Analyse en composantes principales partielle de données séquentielles d'espérance et de matrice de covariance variables dans le temps

Romain Bar, Jean-Marie Monnez

ANALYSE EN COMPOSANTES PRINCIPALES PARTIELLE DE DONNÉES SÉQUENTIELLES D'ESPÉRANCE ET DE MATRICE DE COVARIANCE VARIABLES DANS LE TEMPS.

Romain Bar [1] & Jean-Marie Monnez [2]

[1,2] *Université de Lorraine, CNRS, Institut Elie Cartan de Lorraine (IECL), UMR 7502, BP 239, Vandoeuvre-lès-Nancy, F-54506, France*
INRIA, projet BIGS, Villers-lès-Nancy, F-54600, France
[1] *Romain.Bar@univ-lorraine.fr* [2] *Jean-Marie.Monnez@univ-lorraine.fr*
http ://www.iecl.u-nancy.fr

Résumé. On suppose que des vecteurs de données pouvant être de grande dimension et arrivant séquentiellement dans le temps sont des observations indépendantes d'un vecteur aléatoire d'espérance mathématique et de matrice de covariance variables dans le temps.

On définit alors une méthode récursive d'estimation en ligne de vecteurs directeurs des r premiers axes principaux d'une analyse en composantes principales (ACP) partielle de ce vecteur aléatoire.

On applique ensuite ce résultat au cas particulier de l'analyse canonique généralisée (ACG) partielle après avoir défini un processus d'approximation stochastique de type Robbins-Monro de l'inverse d'une matrice de covariance.

Mots-clés. Big Data, flux de données, données de grande dimension, analyse de données en ligne, analyse en composantes principales partielle, analyse canonique généralisée partielle, approximation stochastique.

Abstract. High dimensional batch data are supposed to be independent observations of a random vector Z, expectation and covariance matrix of which vary with time n.

A recursive method of on-line estimation of direction vectors of the r first principal axes of a partial principal components analysis (PCA) of Z is defined.

This is applied next to the particular case of a partial generalized canonical correlation analysis (gCCA) after defining a stochastic approximation process of the Robbins-Monro type to estimate recursively the inverse of a covariance matrix.

Keywords. Big Data, data flow, high dimensional data, on-line data analysis, partial principal component analysis, partial generalized canonical correlation analysis, stochastic approximation.

1 Introduction

On observe p caractères quantitatifs sur des individus : on obtient des vecteurs de données z_i dans $\mathbb{R}^p$. On se place ici dans le cas où ces vecteurs arrivent séquentiellement dans le temps (données en ligne) : on observe z_n au temps n ; on a une suite de vecteurs de données $z_1, \ldots, z_n, \ldots$.

On suppose que, pour tout $n \geq 1$, z_n est la réalisation d'un vecteur aléatoire (v.a.) Z_n dans $\mathbb{R}^p$, défini sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, d'espérance mathématique et de matrice de covariance variables dans le temps, les v.a. Z_n étant mutuellement indépendants.

Pour tout n, on a la décomposition :

$$Z_n = \theta_n + \Delta_n R_n \quad \text{où :}$$

- $\theta_n = (\theta_n^1 \ldots \theta_n^p)'$ est un vecteur de $\mathbb{R}^p$,

- $\Delta_n = \begin{pmatrix} \delta_n^1 & & & \\ & \cdot & 0 & \\ & & \cdot & \\ & 0 & & \cdot \\ & & & \delta_n^p \end{pmatrix}$ est une matrice diagonale d'ordre p d'éléments non nuls.

  On pose : $\delta_n = (\delta_n^1, \ldots, \delta_n^p)$

- La loi du v.a. R_n ne dépend pas de n, $\mathbb{E}[R_n] = 0$, $Cov(R_n) = C$.

Ceci revient à supposer que les $R_n = \Delta_n^{-1}(Z_n - \theta_n)$ constituent un échantillon i.i.d. d'un vecteur aléatoire R dans $\mathbb{R}^p$ tel que $\mathbb{E}[R] = 0, Cov(R) = C$. On a alors $\mathbb{E}[Z_n] = \theta_n, Cov(Z_n) = \Delta_n C \Delta_n$.

