ACP PROJETÉE DE DONNEES SEQUENTIELLES

Jean-Marie Monnez

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

Résumé. On suppose que des vecteurs de données arrivant séquentiellement dans le temps sont des observations i.i.d. d'un vecteur aléatoire Z = (R,S) partitionné en deux vecteurs R et S. On définit une méthode récursive d'estimation séquentielle des r premiers facteurs de l'ACP projetée de R par rapport à S. On étudie ensuite le cas particulier de l'analyse canonique pour lequel on définit deux autres processus spécifiques, ainsi que celui de l'analyse factorielle discriminante.

Abstract. Suppose that vectors of data are i.i.d. observations which are taken sequentially of a random vector Z = (R,S). We define a recursive method of sequential estimation of the first factors of a projected principal component analysis of R with respect to S. In the particular case of the canonical correlation analysis, we define two specific stochastic approximation processes and we study the case of the factorial discriminant analysis.

Mots-Clés. Analyse de données séquentielles, processus d'approximation stochastique, analyse en composantes principales projetée, analyse canonique, analyse factorielle discriminante.

1 INTRODUCTION

On observe p caractères quantitatifs $r^1, ..., r^p$ et q caractères quantitatifs $s^1, ..., s^q$ sur des individus : on obtient des vecteurs de données (r_i, s_i) dans R^{p+q}. On se place dans le cas où ces vecteurs de données arrivent séquentiellement dans le temps : on observe (r_n, s_n) au temps n ; on a une suite de vecteurs de données $(r_1, s_1), ..., (r_n, s_n), ...$

On suppose que cette suite constitue un échantillon i.i.d. d'un vecteur aléatoire (R, S) de dimension $p+q$ défini sur un espace probabilisé (Ω, A, P) : Ω représente une population d'où on extrait un échantillon.

On étudie l'ACP projetée de R par rapport à S, c'est-à-dire l'ACP du régressé linéaire de R par rapport à S, qui représente l'ACP projetée effectuée sur la population Ω, dont on va chercher à estimer au temps n les résultats à partir de l'échantillon dont on dispose à ce temps.

Soit un résultat θ de cette analyse, par exemple une valeur propre, un facteur (on considère ici le cas d'un facteur). Plutôt que d'effectuer au temps n une estimation de θ en utilisant toutes les données du passé disponibles, on va effectuer une estimation récursive de θ : disposant d'une estimation θ_n obtenue à partir des observations $(r_1, s_1), ..., (r_{n-1}, s_{n-1})$, on introduit l'observation (r_n, s_n) et on définit une nouvelle estimation θ_{n+1} fonction de θ_n et de (r_n, s_n) :

$$\theta_{n+1} = f_n(\theta_n; r_n, s_n).$$

On définit pour cela un processus d'approximation stochastique comme dans l'article de Bouamaine et Monnez (1998).

L'ACP projetée a pour cas particulier l'analyse canonique pour laquelle on peut définir en outre des processus spécifiques, ainsi que l'analyse factorielle discriminante.

2 ACP PROJETEE D'UN VECTEUR ALEATOIRE R PAR RAPPORT A UN VECTEUR ALEATOIRE S

Soit deux vecteurs aléatoires $R = (R^1, ..., R^p)$ de dimension p et $S = (S^1, ..., S^q)$ de dimension q. On effectue la régression linéaire au sens des moindres carrés de R par rapport à S : on détermine la matrice A de dimensions (q, p) et la matrice-colonne D de dimensions $(p, 1)$ qui rendent minimale l'espérance mathématique de $\|R - A'S - D\|^2$. On appelle régressé de R par rapport à S le vecteur aléatoire de dimension p $\hat{E}[R|S] = A'S + D$.

L'ACP projetée de R par rapport à S est l'ACP du régressé $\hat{E}[R|S] = A'S + D$ dans R^p muni d'une métrique M. Le l^{ième} facteur θ_l de cette ACP est vecteur propre de $M\ Cov(\hat{E}[R|S])$ associé à la l^{ième} plus grande valeur propre λ_l.

