Approximation stochastique de vecteurs et valeurs propres. Application à l'ACG en ligne

Jean-Marie Monnez

Approximation stochastique de vecteurs et valeurs propres. Application à l'ACG en ligne.

Jean-Marie Monnez [1,2,*]

[1] *Université de Lorraine, CNRS, Inria*, IECL**, F-54000 Nancy, France*
**Inria, Project-Team BIGS, F-54600 Villers-lès-Nancy*
***IECL, Institut Elie Cartan de Lorraine, F-54506 Vandœuvre-lès-Nancy*
[2] *INSERM U1116, Centre d'Investigation Clinique Plurithématique 1433, Université de Lorraine, Nancy, France*
∗jean-marie.monnez@univ-lorraine.fr
Financement : Programme Investissement d'Avenir ANR-15-RHU-0004

Résumé. Nous avons étendu le domaine d'application du processus d'approximation stochastique de vecteurs propres de Oja en en démontrant la convergence presque sûre sous des hypothèses plus générales. Nous étudions l'application à l'analyse canonique généralisée (ACG) d'un vecteur aléatoire Z dans le cas de données massives ou en flux. Les composantes générales de l'ACG sont les composantes principales de l'ACP de Z avec une métrique particulière M. Nous définissons des processus d'approximation stochastique où l'on peut utiliser à chaque étape toutes les observations de Z effectuées jusqu'à cette étape sans avoir à les stocker au lieu uniquement des nouvelles observations à ce pas, pour estimer simultanément la métrique M, les composantes générales de l'ACG et les valeurs propres associées.

Mots-clés. Analyse canonique généralisée, Approximation stochastique, Données massives, Estimation en ligne, Flux de données, Valeurs propres, Vecteurs propres.

Abstract. We widened the scope of the 0ja's eigenvector stochastic approximation process proving its almost sure convergence under more general assumptions. We study the application to generalized canonical correlation analysis (gCCA) of a random vector Z in the case of big or streaming data. The general components of gCCA are principal components of PCA of Z with a particular metric M. We define stochastic approximation processses using at each step all observations up to this step without storing them instead of the new observations at this step only, to estimate simultaneously the metric M, the general components of gCCA and the corresponding eigenvalues.

Keywords. Big data, Data stream, Eigenvalues, Eigenvectors, Generalized canonical correlation analysis, Online estimation, Stochastic approximation.

1 Analyse de données massives ou en flux

Dans le contexte d'un flux de données, on peut utiliser des algorithmes récursifs pour estimer en ligne des paramètres d'intérêt, par exemple :

- les paramètres d'une fonction de régression linéaire (Duarte et al, 2018) ou logistique (Lalloué et al, 2019) ;

- les centres de classe en classification non supervisée (Cardot et al, 2012) ;

- des composantes principales en ACP (Monnez et Skiredj, 2019).

Le principe en est que chaque vecteur de données entrant est utilisé pour actualiser l'estimation courante des paramètres d'intérêt. Dans le cas d'un tableau de données massives, ce type de méthode peut aussi être utilisé en effectuant dans le temps une succession de tirages au hasard de lignes du tableau.

Les avantages de ces méthodes séquentielles sont multiples :

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

- on peut prendre en compte beaucoup plus de données qu'avec les méthodes non séquentielles durant la même durée de temps ;

- elles utilisent moins de place que les méthodes non séquentielles.

Nous avons montré que l'on pouvait utiliser à chaque étape dans certaines méthodes, au lieu d'un lot de nouvelles données, toutes les données jusqu'à l'étape courante sans avoir à les stocker, donc prendre en compte toute l'information contenue dans les données précédentes et améliorer ainsi en général la vitesse de convergence des processus, comme cela a été vérifié empiriquement dans le cas de la régression linéaire en ligne (Duarte et al, 2018) et l'ACP en ligne (Monnez et Skiredj, 2019).

