Analyse canonique généralisée de données séquentielles d'espérance fixe ou variable dans le temps

Romain Bar, Jean-Marie Monnez

ANALYSE CANONIQUE GÉNÉRALISÉE DE DONNÉES SÉQUENTIELLES D'ESPÉRANCE FIXE OU VARIABLE DANS LE TEMPS.

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

[1,2] *Institut Elie Cartan, UMR 7502, Université de Lorraine, CNRS, INRIA*
BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, 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 i.i.d. d'un vecteur aléatoire. Après avoir défini un processus d'approximation stochastique de type Robbins-Monro de l'inverse d'une matrice de covariance, on définit une méthode récursive d'estimation séquentielle de vecteurs directeurs des r premiers axes principaux de l'analyse canonique généralisée de ce vecteur aléatoire. On étudie ensuite le cas où l'espérance des observations varie dans le temps. On donne finalement des résultats de simulation.

Mots-clés. Données de grande dimension, analyse de données séquentielles, analyse canonique généralisée, approximation stochastique.

Abstract. High dimensional data of a generalized canonical correlation analysis (gCCA) are supposed first to be i.i.d. observations of a random vector Z which are taken sequentially. After defining a stochastic approximation process of the Robbins-Monro type to estimate sequentially the inverse of a covariance matrix, a recursive method of sequential estimation of direction vectors of the principal axes of gCCA is defined. Next, the case where the expectation of the n^{th} observation varies with time n is studied. Finally, simulation results are given.

Keywords. High dimensional data, sequential data analysis, 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 : on observe z_n au temps n ; on a une suite de vecteurs de données $z_1, \ldots, z_n, \ldots$.On suppose d'abord que cette suite constitue un échantillon i.i.d. d'un vecteur aléatoire Z dans $\mathbb{R}^p$ défini sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$. Ω représente une population d'où on extrait un échantillon.

On se place dans le cas où le vecteur aléatoire Z est partitionné en sous-vecteurs $Z^1, \ldots, Z^q$; pour $k = 1, \ldots, q$, Z^k est un vecteur aléatoire dans $\mathbb{R}^{m_k}$, de composantes $Z^{k1}, \ldots, Z^{km_k}$. On souhaite effectuer une ACP de Z dans laquelle les vecteurs aléatoires Z^k aient un rôle équilibré : on

veut éviter que les premiers facteurs soient principalement déterminés à partir de certains vecteurs Z^k. L'analyse canonique généralisée du vecteur aléatoire Z (ACGVA) fournit une solution à ce problème.

L'ACGVA représente l'ACG effectuée sur la population Ω dont on va chercher à estimer au temps n les résultats à partir des données dont on dispose à ce temps. Soit θ un résultat de l'ACGVA, par exemple une valeur propre, un vecteur directeur d'un axe principal,.... Plutôt que d'effectuer à chaque temps n une estimation de θ à partir de l'ensemble des données dont on dispose jusqu'à ce temps, on va effectuer une estimation récursive de θ : disposant d'une estimation θ_n de θ 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 θ_n et z_n une nouvelle estimation θ_{n+1} de θ : $\theta_{n+1} = f_n(\theta_n; z_n)$. On utilise pour cela un processus d'approximation stochastique. On pourra consulter à ce sujet les articles de Robbins et Monro (1951), Benzécri (1969) ou encore Bouamaine et Monnez (1998).

Remarquons que les résultats de cette dernière étude peuvent être aussi appliqués au cas où l'on introduit au temps n, au lieu d'une seule observation z_n de Z, un bloc de r_n observations, $\{z_i, i \in I_n\}$, $\text{card}(I_n) = r_n$, ou à celui où l'on utilise toutes les observations z_i faites jusqu'à ce temps, $\{z_i, i \in \bigcup_{j=1}^{n} I_j\}$.

On suppose ensuite que l'espérance de Z_n, θ_n, varie dans le temps. $Z_n = \theta_n + R_n$, les R_n constituant un échantillon i.i.d. d'un vecteur aléatoire R. Disposant au temps n d'une estimation de θ_n, pouvant être obtenue par approximation stochastique, on effectue dans les mêmes conditions l'ACG de R ou ACG partielle.

2 ACG d'un vecteur aléatoire

Dans tout le paragraphe, on adopte la présentation de Monnez (2008a).