3 CAS OU IL N'Y A PAS DE RELATION AFFINE ENTRE LES COMPOSANTES DE S

On a dans ce cas : $A = (Cov(S))^{-1} Cov(S, R), D = E[R] - A'E[S]$.
En outre : $Cov(\hat{E}[R|S]) = Cov(R, S)(Cov(S))^{-1} Cov(S, R)$.

3.1 Approximation stochastique de A

Soit le vecteur aléatoire S_1 de dimension $q+1$ obtenu en ajoutant à S une dernière composante égale à 1. On établit que la matrice de dimensions $(q+1, p)$, $A_1 = \begin{pmatrix} A \\ D' \end{pmatrix}$ est la solution du système en X :

$$E[S_1 S_1' X - S_1 R'] = 0.$$

$((R_n, S_n))$ étant un échantillon i.i.d. de (R, S) et $S_{1n} = (S_n' \ \ 1)'$, on définit le processus (A_{1n}) d'approximation stochastique de A_1 de type Robbins-Monro dans l'ensemble des matrices $(q+1, p)$:

$$A_{1,n+1} = A_{1n} - a_n (S_{1n} S_{1n}' A_{1n} - S_{1n} R_n'),$$

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

En enlevant pour tout n la dernière ligne de A_{1n}, on obtient le processus (A_n) dans l'ensemble des matrices (q, p) qui converge presque sûrement vers A comme c'est établi dans l'article de Monnez (2008a).

3.2 Approximation stochastique des facteurs

Les facteurs θ_l sont vecteurs propres M^{-1}-orthonormés de

$$B = M\ Cov(R, S)\, A = M(E[RS'] - E[R]E[S'])A.$$

On définit deux processus récursifs (M_n) et (N_n) convergeant presque sûrement respectivement vers M et M^{-1}.

On définit à partir des observations $(R_1, S_1), ...,(R_{n-1}, S_{n-1})$ les moyennes d'ordre n-1, $\overline{R}_{n-1}$ et $\overline{S}_{n-1}$, estimateurs respectifs des espérances de R et S, que l'on calcule récursivement. On définit alors la matrice aléatoire

$$B_n = M_n \left(R_n S_n' - \overline{R}_{n-1} \overline{S}_{n-1}' \right) A_n .$$

La fonction

$$F(x) = \frac{\langle Bx, x \rangle_{M^{-1}}}{\langle x, x \rangle_{M^{-1}}}$$

est maximale pour $x = \theta_1$ et vaut alors λ_1 ; sous la contrainte que x est orthogonal à $\theta_1, ..., \theta_{l-1}$, elle est maximale pour $x = \theta_l$ et vaut alors λ_l. On définit la fonction aléatoire

$$F_n(x) = \frac{\langle B_n x, x \rangle_{N_n}}{\langle x, x \rangle_{N_n}}.$$

En suivant l'article de Bouamaine et Monnez (1998), on définit le processus d'approximation stochastique des r premiers facteurs de l'ACP projetée, $(X_n) = \left(\left(X_n^1, ..., X_n^r \right) \right)$:

$$Y_{n+1}^l = X_n^l + a_n \left(B_n - F_n \left(X_n^l \right) I \right) X_n^l, \; l = 1, ..., r,$$
$$X_{n+1} = orth_{N_n} \left(Y_{n+1} \right).$$

Pour obtenir X_{n+1}, on effectue une orthogonalisation au sens de Gram-Schmidt par rapport à N_n de $Y_{n+1} = \left(Y_{n+1}^1, ..., Y_{n+1}^r \right)$.

On établit la convergence de ce processus pour $\frac{2}{3} < \alpha \leq 1.$

3.3 Variantes

On peut définir des variantes de ce processus en utilisant au pas n, au lieu d'une seule observation (R_n, S_n), plusieurs observations de (R, S), ou encore toutes les observations faites jusqu'à ce pas, en remplaçant dans la définition de B_n le produit $R_n S_n'$ par la moyenne des produits $R_i S_i'$ considérés à ce pas.