Pour $i = 1, \ldots, p$, si l'on note Z_n^i, respectivement R_n^i, la i$^{\text{ème}}$ composante de Z_n, respectivement R_n, on a alors le modèle :

$$Z_n^i = \theta_n^i + \delta_n^i R_n^i, \;\; i = 1, \ldots, p$$

avec $\mathbb{E}[R_n^i] = \mathbb{E}[R^i] = 0$, les R_n^i constituant un échantillon i.i.d de R^i.

Dans la suite, on suppose sans perte de généralité que $Var(R_n^i) = Var(R^i) = 1$.

On pose le problème suivant : estimer les résultats d'une ACP du v.a. R (cf paragraphe 2), aussi appelée ACP partielle, dans $\mathbb{R}^p$ que l'on munit d'une métrique M. On étudie en particulier l'estimation de vecteurs directeurs des r premiers axes principaux de cette analyse qu'on note v_l, $l = 1, 2, \ldots, r$.

Soit v un résultat de l'ACP d'un v.a., par exemple une valeur propre, un vecteur directeur d'un axe principal,....

Plutôt que d'effectuer à chaque temps n une estimation directe de v à partir de l'ensemble des données disponibles à ce temps, qui serait obtenue en effectuant une ACP partielle de ces données, on va effectuer une estimation récursive de v : disposant d'une estimation v_n de v obtenue à partir des observations $z_1, \ldots, z_{n-1}$, on introduit au temps n l'observation z_n et on définit à partir de v_n et z_n une nouvelle estimation v_{n+1} de v : $v_{n+1} = f_n(v_n; z_n)$.

On utilise pour cela un processus d'approximation stochastique. On pourra consulter à ce sujet les articles de Robbins et Monro (1951), Benzécri (1969), Bouamaine et Monnez (1998).

L'intérêt de cette approche récursive est double :

- on n'a pas besoin de stocker les données jusqu'au temps n ;

- la relative simplicité des calculs permet de prendre en compte, dans le même temps de calcul, plus de données que par une méthode classique.

2 ACP d'un vecteur aléatoire

Soit C la matrice de covariance de R et M une métrique quelconque dans $\mathbb{R}^p$.
On suppose qu'il n'existe pas de relation affine entre les composantes du vecteur aléatoire Z.

Le critère de l'ACP du vecteur R est le suivant : pour $l = 1, \ldots, r$, déterminer au pas l une combinaison linéaire des composantes centrées de R, $U_l = \eta_l'(R - \mathbb{E}[R])$, de variance maximale sous les contraintes d'être non corrélée aux précédentes et que η_l soit M^{-1}-unitaire. η_l, appelé $l^{\text{ième}}$ facteur, est vecteur propre associé à la $l^{\text{ième}}$ plus grande valeur propre λ_l de la matrice M^{-1}-symétrique MC et on a $\eta_l' C \eta_l = \lambda_l$.

$v_l = M^{-1}\eta_l$ est un vecteur directeur du $l^{\text{ième}}$ axe principal de cette ACP, vecteur propre de CM.

3 Approximation stochastique des vecteurs v_l

On est ainsi amené à chercher les r premiers vecteurs propres $v_1,\dots,v_r$ de la matrice M-symétrique $B = CM$ associés aux r plus grandes valeurs propres $\lambda_1,\dots,\lambda_r$.

On a $C = Cov(R) = Cov(R_n) = Cov\Big(\Delta_n^{-1}(Z_n - \theta_n)\Big) = \Delta_n^{-1}\mathbb{E}\Big[(Z_n - \theta_n)Z_n'\Big]\Delta_n^{-1}$.

On suppose que l'on a défini trois estimateurs M_n, Θ_n et D_n respectivement de M, θ_n et Δ_n, tels que $M_n \to M$ p.s., $\Theta_n - \theta_n \to 0$ p.s. et $D_n - \Delta_n \to 0$ p.s.

