = 18.14$; (b) Série $\widetilde{X}$: permutation aléatoire de X, $\sigma(\widetilde{X}) = 89.59$ et $\sigma_\theta(\widetilde{X}) = 82.75$ ($\theta = 5$).

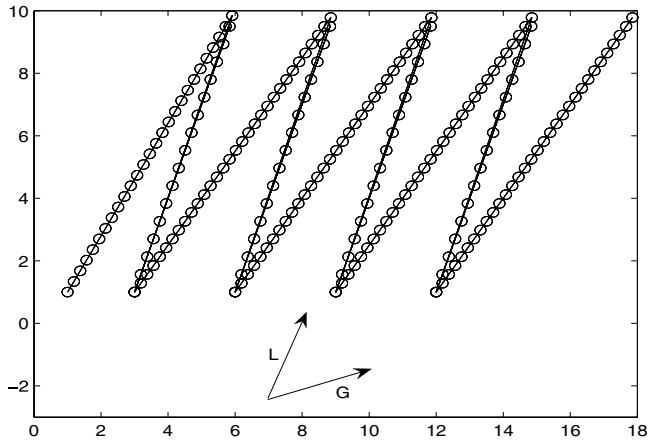


Figure 2. Disposition des individus (petits cercles) avec deux vecteurs propres : l'un issu d'une ACP standard (G) et l'autre d'une ACPR (L) avec $\theta = 5$.

locale de la trajectoire analysée est bien mis en valeur par une ACP relationnelle contrairement à l'ACP standard où le vecteur (G) caractérise la dispersion globale du nuage.

5.2. Espérance relationnelle : exemples

Dans cette section nous proposons deux exemples simples illustrant la notion d'espérance relationnelle. La figure 3 montre une série temporelle X avec son espérance relationnelle tracée en pointillé ($\Omega = [1,100]$ et $\theta = 10$). On constate que la variable coïncide avec son espérance relationnelle sur les intervalles [0,20], [40,60] et [80,100], c'est-à-dire sur les segments où la variable varie linéairement. Il est donc clair que l'espérance relationnelle généralise la notion de moyenne mobile.
La figure 4 présente les deux premières CP C_1 et C_2 d'une ACPR ($\theta = 5$) issues d'un ensemble de variables correspondant

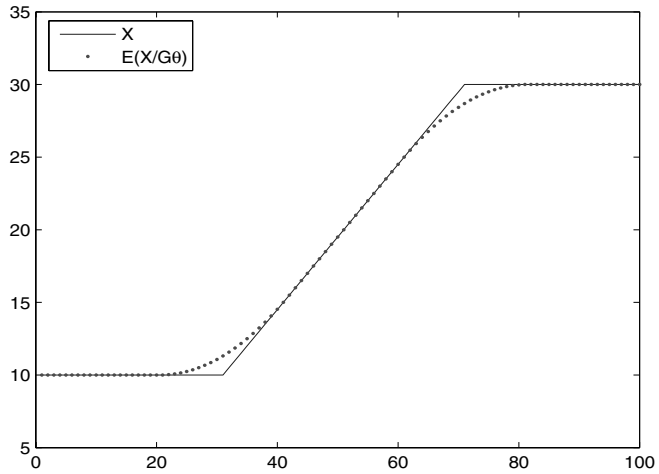


Figure 3. Graphe d'une série temporelle X avec son espérance relationnelle $E(X|G_\theta)$ ($\theta = 10$).

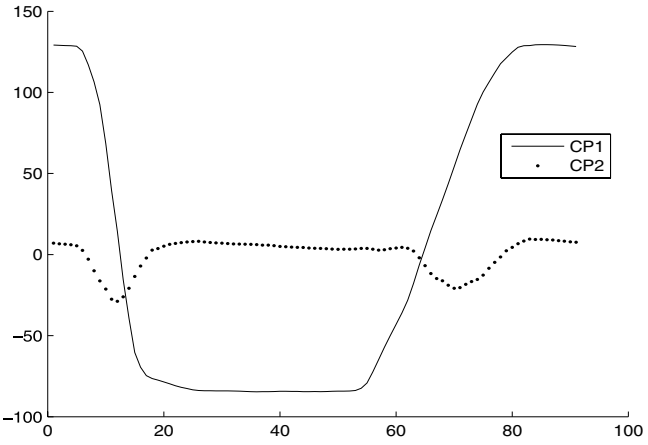


Figure 4. Graphe des deux premières CP d'une ACPR ($\theta = 5$) sur des variables obtenues en enregistrant les coordonnées des membres d'un danseur (mouvement Chute Pliée).

