
Une méthode d'ACP de données en ligne

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

RÉSUMÉ. Des vecteurs de données arrivant en ligne sont considérés comme des réalisations indépendantes d'un vecteur aléatoire. On établit dans ce cadre un résultat de convergence presque sûre d'un processus d'approximation stochastique des facteurs de l'ACP de ce vecteur aléatoire. On peut l'appliquer par exemple à l'analyse factorielle multiple. On étudie ensuite le cas où l'espérance mathématique du vecteur aléatoire varie dans le temps selon un modèle linéaire.

MOTS-CLÉS : analyse de données en ligne, approximation stochastique, analyse en composantes principales, analyse factorielle multiple.

1. Introduction

On observe p caractères quantitatifs sur n individus : on obtient des vecteurs de données z_1, \dots, z_n dans R^p . On peut effectuer une ACP du tableau de données. La métrique utilisée, qui dépend des données, est a priori quelconque : on peut souhaiter effectuer par exemple une ACP normée ou une analyse factorielle multiple (AFM) ou une analyse canonique généralisée (ACG).

On considère ici le cas où les vecteurs de données arrivent séquentiellement dans le temps : on observe z_n au temps n . On a une suite de vecteurs de données z_1, \dots, z_n, \dots .

Supposons dans un premier temps que z_1, \dots, z_n, \dots constituent un échantillon i.i.d. d'un vecteur aléatoire Z défini sur un espace probabilisé (Ω, \mathcal{A}, P) . Ω représente une population d'où on a extrait un échantillon. On peut définir une ACP de ce vecteur aléatoire (ACPVA), présentée dans le paragraphe 2, qui représente l'ACP effectuée sur la population, dont on va chercher à estimer au temps n les résultats à partir de l'échantillon dont on dispose à ce temps. Soit θ un résultat de l'ACPVA, par exemple une valeur propre, un facteur (on considère ici le cas d'un facteur). On peut effectuer une estimation récursive de θ : disposant d'une estimation θ_n de θ obtenue à partir des observations z_1, \dots, z_{n-1} , on introduit l'observation z_n et on définit à partir de θ_n et z_n une nouvelle estimation θ_{n+1} de θ . On utilise pour cela un processus d'approximation stochastique défini dans le paragraphe 4, dont on établit la convergence. Ce processus est une version stochastique d'une méthode itérative de gradient définie dans le paragraphe 3. On présente des variantes de ce processus dans le paragraphe 5.

Considérons dans un deuxième temps le cas où la loi de Z évolue dans le temps. On étudie dans le paragraphe 6 le cas où l'espérance mathématique de Z varie dans le temps selon un modèle linéaire. On estime simultanément les paramètres du modèle linéaire et le résultat de l'ACPVA par des processus d'approximation stochastique.

2. ACP d'un vecteur aléatoire

Soit un vecteur aléatoire Z dans R^p . R^p est muni d'une métrique M . L'ACP du vecteur aléatoire Z consiste à : 1) rechercher une combinaison linéaire des composantes centrées de Z , $f^1(Z-E(Z))$, f^1 appartenant au dual R^{p*} de R^p , de variance maximale sous la contrainte de normalisation $f^1 M^{-1} f^1 = 1$; 2) rechercher une deuxième

combinaison linéaire des composantes de $Z, f^{2'}(Z-E(Z))$, non corrélée à la première, de variance maximale sous la contrainte $f^{2'}M^{-1}f^2 = 1$; 3) et ainsi de suite jusqu'à un rang r au plus égal à p .

La $i^{\text{ème}}$ combinaison linéaire est appelée le $i^{\text{ème}}$ facteur ; on appelle également $i^{\text{ème}}$ facteur le vecteur f^i . Soit

$$C = E((Z - E(Z))(Z - E(Z))') = E(ZZ') - E(Z)E(Z')$$

la matrice de covariance de Z, f^i est vecteur propre M^{-1} unitaire de MC associé à la $i^{\text{ème}}$ plus grande valeur propre.

Si Z a un ensemble fini de N réalisations, l'ACP de Z équivaut à l'ACP usuelle du tableau (N,p) des réalisations, le poids de chaque réalisation étant défini par sa probabilité.