Soit

$$C_n = D_n^{-1}(Z_n - \Theta_n)Z_n'D_n^{-1}.$$

On définit alors $B_n = C_nM_n$ puis récursivement un processus d'approximation stochastique $(X_n) = \Big((X_n^1,\dots,X_n^r)\Big)$ de $(v_1,\dots,v_r)$ par :

$$\begin{array}{rcl} F_n(X_n^l) & = & \frac{\langle B_nX_n^l, X_n^l\rangle_{M_n}}{||X_n^l||^2_{M_n}}, \\ Y_{n+1}^l & = & X_n^l + \frac{a}{n^\alpha}(B_n - F_n(X_n^l)I)X_n^l,\ l = 1,\dots,r, \\ X_{n+1} & = & \mathrm{orth}_{M_n}(Y_{n+1}). \end{array}$$

Pour obtenir X_{n+1}, on effectue une orthogonalisation au sens de Gram-Schmidt par rapport à M_n de $Y_{n+1} = (Y_{n+1}^1,\dots,Y_{n+1}^r)$. On établit la convergence de ce processus pour $\frac{2}{3} < \alpha \leq 1$.

Un cas particulier de modèle d'évolution dans le temps de l'espérance θ_n et de Δ_n est le suivant. Si l'on note θ_n^i la $i^{\text{ème}}$ composante réelle de θ_n $(i = 1,\dots,p)$ et δ_n^i la $i^{\text{ème}}$ composante réelle de δ_n, on définit les modèles linéaires $\theta_n^i = \langle \beta^i, U_n^i\rangle$ et $(\delta_n^i)^2 = \langle \gamma^i, V_n^i\rangle$, $\langle,\rangle$ désignant le produit scalaire euclidien usuel dans $\mathbb{R}^{n_i}$ (resp. $\mathbb{R}^{m_i}$), U_n^i (resp. V_n^i) étant un vecteur de dimension n_i (resp. m_i) de valeurs de fonctions connues du temps n ou de variables explicatives contrôlées et β^i (resp. γ^i) un vecteur inconnu de $\mathbb{R}^{n_i}$ (resp. $\mathbb{R}^{m_i}$).

En s'inspirant de Monnez (2008b), on définit le processus d'approximation stochastique (B_n^i) de β^i tel que :

$$B_{n+1}^i = \Pi_{K_n^i}\Big(B_n^i - a_nU_n^i\Big((U_n^i)'B_n^i - Z_n^i\Big)\Big),$$

K_n^i étant l'ensemble convexe $\{x \in \mathbb{R}^{n_i},\ ||x|| < k_n^i\}$ où $k_n^i = O(n^\beta)$, $\beta > 0$.

On définit comme estimateur de θ_n^i, $\Theta_n^i = \langle B_n^i, U_n^i\rangle$, puis comme estimateur de θ_n, $\Theta_n = (\Theta_n^1,\dots,\Theta_n^p)'$.

On définit également le processus d'approximation stochastique (C_n^i) de γ_i tel que :

$$C_{n+1}^i = \Pi_{L_n^i}\left(C_n^i - a_n V_n^i\left((V_n^i)' C_n^i - (Z_n^i - \Theta_n^i)^2\right)\right),$$

L_n^i étant l'ensemble convexe $\{x \in \mathbb{R}^{m_i},\ ||x|| < l_n^i\}$ où $l_n^i = O(n^\gamma)$, $\gamma > 0$.

On définit enfin comme estimateur de δ_n^i, $D_n^i = (\langle C_n^i, V_n^i \rangle)^{\frac{1}{2}}$, et comme estimateur de Δ_n,

$$D_n = \begin{pmatrix} D_n^1 & & & \\ & \cdot & 0 & \\ & & \cdot & \\ & 0 & & \cdot \\ & & & D_n^p \end{pmatrix}.$$

On établit la convergence de ces processus.

4 Cas particulier : l'Analyse Canonique Généralisée (ACG)

Comme dans Monnez (2008a), on se place dans le cas où le vecteur aléatoire R est partitionné en sous-vecteurs $R^{(1)}, \ldots, R^{(q)}$; pour $k = 1, \ldots, q$, $R^{(k)}$ est un vecteur aléatoire dans $\mathbb{R}^{m_k}$, de composantes $R^{k1}, \ldots, R^{km_k}$. On note également $R^{(k)} = \{R^i, i \in J_k\}$, $card\ (J_k) = m_k$, $\sum_{k=1}^{q} m_k = p$.