aux coordonnées tridimensionnelles de 15 capteurs repérant un danseur qui exécute un mouvement chorégraphique (« Chute Pliée ») [CBV02, HLSJ06]. Ces deux variables sont orthogonales (décorrélées) pour la métrique du Laplacien ($Corr_\theta(C_1,C_2) = 0$). En revanche, la corrélation standard entre les deux variables est non nulle ($Corr(C_1,C_2) = -0.19$), valeur significative pour le nombre d'observations considéré. Cela s'explique par le fait que contrairement à la covariance standard, la covariance relationnelle ne tient pas compte des observations ayant eu lieu entre les dates 20 et 55. Cette période de temps n'a en effet aucun intérêt pour l'analyse locale puisque le danseur est immobile du point de vue de l'une des CP analysées. Dans cette posture, le danseur exprime une attitude mais la covariance locale ne prend en compte que les liens éventuels entre les déplacements des capteurs.
Comme dans l'exemple de la covariance relationnelle (fig. 1), l'espérance relationnelle est calculée suivant un graphe de connexité qui est modifié en fonction de la permutation de la série. On peut vérifier que la différence entre la variable et son espérance relationnelle est nulle dès que la variable varie linéairement en fonction du temps. Donc, les observations pour lesquelles la variable coïncide avec son espérance relationnelle n'influent pas sur la variance relationnelle puisque celle-ci est nulle d'après (23).

6. Conclusion

Notre motivation première dans cet article a été énoncée comme une volonté d'expliciter comment prendre en compte une information exogène de type relationnelle dans l'étude de la liaison entre variables aléatoires. Le contexte de notre travail est celui de l'analyse des covariances ou corrélations par ACP. Nous avons noté que l'ACP classique ignore toute structure extra-

probabiliste sur l'espace Ω. Il est certes possible d'introduire l'information relationnelle *a posteriori* en représentant par exemple le graphe de la relation exogène sur les plans principaux de l'ACP classique. Mais notre intérêt concerne la prise en compte *a priori* de cette information, dans le calcul même des éléments propres. L'ACP relationnelle n'est pas une approche nouvelle, et nous avons donné un ensemble de travaux antérieurs qui présentent cette méthode comme une réponse à notre problématique.

Notre contribution exploite les différentes expressions formelles du critère de la covariance relationnelle pour en décrire ses propriétés et son interprétation. Il apparaît que la spécificité de l'analyse relationnelle réside effectivement dans les propriétés de la covariance relationnelle plutôt que dans celles de l'ACP. Plus précisément, une fois que l'on s'est convaincu que la covariance relationnelle était une covariance à part entière, l'ACP s'applique clairement de façon classique avec toutes ses propriétés usuelles. Par contre, la nature exacte de l'information capturée par la covariance et la variance relationnelles était une chose plus obscure et finalement peu étudiée.

L'information relationnelle exogène s'exprime au travers d'un graphe (orienté) déduit d'un paramètre aléatoire observé sur l'espace fondamental support de l'analyse. L'ACPR ne stipule aucune propriété particulière sur ce graphe. Nous montrons qu'en fait, l'ACPR analyse non pas le graphe étudié, mais plutôt une classe de graphes qui lui sont équivalents dans un sens que nous précisons. La covariance relationnelle sous-entend la notion d'espérance relationnelle, opérateur qui sans avoir la totalité de ses propriétés, présente de fortes similitudes avec la notion classique d'espérance conditionnelle.

Au-delà des propriétés générales de la covariance relationnelle, valables quelle que soit la relation étudiée, des propriétés spécifiques peuvent être proposées lorsque la relation exogène étudiée est d'une forme particulière. Ainsi, nous étudions le cas de relations d'ordre et surtout le cas d'une relation d'équivalence, pour laquelle espérance relationnelle et espérance conditionnelle coïncident.

Nous appliquons cette technique dans des domaines divers d'analyse du mouvement dansé (trajectoires de groupes de capteurs équipant les danseurs), de données en biologie marine (trajectoires de poissons) et enfin en chimie (analyse de lits de particules). L'ACP Relationnelle généralise donc l'ACP standard et son caractère « local » s'étend à d'autres méthodes factorielles telles que l'analyse discriminante, l'analyse canonique ou encore l'analyse d'un tableau de distances.

Références

[BE90] H. BENALI and B. ESCOFIER, Analyse factorielle lissée et analyse factorielle des différences locales. *Revue de Statistique Appliquée*, 38(2) :55-76, 1990.

[BL84] T. ALUJA BANET and L. LEBART, Local and partial principal component analysis and correspondence analysis. *Proceedings in Computational Statistics (COMPSTAT'84)*, pp. 113-118, 1984.

[Bol98] B. BOLLODAS, *Graph Theory*. Springer Verlag, 1998.

[BWOS96] M.A. BALSERA, W. WRIGGERS, Y. OONO and K. SCHULTEN, Principal component analysis and long time protein dynamics. *Journal of Physical and Chemical*, 100 :2567-2772, 1996.

[CBV02] F. CHENEVIERE, S. BOUKIR and B. VACHON, A HMM-based dance gesture recognition. *IWSIP'02*, pp. 322-326, 2002.