4 CAS DE L'ANALYSE CANONIQUE DE DEUX VECTEURS ALEATOIRES

On suppose qu'il n'existe pas de relation affine entre les composantes du vecteur aléatoire (R, S). Le critère de l'analyse canonique est : pour $l = 1, ..., r$, déterminer au pas l une combinaison linéaire des composantes centrées de R, $U_l = \theta_l' (R - E[R])$, et une combinaison linéaire des composantes centrées de S, $V_l = \eta_l' (S - E[S])$, telles que :

$$\rho(U_l, V_l) \quad \max,$$
$$Var(U_l) = 1, \; Cov(U_l, U_j) = 0, \; j = 1, ..., l-1,$$
$$Var(V_l) = 1, \; Cov(V_l, V_j) = 0, \; j = 1, ..., l-1.$$

Les facteurs canoniques θ_l, vecteurs propres de $(Cov(R))^{-1} (Cov(R,S)) (Cov(S))^{-1} Cov(S,R))$ sont

aussi les facteurs de l'ACP projetée de R par rapport à S en prenant dans R^p la métrique $(Cov(R))^{-1}$. De même, les facteurs canoniques η_l sont les facteurs de l'ACP projetée de S par rapport à R en prenant dans R^q la métrique $(Cov(S))^{-1}$.
On peut estimer les facteurs θ_l et η_l séparément ou simultanément.

4.1 Application de la méthode générale

On peut estimer *Cov(R)* au temps *n* par la matrice de covariance empirique N_n de l'échantillon $(R_1, ...,R_{n-1})$ calculée récursivement. On peut estimer $(Cov(R))^{-1}$ au temps *n* par l'inverse M_n de N_n, que l'on peut calculer récursivement. Une autre solution est la suivante.

Soit le vecteur aléatoire R_1 de dimension $p+1$ obtenu en ajoutant à R une dernière composante égale à 1. Soit J la matrice de dimensions $(p+1, p)$ obtenue en ajoutant à la matrice-identité d'ordre p une ligne de zéros. On établit que la matrice de dimensions $(p+1, p)$

$$F_1 = \begin{pmatrix} (Cov(R))^{-1} \\ -E[R'](Cov(R))^{-1} \end{pmatrix}$$

est la solution du système en X :

$$E[R_1 R_1' X - J] = 0 .$$

On définit alors le processus (M_{1n}) d'approximation stochastique de F_1 de type Robbins-Monro dans l'ensemble des matrices $(p+1, p)$:

$$M_{1,n+1} = M_{1n} - a_n (R_{1n} R_{1n}' M_{1n} - J) .$$

En enlevant pour tout n la dernière ligne de M_{1n}, on obtient le processus (M_n) d'approximation stochastique de $(Cov(R))^{-1}$.

Pour estimer les r premiers facteurs θ_l, on utilise alors le processus défini dans le paragraphe 3.2.

4.2 Méthode spécifique

On note : $G = (Cov(R))^{-1} Cov(R, S)$.
Les facteurs θ_l sont vecteurs propres de GA et les facteurs η_l de AG.

On a défini un processus (A_n) d'approximation stochastique de A. En permutant les rôles de R et S, on définit de même un processus (G_n) d'approximation stochastique de G.

On pose alors $B_n = G_n A_n$, estimateur d'ordre n de GA.
On introduit cette définition de B_n dans le processus d'approximation stochastique de facteurs défini dans le paragraphe 3.2.

4.3 Estimation simultanée des facteurs en R et en S

On a les relations de transition suivantes entre les facteurs :

$$\eta_l = \frac{1}{\sqrt{\lambda_l}} A\theta_l , \theta_l = \frac{1}{\sqrt{\lambda_l}} G\eta_l .$$

On établit que $\xi_l = \begin{pmatrix} \theta_l \\ \eta_l \end{pmatrix}$ est vecteur propre de la matrice carrée d'ordre $p+q$, $B = \begin{pmatrix} I & G \\ A & I \end{pmatrix}$, associé à la l[ième] plus grande valeur propre $1+\sqrt{\lambda_l}$. ξ_l est le l[ième] facteur de l'analyse canonique généralisée du vecteur aléatoire $Z = (R, S)$ étudiée dans l'article de Monnez (2008a).

