Analyse canonique généralisée d'un flux de données d'espérance variable dans le temps

Jean-Marie Monnez, Romain Bar

Analyse Canonique Généralisée partielle d'un flux de données d'espérance variable dans le temps

Romain Bar*, Jean-Marie Monnez**

*Institut Elie Cartan, UMR 7502, Université de Lorraine, CNRS, INRIA
BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, France
Romain.Bar@univ-lorraine.fr,
http://www.iecn.u-nancy.fr
** Jean-Marie.Monnez@univ-lorraine.fr

Résumé. On suppose que des vecteurs de données z_n arrivant séquentiellement dans le temps sont les observations respectives d'un vecteur Z_n dont l'espérance θ_n varie dans le temps. On note $\tilde{Z}_n = Z_n - \theta_n$ et on suppose que les vecteurs $\tilde{Z}_n$ forment un échantillon i.i.d. d'un vecteur aléatoire $\tilde{Z}$. On définit des processus d'approximation stochastique utilisant des blocs de données pour estimer des vecteurs directeurs des axes principaux de l'analyse canonique généralisée (ACG) de $\tilde{Z}$.

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 dans le cas où ces vecteurs arrivent séquentiellement dans le temps : on observe z_n au temps n ; on a donc une suite de vecteurs de données $z_1, \ldots, z_n, \ldots$. On suppose que :

- pour tout n, z_n est la réalisation d'un vecteur aléatoire Z_n défini sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$, d'espérance mathématique θ_n variable dans le temps ;
- les vecteurs aléatoires Z_n sont mutuellement indépendants ;
- les vecteurs aléatoires $\tilde{Z}_n = Z_n - \theta_n$ constituent un échantillon i.i.d d'un vecteur aléatoire $\tilde{Z}$ dans $\mathbb{R}^p$ d'espérance nulle et de matrice de covariance ne dépendant pas de n ;
- le vecteur aléatoire $\tilde{Z}$ est partitionné en sous-vecteurs $\tilde{Z}^1, \ldots, \tilde{Z}^q$; pour $k = 1, \ldots, q$, $\tilde{Z}^k$ est un vecteur aléatoire dans $\mathbb{R}^{m_k}$, de composantes $\tilde{Z}^{k1}, \ldots, \tilde{Z}^{km_k}$; on a $m_1 + \cdots + m_q = p$.

On pose le problème suivant : réaliser une ACP du vecteur aléatoire $\tilde{Z}$ dans laquelle les vecteurs aléatoires $\tilde{Z}^k$ aient un rôle équilibré : on veut éviter que les premiers facteurs soient principalement déterminés à partir de certains vecteurs $\tilde{Z}^k$. L'analyse canonique généralisée du vecteur aléatoire (ACGVA) $\tilde{Z}$, ou ACG partielle, présentée au paragraphe 2, 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 v un résultat de l'ACGVA, par exemple un vecteur directeur d'un axe principal, un facteur principal, une valeur propre.

Plutôt que d'effectuer à chaque temps n une estimation de v à partir de l'ensemble des données dont on dispose jusqu'à ce temps, on va effectuer une estimation récursive de v : disposant d'une estimation v_n de v obtenue à partir des observations $z_1,\dots,z_{n-1}$, 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, en supposant que l'on dispose au temps n d'un estimateur Θ_n de θ_n, que l'on peut aussi éventuellement obtenir par approximation stochastique.
On considère dans le paragraphe 3 le cas où l'on utilise au temps n un bloc de r_n nouvelles observations indépendantes $z_{R_{n-1}+1},\dots,z_{R_n}$ avec $R_n = \sum_{j=1}^{n} r_j$.
On considère dans le paragraphe 4 le cas où l'on utilise au temps n toutes les observations faites jusqu'à ce temps ; dans le paragraphe 5, on présente le cas particulier d'un modèle linéaire de variation de l'espérance.