[Cre93] N. CRESSIE, *Statistics for Spatial Data*. New York, Wiley, 1993.

[Fol82] Y. LE FOLL, Pondération des distances en analyse factorielle. *Statistique et Analyse des données*, 1(7) :13-21, 1982.

[Gea54] R.C. GEARY, The contiguity ratio and statistical mapping. *The Incorporated Statistician*, 5 :115-145, 1954.

[HLSJ06] S. HIDOT, J.Y. LAFAYE and C. SAINT-JEAN, ACP relationnelle pour l'analyse du mouvement : application à la danse. *Actes électroniques de RFIA'06*, p. 64, 9 pages, 2006.

[JN02] D. JENSEN and J. NEVILLE, Linkage and autocorrelation cause feature selection bias in relationnal learning. *Nineteenth International Conference of Machine Learning ICML'02*, pages 259-266, 2002.

[Jol86] I.T. JOLIFFE, *Principal Component Analysis*. Springer Verlag, 1986.

[KMMT01] S. KSHIRAGAR, T. MOLLET and N. MAGNENAT-THALMANN, Principal components of expressive speech animation. *In Proceedings of CG'01*, pages 38-46, 2001.

[Kri51] D.G. KRIGE, A Statistical approach to some basic mine valuation problems on the witwatersrand. *Metallurgical and Mining Society of South Africa*, 52 :119-139, 1951.

[Leb69] L. LEBART, Analyse statistique de la contiguïté. *Institut de Statistique de l'Université de Paris*, 28 :81-112, 1969.

[LMP00] L. LEBART, A MORINEAU and M. PIRON, *Statistique exploratoire multidimensionnelle*. Dunod, 2000.

[Mat65] G. MATHERON. *Les variables régionalisées et leur estimation*. Masson, Paris, 1965.

[Moh91] B. MOHAR, The Laplacian spectrum of graphs. *In Graph Theory, Combinatorics and Application*, 2 :871-898, 1991.

[Mor48] P.A.P. MORAN, The interpretation of statistical maps. *Journal of the Royal Statistical Society, series B*, 10 :243-251, 1948.

[PL97] N. PECH and F. LAHOE, Use of principal component analysis with instrumental variables (PCAIV) to analyse fisheries catch data. *ICES Journal of Marine Science*, pages 32-47, 1997.

[RSV90] M. ROUX and S. SERVANT-VILDARY, Multivariate analysis of diatomes and water chemistry in bolivian salin lakes hydrobiologia. *Hydrobiologia*, 197 :267-290, 1990.

[Sab87] R. SABATIER, Analyse factorielle de données structurées et métriques. *Statistique et Analyse des Données*, 3 :75-96, 1987.

[SSM96] B. SCHOLKOPF, A. SMOLA and K.R. MULLER, Nonlinear componant analysis as a Kernel eignevalue problem. Technical report, Max-Planck-Institute fur biologische Kybernetik, 1996.

[TCC02] J. THIOULOUSE, D. CHESSEL and S. CHAMPELY, Multivariate analysis of spatial patterns : a unified approach to local and global structures. *Environnemental and Ecological Statistics*, 27 :1-14, 2002.

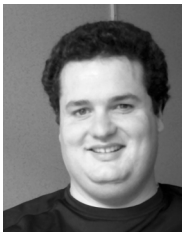
[WHT02] L. WANG, W. HU and T. TAN, A new attempt to gait-based human identification. *In Proceedings of ICPR'02*, 1 :115-118, 2002.

Sullivan **Hidot**

Sullivan Hidot est doctorant au Laboratoire L3i de l'Université de La Rochelle depuis 2004. Ses travaux portent sur l'analyse de données spatio-temporelles et en particulier l'analyse de trajectoires. Il s'intéresse aux méthodes factorielles, statistiques et à l'approche fractale pour l'étude du mouvement. Son domaine d'application est l'analyse du mouvement dansé et du déplacement de poissons.

Jean-Yves **Lafaye**

Prof. J.Y. Lafaye est responsable du projet L3i « Compréhension des comportements complexes » à l'Université de La Rochelle. Sa principale contribution concerne la modélisation de systèmes avec des applications en écologie, biologie marine et plus récemment en analyse du mouvement. Après un doctorat en statistique et modèles stochastiques, il est maintenant impliqué dans l'enseignement et la recherche universitaires dans le domaine du génie logiciel et plus spécialement en spécifications formelles, qualité et test.



Christophe **Saint-Jean**

Christophe Saint-Jean est Maître de conférences au L3i à l'Université de La Rochelle depuis 2002. Après une thèse sur la classification robuste partiellement supervisée, il s'intéresse maintenant à l'analyse de données spatio-temporelles pour la segmentation, caractérisation, reconnaissance du mouvement. Il investigue l'utilisation de méthodes de l'analyse de données pour l'analyse et la compréhension des images.