3. Une méthode itérative de détermination des facteurs

On suppose dans ce paragraphe que la matrice de covariance C et la métrique M sont connues.

La fonction $F(x) = \frac{\langle MCx, x \rangle_{M^{-1}}}{\langle x, x \rangle_{M^{-1}}}$ est maximale pour $x = f^1$ et minimale pour $x = f^p$, de gradient

$$G(x) = \frac{2M^{-1}}{x'M^{-1}x} (MC - F(x)I)x.$$

Pour déterminer f^1 , on peut utiliser un processus de gradient (X_n) défini récursivement par

$$X_{n+1} = X_n + a_n (MC - F(X_n)I)X_n.$$

Pour déterminer les r premiers facteurs, on peut utiliser le processus suivant :

$$X_{n+1}^i = \text{orth}_{M^{-1}}(X_n^i + a_n (MC - F(X_n^i)I)X_n^i), i = 1, \dots, r.$$

$X_{n+1}^i = \text{orth}(Y_{n+1}^i)$ signifie que $(X_{n+1}^1, \dots, X_{n+1}^i)$ est obtenu à partir de $(Y_{n+1}^1, \dots, Y_{n+1}^i)$ par une orthogonalisation de Gram-Schmidt au sens de M^{-1} . En supposant les r plus grandes valeurs propres de MC

distinctes, alors, pour $i=1, \dots, r$, le processus $(\frac{X_n^i}{\|X_n^i\|_{M^{-1}}})$ converge vers f^i , en prenant la suite (a_n) telle que

$$a_n > 0, \quad \sum_1^\infty a_n = \infty, \quad \sum_1^\infty \frac{a_n}{\sqrt{n}} < \infty, \quad \sum_1^\infty a_n^2 < \infty.$$

4. Approximation stochastique des facteurs

On suppose maintenant que $E(Z), C$ et M sont inconnus et que l'on dispose d'une suite d'observations (Z_1, \dots, Z_n, \dots) arrivant dans le temps et constituant un échantillon i.i.d. de Z .

Soit, au temps n, M_n un estimateur de M et Θ_n un estimateur de $E(Z)$ fonctions de Z_1, \dots, Z_{n-1} . Soit

$$B_n = M_n(Z_n Z_n' - \Theta_n \Theta_n'), \quad F_n(X_n^i) = \frac{\langle B_n X_n^i, X_n^i \rangle_{M_n^{-1}}}{\langle X_n^i, X_n^i \rangle_{M_n^{-1}}}.$$

On définit le processus d'approximation stochastique :

$$X_{n+1}^i = \text{orth}_{M_n^{-1}}(X_n^i + a_n (B_n - F_n(X_n^i)I)X_n^i), \quad i = 1, \dots, r.$$

Sous les hypothèses précédentes sur la suite (a_n) et les hypothèses complémentaires

$$M_n \xrightarrow{p.s.} M, \quad \sum_1^\infty a_n \|M_n - M\| < \infty \text{ p.s.},$$

$$\Theta_n - E(Z) \xrightarrow{p.s.} 0, \quad \sum_1^\infty a_n \|\Theta_n - E(Z)\| < \infty \text{ p.s.},$$

on établit à partir d'un théorème démontré dans [BOU 98] la convergence presque sûre du processus

$$\left(\frac{X_n^i}{\|X_n^i\|_{M_n^{-1}}} \right) \text{ vers } f^i \text{ pour } i=1, \dots, r.$$

Par exemple, dans le cas de l'analyse factorielle multiple de Z , qui est une ACP de Z avec un choix particulier de métrique M , on peut définir un processus d'approximation stochastique (M_n) convergeant presque sûrement vers M et établir alors la convergence presque sûre en direction du processus (X_n^1, \dots, X_n^r) vers les r premiers facteurs [MON 06].