2 L'algorithme de Oja

Soit B une matrice (p,p) symétrique de vecteurs propres normés $V_1, ..., V_p$ associés aux valeurs propres $\lambda_1 > ... > \lambda_p$. Soit (a_n) une suite de nombres réels positifs. $\|x\|$ désigne la norme euclidienne usuelle d'un vecteur x de $\mathbb{R}^p$, la norme matricielle est la norme spectrale.

Supposons que B soit inconnue et qu'il existe une suite $(B_1, ..., B_n, ...)$ de matrices aléatoires symétriques mutuellement indépendantes, bornées presque sûrement et d'espérance mathématique B. Soit le processus stochastique normé $(X_n, n \geqslant 1)$ défini par Oja et Karhunen (1985) tel que :

$$X_{n+1} = \frac{(I + a_n B_n) X_n}{\|(I + a_n B_n) X_n\|}. \tag{1}$$

La convergence presque sûre de ce processus vers un vecteur propre normé associé à la plus grande valeur propre de B a été établie par Oja et Karhunen (1985). Sa rapidité de convergence est étudiée dans Balsubramani et al (2013). Remarquons que, T_n étant la tribu engendrée par $X_1, B_1, ..., B_{n-1}$, on a $E[B_n \mid T_n] = B$.

Dans le cas de l'ACP d'un vecteur aléatoire Z, $\mathbb{R}^p$ étant muni d'une métrique M qui peut dépendre de caractéristiques de Z, pour déterminer les composantes principales on recherche les premiers vecteurs propres de la matrice M^{-1}symétrique $B =$

$ME\left[(Z-E[Z])(Z-E[Z])'\right]$. Dans le cas d'un flux d'observations de Z, $E[Z]$ et M ne sont pas connues a priori, mais peuvent être estimées en ligne. L'hypothèse d'indépendance des B_n n'est alors pas vérifiée.

De façon générale, soit une matrice Q-symétrique B, pour $n>1$, Q_n une métrique connue à l'étape $n-1$ convergeant presque sûrement vers Q, $\langle .,. \rangle_n$ et $\|.\|_n$ le produit scalaire et la norme induits par la métrique Q_n. Nous établissons dans (Monnez, 2020) un théorème de convergence presque sûre de processus $(X_n^i), i=1,2,...,r$, obtenus par une orthonormalisation de Gram-Schmidt à l'étape n par rapport à Q_{n+1}, vers $\pm V_i$ et $(\Lambda_n^i), i=1,2,...,r$, vers λ_i, tels que :

$$\widetilde{Y}_{n+1}^i = (I+a_nB_n)\widetilde{X}_n^i \tag{2}$$

$$\widetilde{X}_{n+1}^i = \widetilde{Y}_{n+1}^i - \sum_{j<i}\left\langle \widetilde{Y}_{n+1}^j, X_{n+1}^j\right\rangle_{n+1} X_{n+1}^j, \quad X_{n+1}^i = \frac{\widetilde{X}_{n+1}^i}{\left\|\widetilde{X}_{n+1}^i\right\|_{n+1}} \tag{3}$$

$$\Lambda_{n+1}^i = (1-a_n)\Lambda_n^i + a_n\left\langle B_nX_n^i, X_n^i\right\rangle_n. \tag{4}$$

Ce théorème peut être utilisé dans des cas où $E[B_n \mid T_n]$ converge p.s. vers B ou B_n converge p.s. vers B.
Il étend ou complète ceux donnés par Benzécri (1969), Oja et Karhunen (1985), Duflo (1997) et Brandière (1998) pour l'ACP, Monnez et Skiredj (2019). Nous en énonçons un corollaire dans le cas où $B_n - B = B_n^1 + B_n^2$ en supposant :

H1 (a) B est Q-symétrique. (b) Les valeurs propres de B sont simples.

H2 (a) $\sum_1^\infty a_n\|B_n^1\| < \infty$ p.s., $E[B_n^2 \mid T_n] = 0$ p.s., $E\left[\sup_n \|B_n^2\|^2\right] < \infty$.(b) Pour tout n, $I+a_nB_n$ est inversible.