2 ACG d'un vecteur aléatoire

Dans tout le paragraphe, on adopte la présentation de [4].
On suppose qu'il n'existe pas de relation affine entre les composantes du vecteur aléatoire $\tilde{Z}$.
Le critère de l'ACG est le suivant : pour $l = 1,\dots,r$, déterminer au pas l une combinaison linéaire des composantes centrées de $\tilde{Z}$, $U_l = \xi_l'(\tilde{Z} - \mathbb{E}[\tilde{Z}])$, et pour $k = 1,\dots,q$, une combinaison linéaire des composantes centrées de $\tilde{Z}^k$, $V_l^k = (\eta_l^k)'(\tilde{Z}^k - \mathbb{E}[\tilde{Z}^k])$, telles que :

$$\begin{aligned}
&\sum_{k=1}^{q} \rho^2(U_l, V_l^k) \text{ soit maximal, avec :}\\
&Var(U_l) = 1,\\
&Cov(U_l, U_j) = 0,\ j = 1,\dots,l-1,\\
&Var(V_l^k) = 1,\ k = 1,\dots,q.
\end{aligned}$$

Soit C la matrice de covariance de $\tilde{Z}$, C^k celle de $\tilde{Z}^k$ et M la métrique diagonale par blocs d'ordre p définie par :

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

ξ_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 : ξ_l est le $l^{\text{ième}}$ facteur de l'ACP de $\tilde{Z}$ dans $\mathbb{R}^p$ muni de la métrique M. $v_l = M^{-1}\xi_l$ 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, $\tilde{Z}^k$ est de dimension 1, on retrouve l'ACP normée.

3 Approximation stochastique des vecteurs v_l en utilisant à chaque temps un paquet de données

On suppose qu'au temps n, on dispose d'un bloc de r_n nouvelles observations $z_{R_{n-1}+1},\ldots,z_{R_n}$ et d'estimateurs $(\Theta_{R_{n-1}+1},\ldots,\Theta_{R_n})$ de $(\theta_{R_{n-1}+1},\ldots,\theta_{R_n})$. On note $I_n=\{R_{n-1}+1,\ldots,R_n\}$.

1. On utilise au temps n un estimateur convergent M_n de M, obtenu à partir des observations $z_1,\ldots,z_{R_{n-1}}$.
   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 : $M_{n+1}^k=M_n^k-a_{1_n}((\frac{1}{r_n}\sum_{i\in I_n}(Z_i^k-\Theta_i^k)(Z_i^k)')M_n^k-I)$
   avec $(a_{1_n}=\frac{a}{n^\alpha}$ avec $a>0,\ \frac{1}{2}<\alpha<1)$ ou $(a_{1_n}=\frac{a}{n}$ et $a>\frac{1}{\lambda_{\min}(C)})$.
   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.

2. En suivant [1] et [2], on définit récursivement un processus d'approximation stochastique $(X_n)=((X_n^1,\ldots,X_n^r))$ de $(v_1,\ldots,v_r)$ par :

$$\begin{aligned} B_n &= (\tfrac{1}{r_n}\sum_{i\in I_n}(Z_i-\Theta_i)Z_i')M_n,\\ F_n(X_n^l) &= \tfrac{<B_nX_n^l,X_n^l>_{M_n}}{||X_n^l||^2_{M_n}},\\ Y_{n+1}^l &= X_n^l+a_{2_n}(B_n-F_n(X_n^l)I)X_n^l,\ l=1,\ldots,r,\\ X_{n+1} &= \text{orth}_{M_n}(Y_{n+1}). \end{aligned}$$

Sous les conditions adéquates (en particulier $a_{2_n}>0$, $\sum_{n=1}^{+\infty}a_{2_n}=\infty$, $\sum_{n=1}^{+\infty}(a_{2_n})^2<\infty$), on établit la convergence presque sûre de (X_n).