On pose alors $B_n = \begin{pmatrix} I & G_n \\ A_n & I \end{pmatrix}$.
On introduit cette définition de B_n dans le processus d'approximation stochastique de facteurs défini dans le paragraphe 3.2.

5 CAS OU LES COMPOSANTES DE S SONT LES INDICATRICES DES MODALITES EXCLUSIVES D'UNE VARIABLE ALEATOIRE NOMINALE

Dans ce cas, on a : $E[R|S] = A_2 'S$, avec $A_2 = (E[SS'])^{-1} E[SR']$.
En outre : $Cov(E[R|S]) = Cov(R,S)\, A_2$.

5.1 Approximation stochastique des facteurs

Les facteurs de l'ACP projetée de R par rapport à S dans R^p muni d'une métrique M sont vecteurs propres de la matrice M^{-1}-symétrique $B = M\, Cov(R, S)\, A_2$.

On définit deux processus récursifs (M_n) et (N_n) convergeant presque sûrement respectivement vers M et M^{-1}.

A_2 est solution du système en X :

$$E[SS'\, X - SR'] = 0 .$$

On définit le processus (A_{2n}) d'approximation stochastique de A_2 :

$$A_{2,n+1} = A_{2n} - a_n \left(S_n S_n' A_{2n} - S_n R_n'\right).$$

On pose alors : $B_n = M_n \left(R_n S_n' - \overline{R}_{n-1} \overline{S}_{n-1}'\right) A_{2n}$.
On introduit cette définition de B_n dans le processus d'approximation stochastique de facteurs défini dans le paragraphe 3.2.

5.2 Cas particulier de l'analyse factorielle discriminante

On prend dans ce cas particulier la métrique $M = (Cov(R))^{-1}$.

On peut utiliser comme estimateurs respectifs de M et M^{-1} les estimateurs (M_n) et (N_n) définis dans le paragraphe 4.1. On utilise alors le processus défini dans le paragraphe 5.1.

Un autre processus est le suivant. Les facteurs de l'analyse sont vecteurs propres de

$$B = (Cov(R))^{-1}\, Cov(R,S) \left(E[SS']\right)^{-1}\, E[SR'] = G\, A_2 .$$

On utilise le processus (G_n) d'approximation stochastique de G défini dans le paragraphe 4.2.

On considère alors, comme on l'a vu dans le paragraphe 4.2, l'estimateur de B, $B_n = G_n\, A_{2n}$.

On introduit cette définition de B_n dans le processus d'approximation stochastique de facteurs défini dans le paragraphe 3.2.

6. CONCLUSION

En supposant que les observations d'un vecteur aléatoire *(R, S)* arrivent séquentiellement dans le temps, on a défini un processus d'approximation stochastique des facteurs d'une ACP projetée du vecteur aléatoire *R* par rapport au vecteur aléatoire *S*, en considérant le cas où il n'existe pas de relation affine entre les composantes de *S* et celui où les composantes de *S* sont les indicatrices des modalités exclusives d'une variable aléatoire nominale. On a traité les cas particuliers de l'analyse canonique et de l'analyse factorielle discriminante, pour lesquels on a défini des processus spécifiques.

Une extension concernant l'analyse canonique est l'approximation stochastique des facteurs d'une analyse canonique généralisée.

On peut également étudier des cas où la loi des observations effectuées n'est pas stationnaire dans le temps, comme ceci a été entrepris dans le cas de l'ACP dans l'article de Monnez (2008b).

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] Krasulina, T.P. (1970) Method of stochastic approximation in the determination of the largest eigenvalue of the mathematical expectation of random matrices, *Automation and remote control*, 2, 215-221.
[4] Monnez, J.M. (2008a) Stochastic approximation of the factors of a generalized canonical correlation analysis, *Statistics & Probability Letters*, 78, 2210-2216.
[5] Monnez, J.M. (2008b) Analyse en composantes principales d'un flux de données d'espérance variable dans le temps, *RNTI*, C-2, 43-56.
[6] Robbins, H. et Monro, S. (1954) A stochastic approximation method, *AMS*, 22, 400-407.