H3 (a) $a_n > 0, \sum_1^\infty a_n = \infty, \sum_1^\infty a_n^2 < \infty$.

H3 (b) $a_n = \frac{a}{n^\alpha}$ avec $a>0, \frac{2}{3} < \alpha \leq 1$ et $a > \frac{1}{2}$ pour $\alpha = 1$ (une hypothèse plus générale peut être formulée).

H4 $Q_n \longrightarrow Q, \sum_1^\infty a_n\|Q_n - Q\| < \infty$ p.s.

H5 Pour $i=1,...,r, X_1^i$ est une variable aléatoire absolument continue indépendante de $B_1,...,B_n,....$

Théorème 1 *Sous les hypothèses H1a,b,2a,b,3b,4,5, on a presque sûrement pour* $i=1,...,r$ *:* $X_n^i \longrightarrow V_i$ ou $-V_i$, $\Lambda_n^i \longrightarrow \lambda_i$, $\sum_1^\infty a_n|\langle BX_n^i, X_n^i\rangle_n - \lambda_i| < \infty$. *Dans le cas où* $B_n^2 = 0$, *on a les mêmes conclusions en remplaçant H3b par H3a ; en outre* $\sum_1^\infty a_n|\langle B_nX_n^i, X_n^i\rangle_n - \lambda_i| < \infty$ *et* $\sum_1^\infty a_n|\Lambda_n^i - \lambda_i| < \infty$ *presque sûrement.*

3 L'analyse canonique généralisée

3.1 Formulation probabiliste

Supposons que l'ensemble des composantes d'un vecteur aléatoire Z dans $\mathbb{R}^p$ soit partitionné en q sous-ensembles de variables aléatoires réelles $\left\{Z^{k1},...,Z^{kr_k}\right\}$, $k=1,...,q$. Soit Z^k le vecteur aléatoire dans $\mathbb{R}^{r_k}$ dont les composantes sont $Z^{k1},...Z^{kr_k}$. Soit C^k la matrice de covariance de Z^k, C celle de Z. Supposons qu'il n'y ait pas de relation affine entre les composantes de Z, ainsi $C^k, k=1,...,q$ et C sont inversibles.

Soit le problème suivant : pour $l=1,...,r\leq p$, déterminer au pas l une combinaison linéaire de toutes les composantes centrées de Z, $U_l=\theta_l'\left(Z-E\left[Z\right]\right)$, appelée $l^{ième}$ composante générale, de variance 1 et non corrélée à $U_1,...,U_{l-1}$, et, pour $k=1,...,q$, une combinaison linéaire de variance 1 des composantes centrées de Z^k, $V_l^k=\left(\eta_l^k\right)'\left(Z^k-E\left[Z^k\right]\right)$, appelée $l^{ème}$ composante canonique du $k^{ième}$ sous-ensemble de variables, qui maximisent $\sum_{k=1}^q\rho^2\left(U_l,V_l^k\right)$, ρ désignant le coefficient de corrélation linéaire.

Soit M la matrice inconnue d'ordre p diagonale par blocs dont le $k^{ième}$ bloc diagonal est $M^k=\left(C^k\right)^{-1}$. Soit $\theta_l=\left(\left(\theta_l^1\right)'...\left(\theta_l^q\right)'\right)', \theta_l^k\in\mathbb{R}^{m_k}, k=1,...,q$. On montre que θ_l est un vecteur propre C-normé de la matrice M^{-1}-symétrique $B=MC$ correspondant à sa $l^{ième}$ plus grande valeur propre λ_l et que pour $k=1,...,q$, il existe $\alpha_l^k\in\mathbb{R}$ tel que $\eta_l^k=\alpha_l^k\theta_l^k$. Remarquons que $\sqrt{\lambda_l}\left(\theta_l\right)'\left(Z-E\left[Z\right]\right)$ est la $l^{ième}$ composante principale de l'ACP de Z avec la métrique M. L'objectif est donc de réaliser l'ACP en ligne de Z en utilisant à l'étape n un estimateur convergent M_n de M.