5. Variantes

1) Au pas n , on peut utiliser plusieurs observations Z_{n1}, \dots, Z_{nm_n} de Z . On définit alors :

$$B_n = M_n \left(\frac{1}{m_n} \sum_{k=1}^{m_n} Z_{nk} Z_{nk}' - \Theta_n \Theta_n' \right).$$

2) Au pas n , on peut utiliser toutes les observations faites jusqu'à ce pas. On définit alors :

$$B_n = M_n \left(\frac{1}{n} \sum_{i=1}^n Z_i Z_i' - \Theta_n \Theta_n' \right).$$

6. Cas où l'espérance de Z_n est fonction du temps n

On suppose que l'on dispose d'une suite d'observations (Z_1, \dots, Z_n, \dots) arrivant dans le temps telles que $E(Z_n) = \theta_n$ dépende du temps n et que les vecteurs $R_n = Z_n - E(Z_n)$ constituent un échantillon i.i.d. d'un vecteur aléatoire R de matrice de covariance C . Les facteurs de l'ACP de R sont vecteurs propres de MC .

Considérons le cas d'un modèle linéaire d'évolution de l'espérance de Z_n défini de la façon suivante.

Soit $\theta_n^1, \dots, \theta_n^p$ les composantes de l'espérance θ_n de Z_n . On suppose que pour $k=1, \dots, p$,

$$\theta_n^k = \langle \beta^k, U_n^k \rangle, \beta^k \in \mathbb{R}^{q_k}, U_n^k \in \mathbb{R}^{q_k};$$

β^k est un vecteur inconnu et U_n^k un vecteur connu au temps n à q_k composantes.

Pour estimer les paramètres β^k , on utilise les processus d'approximation stochastique (B_n^k) tels que

$$B_{n+1}^k = B_n^k - a_n U_n^k (U_n^k B_n^k - Z_n^k), k = 1, \dots, p.$$

Soit $\Theta_n^k = \langle B_n^k, U_n^k \rangle$, $\Theta_n = (\Theta_n^1, \dots, \Theta_n^p)'$, $B_n = M_n (Z_n Z_n' - \Theta_n \Theta_n')$. On définit le processus (X_n^1, \dots, X_n^r) comme dans le paragraphe 3 ; on en établit la convergence presque sûre vers les facteurs de l'ACP de R en faisant des hypothèses complémentaires portant sur les U_n^k [MON 08b].

7. Conclusion

Dans le cas où la loi de Z n'évolue pas dans le temps, on a défini un processus d'approximation stochastique des facteurs et donné un résultat général de convergence qui a été appliqué à l'ACP, l'AFM et l'ACG.

Ce résultat étend au cas de plusieurs facteurs et au cas où l'espérance de Z et la métrique M sont inconnues un résultat de convergence vers le premier facteur lorsque l'espérance de Z est connue et la métrique M est l'identité que l'on déduit d'un théorème de Krasulina [KRA 70]. Dans le cas où la métrique M est connue, la méthode d'orthogonalisation a été utilisée par Benzécri [BEN 69] dans le cadre d'un autre processus.

Dans le cas où l'espérance mathématique de Z évolue dans le temps selon un modèle linéaire, on a établi un résultat de convergence qui a été appliqué à l'ACP normée [MON 08b]. Dans une autre étude en préparation, on considère l'application à l'ACG ; on traite également le cas de modèles non linéaires.

On peut mettre en œuvre ces processus pour effectuer des ACP en ligne de données arrivant en ligne.

8. Bibliographie

- [BEN 69] BENZECRI J.P., "Approximation stochastique dans une algèbre normée non commutative", *Bulletin de la SMF*, vol. 97, 1969, p. 225-241.
- [BOU 98] BOUAMAIN A., MONNEZ J.M., "Approximation stochastique de vecteurs et valeurs propres", *Publications de l'ISUP*, vol. 42, n° 2-3, 1998, p. 15-38.
- [KRA 70] KRASULINA T.P., "Method of stochastic approximation in the determination of the largest eigenvalue of the mathematical expectation of random matrices", *Automation and Remote Control*, vol. 2, 1970, p. 215-221.
- [MON 06] MONNEZ J.M., "Approximation stochastique en analyse factorielle multiple", *Publications de l'ISUP*, vol. 50, n° 3, 2006, p. 27-45.
- [MON 08a] MONNEZ J.M., "Stochastic approximation of the factors of a generalized canonical correlation analysis", *Statistics & Probability Letters*, vol. 78, n° 14, 2008, p. 2210-2216.
- [MON 08b] MONNEZ J.M., "Analyse en composantes principales d'un flux de données d'espérance variable dans le temps", *Revue des Nouvelles Technologies de l'Information*, Vol C-2, 2008, p. 43-56.