4 Approximation stochastique des vecteurs v_l en utilisant à chaque temps toutes les données observées depuis le début

On peut aussi utiliser au temps n toutes les observations faites jusqu'à ce pas inclus. Pour cela, on définit récursivement le processus (C_n^k) d'approximation stochastique de C^k, de type Robbins-Monro [5], dans l'ensemble des matrices (m_k,m_k) : $C_{n+1}^k=C_n^k-b_n(C_n^k-(Z_n^k-\Theta_n^k)(Z_n^k)')$
avec $(b_n=\frac{b}{n^\alpha}$ avec $b>0,\ \frac{1}{2}<\alpha<1)$ ou $(b_n=\frac{b}{n}$ et $b>1)$.
On définit ensuite récursivement le processus d'approximation stochastique de $(C^k)^{-1}$, (M_n^k), par : $M_{n+1}^k=M_n^k-b_n(C_n^kM_n^k-I)$.
On utilise toujours 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 note $B_n=C_nM_n$, C_n étant le processus d'approximation stochastique de $C=Var(\tilde{Z})$, de type Robbins-Monro, dans l'ensemble des matrices (p,p) : $C_{n+1}=C_n-b_n(C_n-(Z_n-\Theta_n)(Z_n)')$.
On définit alors le même processus d'approximation stochastique qu'au paragraphe précédent $(X_n)=((X_n^1,\ldots,X_n^r))$ de $(v_1,\ldots,v_r)$ avec la nouvelle définition de B_n.

5 Cas particulier d'un modèle linéaire de variation de l'espérance

A l'instar de [3], on choisit un modèle linéaire pour représenter l'espérance. Plus précisément, pour $i = 1, \ldots, p$, on suppose qu'il existe un vecteur β^i inconnu de $\mathbb{R}^{n_i}$ et, pour tout n, un vecteur U_n^i de $\mathbb{R}^{n_i}$ connu au temps n tels que la i$^{\text{ème}}$ composante réelle de θ_n, θ_n^i, soit égale à $< \beta^i, U_n^i >$.
On définit le processus d'a.s. (B_n^i) de β^i tel que : $B_{n+1}^i = B_n^i - a_n U_n^i((U_n^i)' B_n^i - Z_n^i)$.
On définit aussi $\Theta_n^i =< B_n^i, U_n^i >$, $\Theta_n = (\Theta_n^1, \ldots, \Theta_n^p)'$ que l'on introduit dans la définition des processus d'approximation stochastique des vecteurs v_l.

6 Conclusion

On a présenté deux processus d'approximation stochastique de vecteurs directeurs des axes principaux d'une analyse canonique généralisée.
Des tests par simulation ont été réalisés en langage R pour comparer les variantes possibles ; plusieurs choix des suites (a_{1_n}), (a_{2_n}), (b_n) et différentes initialisations des algorithmes ont également été mis en oeuvre.
Un prolongement de cette étude est de considérer le cas où la matrice de covariance des observations varie aussi dans le temps.

Références

[1] J.P. Benzecri. Approximation stochastique dans une algèbre normée non commutative. *Bulletin de la SMF*, 97 :225–241, 1969.

[2] A. Bouamaine and J.M Monnez. Approximation stochastique de vecteurs et valeurs propres. *Publications de l'ISUP*, 42, n° 2-3 :15–38, 1998.

[3] J.M. Monnez. Analyse en composantes principales d'un flux de données d'espérance variable dans le temps. *RNTI*, C-2 :43–56, 2008.

[4] J.M. Monnez. Stochastic approximation of the factors of a generalized canonical correlation analysis. *Statistics & Probability Letters*, 78 :2210–2216, 2008.

[5] H. Robbins and S. Monro. A stochastic approximation method. *AMS*, 22 :400–407, 1951.

Summary

Consider a data stream and suppose that each multidimensional data z_n is a realization of a random vector Z_n whose expectation θ_n varies with time. Let $\tilde{Z}_n = Z_n - \theta_n$ and suppose that the vectors $\tilde{Z}_n$ form an i.i.d. sample of a random vector $\tilde{Z}$. Stochastic approximation processes using data blocks are used to estimate on-line direction vectors of the principal axes of the generalized canonical correlation analysis of $\tilde{Z}$.