Dans le cas où pour tout k, $m_k=1$, cette analyse est équivalente à l'ACP normée.

Dans le cas $q=2$, cette analyse est équivalente à l'analyse canonique de deux ensembles de variables, qui a pour cas particuliers l'analyse factorielle discriminante et l'analyse factorielle des correspondances. Pour $k=1,2$, $\left(\theta_l^k\right)'\left(Z^k-E\left[Z^k\right]\right)$ est colinéaire à la $l^{ième}$ composante canonique du $k^{ième}$ ensemble de variables.

3.2 Approximation stochastique dans le cas d'un flux

Soit $(Z_{11},...,Z_{1m_1},...,Z_{n1},...,Z_{nm_n},...)$ un échantillon i.i.d. de Z avec $Z_{ij}=\left(Z_{ij}^1,...,Z_{ij}^q\right)$; $Z_{n1},...,Z_{nm_n}$ sont observés à l'étape n du processus. On note T_n la tribu du passé à l'étape n, par rapport à laquelle $Z_{11,}...,Z_{n-1,m_{n-1}}$ sont mesurables. On note $\overline{Z}_n$ la moyenne des Z_i observés jusqu'à l'étape n, et $\overline{Z}_n^k$ celle des Z_i^k pour $k=1,...,q$. On note C_n la matrice de covariance des Z_i observés jusqu'à l'étape n, et C_n^k celle des Z_i^k pour $k=1,...,q$, qui peuvent être calculées de façon récursive :

$$C_n^k=\frac{1}{\mu_n}\sum_{i=1}^n\sum_{j=1}^{m_i}Z_{ij}^kZ_{ij}^{k\prime}-\overline{Z}_n^k\overline{Z}_n^{k'},\mu_n=\sum_{i=1}^n m_i. \tag{5}$$

$M^k = \left(C^k\right)^{-1}$ est solution de l'équation en X : $C^k X - I = 0$ ou

$$E\left[\left(Z^k Z^{k\prime} - E\left[Z^k\right] E\left[Z^k\right]'\right) X - I\right] = 0. \tag{6}$$

Pour estimer $M^k = \left(C^k\right)^{-1}$, on définit le processus d'approximation stochastique $\left(M_n^k, n \geqslant 1\right)$ tel que :

$$M_n^k = M_{n-1}^k - a_n \left(C_n^k M_{n-1}^k - I\right). \tag{7}$$

On fait les hypothèses :

H6 (a) Il n'y a pas de relation affine entre les composantes de Z ; (b) Z admet des moments d'ordre 4.

H3 (c) $\sum_1^\infty \frac{a_n}{\sqrt{n}} < \infty$.

Théorème 2 *Sous les hypothèses H6a,b,3a,c, on a presque sûrement :* $M_n^k \longrightarrow \left(C^k\right)^{-1}$, $\sum_{n=1}^\infty a_n \left\| M_n^k - \left(C^k\right)^{-1} \right\| < \infty$.

Soit M_n et N_n les matrices diagonales par bloc dont les $k^{ièmes}$ blocs diagonaux sont respectivement M_n^k et $C_n^k, k = 1, ...q$. N_n est symétrique positive de plein rang à partir d'un certain n que l'on suppose égal à 1. Soit :

$$B_n = M_n \left(\omega_{1n} C_n + \omega_{2n} \left(\frac{1}{m_n} \sum_{j=1}^{m_n} Z_{nj} Z_{nj}' - \overline{Z}_n \overline{Z}_n'\right)\right), \omega_{1n} \geqslant 0, \omega_{2n} \geqslant 0, \omega_{1n} + \omega_{2n} = 1. \tag{8}$$

B_n est la somme de deux termes, le premier correspondant à l'utilisation de toutes les observations jusqu'à l'étape n et le deuxième à celle d'un lot de m_n observations entrées à cette étape, pondérés par un poids compris entre 0 et 1 et dépendant de n. On peut aussi définir B_n de telle manière qu'à partir d'une étape N, on ne tienne plus compte des observations effectuées avant cette étape. On définit les processus $(X_n^i, n \geqslant 1)$ et $(\Lambda_n^i, n \geqslant 1)$ comme précédemment en faisant l'orthonormalisation de Gram-Schmidt à l'étape n par rapport à N_n (on a $Q_{n+1} = N_n$). On fait l'hypothèse :

H6 (c) Z est p.s. bornée.