On suppose qu'il n'existe pas de relation affine entre les composantes du vecteur aléatoire Z. Le critère de l'ACG est le suivant : pour $l = 1,\ldots,r$, déterminer au pas l une combinaison linéaire des composantes centrées de Z, $U_l = \theta_l'(Z - \mathbb{E}[Z])$, et pour $k = 1,\ldots,q$, une combinaison linéaire des composantes centrées de Z^k, $V_l^k = (\eta_l^k)'(Z^k - \mathbb{E}[Z^k])$, telles que :

$$\begin{aligned}
&\sum_{k=1}^{q} \rho^2(U_l, V_l^k) \ \max\,, \\
&Var(U_l) = 1, \\
&Cov(U_l, U_j) = 0,\ j = 1,\ldots,l-1, \\
&Var(V_l^k) = 1,\ k = 1,\ldots,q.
\end{aligned}$$

Soit C la matrice de covariance de Z, C^k celle de Z^k et M la métrique diagonale par blocs d'ordre p :

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

θ_l, appelé $l^{\text{ième}}$ facteur général, est vecteur propre de la matrice MC associé à la $l^{\text{ième}}$ plus grande valeur propre. On peut interpréter ce résultat de la façon suivante : θ_l est le $l^{\text{ième}}$ facteur de l'ACP de Z dans $\mathbb{R}^p$ muni de la métrique M. $v_l = M^{-1}\theta_l$ est un vecteur directeur du $l^{\text{ième}}$ axe principal de cette ACP, vecteur propre de CM. Dans le cas particulier où, pour tout k, Z^k est de dimension 1, on retrouve l'ACP normée.

3 Approximation stochastique des vecteurs v_l

On suppose qu'au temps n, on dispose d'un bloc de r_n nouvelles observations i.i.d de Z, $Z_{R_{n-1}+1}, \ldots, Z_{R_n}$, avec $R_n = \sum_{j=1}^{n} r_j$. On note $I_n = \{R_{n-1}+1, \ldots, R_n\}$.

Pour définir le processus d'approximation stochastique, on utilise au temps n un estimateur convergent M_n de M, obtenu à partir des observations $Z_1, \ldots, Z_{R_{n-1}}$.

Soit le vecteur aléatoire Z_1^k de dimension $m_k + 1$, obtenu en ajoutant au vecteur Z^k une dernière composante égale à 1. Soit J la matrice (m_k+1, m_k) obtenue en ajoutant à la matrice-identité d'ordre m_k une dernière ligne de zéros. On établit que la matrice (m_k+1, m_k),

$$X^k = \begin{pmatrix} (C^k)^{-1} \\ -(\mathbb{E}[Z^k])'(C^k)^{-1} \end{pmatrix},$$

est solution de l'équation en X : $\mathbb{E}[Z_1^k (Z_1^k)'X - J] = 0$.

On définit alors récursivement le processus (M_{1n}^k) d'approximation stochastique de X^k, de type Robbins-Monro, dans l'ensemble des matrices (m_k+1, m_k) :

$$M_{1,n+1}^k = M_{1n}^k - a_n((\frac{1}{r_n} \sum_{l=R_{n-1}+1}^{R_n} Z_{1l}^k (Z_{1l}^k)')M_{1n}^k - J),$$

$$a_n > 0, \sum_1^\infty a_n = \infty, \sum_1^\infty (a_n)^2 < \infty.$$

Soit M_n^k la matrice obtenue à partir de M_{1n}^k en enlevant la dernière ligne ; on définit 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.

Dans le cas où $r_n = 1$ pour tout n, on peut aussi utiliser comme estimateur de $(C^k)^{-1}$, à l'instar de Nguyen et Saracco (2010), l'inverse de la matrice de covariance empirique obtenue à partir des observations faites jusqu'au pas $(n-1)$ que l'on peut calculer de façon récursive.

Un estimateur de $\mathbb{E}[Z]$ au pas n est la moyenne empirique $\overline{Z}_{R_n}$ des observations $Z_1, \ldots, Z_{R_n}$, que l'on calcule récursivement.