On souhaite effectuer une ACP de R dans laquelle les vecteurs aléatoires $R^{(k)}$ aient un rôle équilibré : on veut éviter que les premiers facteurs soient principalement déterminés à partir de certains vecteurs $R^{(k)}$. L'analyse canonique généralisée du vecteur aléatoire R fournit une solution à ce problème grâce à un choix particulier de métrique.

Le vecteur aléatoire Z_n observé est partitionné en sous-vecteurs $Z_n^{(1)}, \ldots, Z_n^{(q)}$, de dimensions respectives $m_1, \ldots, m_q$, avec :

$$Z_n^{(k)} = \theta_n^{(k)} + \Delta_n^{(k)} R_n^{(k)},$$

$\Delta_n^{(k)}$ étant la matrice diagonale d'ordre m_k des δ_n^i, $i \in J_k$, $\theta_n^{(k)}$ étant le vecteur des θ_n^i, $i \in J_k$, définis dans le paragraphe 3 et les $R_n^{(k)}$ constituant un échantillon i.i.d. de $R^{(k)}$.

On définit $C^k = \mathbb{E}\left[R^{(k)}(R^{(k)})'\right]$, matrice de covariance de $R^{(k)}$. L'ACG de R est une ACP avec la métrique diagonale par blocs d'ordre p, M :

$$M = \begin{pmatrix} (C^1)^{-1} & & & \\ & \cdot & 0 & \\ & & \cdot & \\ & 0 & & \cdot \\ & & & (C^q)^{-1} \end{pmatrix}$$

Pour tout n, pour $k = 1, \ldots, q$, $(C^k)^{-1}$ est solution de l'équation en X :

$$\mathbb{E}\left[R_n^{(k)}(R_n^{(k)})'X - I\right] = \mathbb{E}\left[(\Delta_n^{(k)})^{-1}(Z_n^{(k)} - \theta_n^{(k)})(Z_n^{(k)})'(\Delta_n^{(k)})^{-1}X - I\right] = 0$$

où I est la matrice-identité d'ordre m_k.

Soit $\Theta_n^{(k)}$ le vecteur des $\Theta_n^i,\ i \in J_k$, définis dans le paragraphe 3.
Soit $D_n^{(k)}$ la matrice diagonale des $D_n^i,\ i \in J_k$, définis dans le paragraphe 3.

On définit récursivement le processus d'approximation stochastique de $(C^k)^{-1}$, (M_n^k), par :

$$M_{n+1}^k = M_n^k - \frac{a}{n^\alpha}\left((D_n^{(k)})^{-1}(Z_n^{(k)} - \Theta_n^{(k)})(Z_n^{(k)})'(D_n^{(k)})^{-1}M_n^k - I\right).$$

On établit que $M_n^k - (C^k)^{-1} \to 0$ p.s. et $\sum_{n=1}^{\infty} \frac{1}{n^\alpha}||M_n^k - (C^k)^{-1}|| < \infty$ p.s. pour $\frac{2}{3} < \alpha \leq 1$.

On définit alors comme estimateur de M au pas n la matrice diagonale par blocs M_n qui a pour $k^{\text{ième}}$ bloc diagonal M_n^k.

On déduit du résultat précédent sur les $M_n^k,\ k = 1, \ldots, q$, que $M_n - M \to 0$ p.s. et $\sum_{n=1}^{\infty} \frac{1}{n^\alpha}||M_n - M|| < \infty$ p.s. pour $\frac{2}{3} < \alpha \leq 1$.

Bibliographie

[1] Benzecri, J.P. (1969), Approximation stochastique dans une algèbre normée non commutative, Bulletin de la SMF, 97, 225-241.
[2] Bouamaine, A. et Monnez, J.M. (1998), Approximation stochastique de vecteurs et valeurs propres, Publications de l'ISUP, 42, n°2-3, 15-38.
[3] Monnez, J.M. (2008a), Stochastic approximation of the factors of a generalized canonical correlation analysis, Statistics & Probability Letters, 78, 2210-2216.
[4] Monnez, J.M. (2008b), Analyse en composantes principales d'un flux de données d'espérance variable dans le temps, RNTI, C-2, 43-56.
[5] Robbins, H. et Monro, S. (1951), A stochastic approximation method, AMS, 22, 400-407.