Théorème 3 *Sous les hypothèses H1b,3b,c,5,6a,c, on a presque sûrement pour* $i = 1, ..., r$ *:* $X_n^i \longrightarrow V_i$ ou $-V_i$, $\Lambda_n^i \longrightarrow \lambda_i$, $\sum_1^\infty a_n \left|\langle B X_n^i, X_n^i\rangle_n - \lambda_i\right| < \infty$. *Dans le cas où* $\omega_{2n} = 0$, *on a les mêmes conclusions en remplaçant H3b par H3a et H6c par H6b ; on a en outre* $\sum_1^\infty a_n \left|\langle B_n X_n^i, X_n^i\rangle_n - \lambda_i\right| < \infty$ *et* $\sum_1^\infty a_n \left|\Lambda_n^i - \lambda_i\right| < \infty$ *presque sûrement.*

4 Conclusion

Nous avons présenté un corollaire et une application du théorème général de convergence presque sûre du processus de Oja établi dans (Monnez, 2020) à l'analyse canonique généralisée en ligne d'un vecteur aléatoire en estimant par un processus d'approximation

stochastique la métrique inconnue et en utilisant à chaque étape des processus définis toutes les observations du vecteur aléatoire effectuées jusqu'à cette étape. On peut aussi utiliser à chaque étape seulement les nouvelles observations entrantes à cette étape ($\omega_{1n} = 0$). Les expériences effectuées dans le cas de l'ACP en ligne dans les cas $\omega_{1n} = 0$ (utilisation d'un lot de nouvelles observations à l'étape courante) ou $\omega_{2n} = 0$ (utilisation de toutes les observations jusqu'à l'étape courante) ont montré en général une plus grande rapidité de convergence des processus avec $\omega_{2n} = 0$ (Monnez et Skiredj, 2020).

5 Bibliographie

Balsubramani, A., Dasgupta S. et Freund, Y. (2013), The fast convergence of incremental PCA, *NIPS*, pp. 3174-3182.

Benzécri, J.P. (1969), Approximation stochastique dans une algèbre normé non commutative, *Bull. Soc. Math. France,* 97, pp. 225-241.

Brandière, O. (1998), Some pathological traps for stochastic approximation, *Siam J. Control Optim.,* 36, No. 4, pp. 1293-1314.

Cardot, H., Cénac, P. et Monnez, J.M. (2012), A fast and recursive algorithm for clustering large datasets with k-medians, *Computational Statistics and Data Analysis*, 56, pp. 1434-1449.

Duarte, K., Monnez, J.M. et Albuisson, E. (2018), Sequential linear regression with online standardized data, *PLoS ONE*, 13 (1) : e0191186.

Duflo, M. (1997), *Random Iterative Models*, Applications in Mathematics, 34, Springer-Verlag, Berlin.

Lalloué, B., Monnez, J.M. et Albuisson, E. (2019), Streaming constrained binary logistic regression with online standardized data, *Soumis*.

Monnez, J.M. (2020), Stochastic approximation of eigenvectors and eigenvalues of the Q-symmetric expectation of a random matrix, *Pré-publication*.

Monnez, J.M. et Skiredj, A. (2019), Convergence of a normed eigenvector stochastic approximation process and application to online principal component analysis of a data stream, *hal-01844419.*

Monnez, J.M. et Skiredj, A. (2020), Widening the scope of an eigenvector stochastic approximation process and application to streaming PCA and related methods, *Soumis.*

Oja, E. et Karhunen, J. (1985), On stochastic approximation of the eigenvectors and eigenvalues of the expectation of a random matrix, *Journal of Mathematical Analysis and Applications,* 106, pp. 69-84.