Un vecteur directeur v_l du $l^{\text{ième}}$ axe principal de l'ACG est vecteur propre de la matrice M-symétrique $B = CM = (\mathbb{E}[ZZ'] - \mathbb{E}[Z]\mathbb{E}[Z'])M$ associé à la $l^{\text{ième}}$ plus grande valeur propre. En suivant Bouamaine et Monnez (1998), on définit récursivement le processus d'approximation stochastique $(X_n) = ((X_n^1, \ldots, X_n^r))$ de $(v_1, \ldots, v_r)$:

$$\begin{aligned} B_n &= (\frac{1}{r_n} \sum_{l=R_{n-1}+1}^{R_n} Z_l Z_l' - \overline{Z}_{R_n} \overline{Z}_{R_n}')M_n, \\ F_n(X_n^l) &= \frac{<B_n X_n^l, X_n^l>_{M_n}}{||X_n^l||_{M_n}^2}, \\ Y_{n+1}^l &= X_n^l + \frac{a}{n^\alpha}(B_n - F_n(X_n^l)I)X_n^l,\ l = 1, \ldots, r\ , \\ X_{n+1} &= \text{orth}_{M_n}(Y_{n+1}). \end{aligned}$$

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

On peut aussi utiliser au temps n toutes les observations faites jusqu'à ce pas inclus, en définissant $B_n = C_n M_n$, C_n étant la matrice de covariance empirique de Z calculée à partir des observations effectuées jusqu'au pas n.

4 Cas où l'espérance des observations varie dans le temps

On suppose que, pour tout $n \geq 1$, l'espérance mathématique de Z_n, θ_n, dépend du temps n : $Z_n = \theta_n + R_n$, les R_n constituant un échantillon i.i.d. d'un vecteur aléatoire R d'espérance nulle et de matrice de covariance C. On note $\theta_n^k = \mathbb{E}[Z_n^k]$, de dimension m_k, $k = 1, \ldots, q$.

On effectue l'ACG de R, appelée ACG partielle. Des vecteurs directeurs v_l des axes principaux sont vecteurs propres de CM, avec $C = \mathbb{E}[(Z_n - \theta_n)Z_n']$ pour tout $n \geq 1$, M étant la métrique définie dans le paragraphe 2 avec, pour $k = 1, \ldots, q$, $C^k = \mathbb{E}[(Z_n^k - \theta_n^k)(Z_n^k)']$ pour tout $n \geq 1$.

On suppose que l'on dispose au temps n d'un estimateur Θ_n de θ_n (ou, pour $k = 1, \ldots, q$, d'un estimateur Θ_n^k de θ_n^k) vérifiant certaines hypothèses.

Pour $k = 1, \ldots, q$, $(C^k)^{-1}$ est solution de l'équation en X : $\mathbb{E}[(Z_n^k - \theta_n^k)(Z_n^k)'X - I] = 0$ où I est la matrice-identité d'ordre m_k. On définit récursivement le processus d'approximation stochastique de $(C^k)^{-1}$, (M_n^k), par :

$$M_{n+1}^k = M_n^k - a_n((Z_n^k - \Theta_n^k)(Z_n^k)'M_n^k - I).$$

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

On définit récursivement un processus d'approximation stochastique $(X_n) = ((X_n^1, \ldots, X_n^r))$ de $(v_1, \ldots, v_r)$ par :

$$\begin{aligned} B_n &= (Z_n Z_n' - \Theta_n \Theta_n')M_n, \\ F_n(X_n^l) &= \frac{<B_n X_n^l, X_n^l>_{M_n}}{||X_n^l||_{M_n}^2}, \\ Y_{n+1}^l &= X_n^l + \frac{a}{n^\alpha}(B_n - F_n(X_n^l)I)X_n^l,\ l = 1, \ldots, r, \\ X_{n+1} &= \mathrm{orth}_{M_n}(Y_{n+1}). \end{aligned}$$

Un cas particulier de modèle d'évolution dans le temps de l'espérance θ_n de Z_n est le suivant. Si l'on note $\theta_n^{(i)}$ la $i^{\text{ème}}$ composante réelle de θ_n $(i = 1, \ldots, p)$, on définit le modèle linéaire $\theta_n^{(i)} =< \beta^i, U_n^i >$, $< \cdot, \cdot >$ désignant le produit scalaire euclidien usuel dans $\mathbb{R}^{n_i}$, U_n^i étant un vecteur de dimension n_i de valeurs de fonctions connues du temps n ou de variables explicatives contrôlées et β^i un vecteur inconnu de $\mathbb{R}^{n_i}$.

On définit alors comme dans Monnez (2008b), pour $i = 1, \ldots, p$, le processus d'approximation stochastique (B_n^i) de β^i et l'estimateur $\Theta_n^{(i)}$ de $\theta_n^{(i)}$ par :

$$\begin{aligned} B_{n+1}^i &= B_n^i - a_n U_n^i((U_n^i)' B_n^i - Z_n^{(i)}), \\ \Theta_n^{(i)} &=< B_n^i, U_n^i >, \end{aligned}$$

$Z_n^{(i)}$ étant la $i^{\text{ème}}$ composante réelle de Z_n.

5 Mise en oeuvre et conclusion

Les simulations ont été effectuées avec le logiciel R et consistent à tester la précision de notre méthode pour estimer des vecteurs directeurs d'axes principaux de l'ACG dans le cas où les observations sont i.i.d. en la comparant à une méthode classique. L'idée générale en est la suivante :

1) On fixe les paramètres du programme en choisissant le temps durant lequel va tourner l'algorithme (en supposant que le flux de données est continu), la dimension du vecteur Z dont on observe des réalisations et le nombre r de vecteurs à estimer.

2) Initialisation : on prend en compte un petit nombre d'observations afin de calculer une première estimation de la matrice de covariance C, C_0, de la métrique M, M_0, et de vecteurs directeurs $v^1, \ldots, v^r$ des axes principaux, $v_0^1, \ldots, v_0^r$.

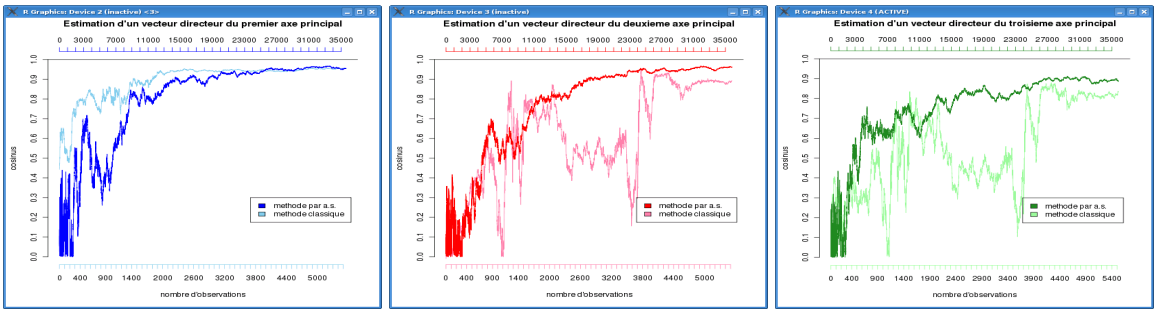
3) Pas n : On introduit un nouveau vecteur d'observations puis on met à jour la matrice de covariance empirique C_n et la métrique empirique M_n à l'aide de formules récursives.

Dans un premier programme, on calcule grâce à la routine Lapack utilisée dans R (mais aussi Scilab, ...) les r premiers vecteurs propres de la matrice $C_n M_n$ qui sont des estimations de vecteurs directeurs des axes principaux de l'ACG de Z.

Dans un deuxième programme, on met en oeuvre le processus décrit à la fin du paragraphe 3 (en utilisant toutes les observations faites jusqu'au pas n) pour obtenir, pour $l = 1, \ldots, r$, une estimation d'un vecteur directeur du $l^{\text{ième}}$ axe principal de l'ACG de Z.

4) Pour un même temps d'exécution, on compare alors la précision des deux méthodes via la valeur du cosinus de l'angle formé par les vecteurs théoriques et calculés en fonction du nombre d'observations prises en compte.

Voici par exemple les résultats obtenus pour l'estimation de vecteurs directeurs des trois premiers axes principaux de l'ACG d'un vecteur Z de dimension 190 réparti en 6 sous-vecteurs de dimension respective $(25, 25, 37, 33, 38, 32)$. On utilise dans ce cas le processus moyennisé avec un pas $a_n = (\frac{10}{n+10})^{0.9999}$ et pendant une durée de 800 secondes.



On réalise le même type de simulation lorsque l'espérance des observations varie dans le temps, en s'appuyant cette fois sur le processus du paragraphe 4.

Un prolongement de cette étude est de considérer le cas où la matrice de covariance des observations varie aussi dans le temps.

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] Nguyen, T.M.N. et Saracco, J. (2010), Estimation récursive en régression inverse par tranches (sliced inverse regression), Journal de la Société Française de Statistique, 151(2), 19-46.

[6] Robbins, H. et Monro, S. (1951), A stochastic approximation method, AMS, 22, 400-407.