Convergence du processus de Oja et ACP en ligne

Jean-Marie Monnez

Convergence du processus de Oja et acp 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. Le processus de Oja est couramment utilisé pour estimer séquentiellement un vecteur propre associé à la plus grande valeur propre de l'espérance mathématique d'une matrice aléatoire symétrique en en utilisant un échantillon i.i.d., puis des vecteurs propres associés aux valeurs propres suivantes en ordre décroissant. Nous proposons deux extensions des hypothèses de convergence presque sûre de ce processus. Dans l'ACP en ligne d'un flux de données, ces extensions permettent de traiter des cas où la métrique utilisée est inconnue et est estimée en ligne, et également d'établir la convergence d'un processus où, au lieu d'utiliser plusieurs observations à chaque étape, on utilise toutes les observations jusqu'à l'étape courante, donc toute l'information contenue dans les données précédentes, sans avoir à les stocker.

Mots-clés. Algorithmes stochastiques, Analyse en composantes principales, Estimation en ligne, Flux de données, Vecteurs propres.

Abstract. Using an i.i.d. sample of a random matrix, an eigenvector corresponding to its largest eigenvalue can be sequentially estimated by the currently used process of Oja, then eigenvectors corresponding to its eigenvalues in decreasing order. We propose two extensions of the almost sure convergence assumptions of this process. Applying these extensions to online PCA of a data stream, this allows to study the case where the metrics used is unknown and is estimated online and also to prove the convergence of a process using, instead a batch of observations at each step, all observations until the current step, thus the information contained in the previous data without storing them.

Keywords. Data stream, Eigenvectors, Online estimation, Principal component analysis, Stochastic algorithms.

1 Théorèmes de convergence du processus de Oja

Soit B une matrice (p,p) symétrique de vecteurs propres normés $V_1,...,V_p$ associés aux valeurs propres $\lambda_1 > ... > \lambda_p$. $\|x\|$ désigne la norme euclidienne usuelle d'un vecteur x

de $\mathbb{R}^p$, la norme matricielle est la norme spectrale. Soit (a_n) une suite de nombres réels vérifiant l'hypothèse classique

H1 (a) $a_n > 0, \sum_{n=1}^{\infty} a_n = \infty, \sum_{n=1}^{\infty} a_n^2 < \infty$.

Soit le processus déterministe normé $(x_n, n \geqslant 1)$ dans $\mathbb{R}^p$ tel que : $x_{n+1} = \frac{(I+a_nB)x_n}{\|(I+a_nB)x_n\|}$. On établit que (x_n) converge vers V_1 ou $-V_1$ à condition que x_1 ne soit pas orthogonal à V_1.

Supposons maintenant que B soit l'espérance mathématique d'une matrice aléatoire symétrique A supposée bornée p.s. dont on peut observer un échantillon i.i.d. $(A_1, ..., A_n, ...)$ et que B soit inconnue. Soit le processus normé $(X_n, n \geqslant 1)$ défini par Oja et Karhunen (1985) tel que :

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

On établit la convergence preque sûre de ce processus. Sa rapidité de convergence est étudiée dans Balsubramani et al (2013). T_n étant la tribu du passé à l'étape n, on a donc $E[A_n \mid T_n] = B$.

Dans le cas de l'ACP d'un vecteur aléatoire Z présentée dans le paragraphe suivant, $\mathbb{R}^p$ étant muni d'une métrique M qui peut dépendre de caractéristiques de Z, comme la métrique diagonale des inverses des variances des composantes de Z en ACP normée, pour déterminer les composantes principales on recherche les premiers vecteurs propres de la matrice $B = ME\left[(Z - E[Z])(Z - E[Z])'\right]$. La matrice B est Q-symétrique i.e. $(QB)' = QB$, avec $Q = M^{-1}$. Dans le cas d'un flux de données $(z_1, ..., z_n, ...)$, $E[Z]$ et M ne sont pas connues a priori, mais peuvent être estimées en ligne au fur et à mesure de l'arrivée de nouvelles données.

De façon générale, considérons une matrice Q-symétrique B, notons Q_n une métrique T_n-mesurable estimateur de la métrique Q, $\langle ., . \rangle_n$ et $\|.\|_n$ le produit scalaire et la norme induits par la métrique Q_n. Considérons le processus $(X_n, n \geqslant 1)$ normé par rapport à Q_n et le processus réel (Λ_n) tels que :

$$X_{n+1} = \frac{(I + a_n B_n) X_n}{\|(I + a_n B_n) X_n\|_n}, \tag{2}$$

$$\Lambda_{n+1} = (1 - a_n)\Lambda_n + a_n \langle B_n X_n, X_n \rangle_n, \tag{3}$$

$(B_1, ..., B_n, ...)$ n'étant plus de façon générale un échantillon i.i.d. d'une matrice aléatoire. Monnez et Skiredj (2018) établissent des résultats de convergence p.s. de (X_n) dans le cas $Q_n = Q$.

Nous donnons ici un théorème de convergence p.s. de ces processus sous :

H1 (b) $a_n = \frac{c}{n^\alpha}, c \geqslant 1, \frac{3}{4} < \alpha \leq 1$.

H2 (a) B est Q-symétrique, (b) B a p valeurs propres distinctes $\lambda_1 > ... > \lambda_p$, (c) $\|B\| = \lambda_1$.

H3 (a) $\sum_1^\infty a_n \|E[B_n \mid T_n] - B\| < \infty$ p.s., (b) Pour tout n, $I + a_n B_n$ est p.s. inversible, (c) $E\left[\sup_n \|B_n\|^2\right] < \infty$.

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

H5 (a) X_1 est absolument continu, indépendant de $B_1, ..., B_n, ...$

Théorème 1 *Sous les hypothèses H1b, H2a,b,c, H3a,b,c, H4a, H5a, X_n converge p.s. vers V_1 ou $-V_1$, $\sum_1^\infty a_n |\lambda_1 - \langle BX_n, X_n\rangle_n| < \infty$ p.s. et Λ_n converge p.s. vers λ_1.*

L'hypothèse "Les B_n constituent un échantillon i.i.d. d'une matrice aléatoire" est remplacée par l'hypothèse H3a, qui pourra être utilisée dans des cas où $E[B_n \mid T_n]$ converge p.s. vers B. L'hypothèse H1b peut être remplacée par une hypothèse plus générale non énoncée ici. On remarque que H2c est vérifiée si les valeurs propres de B sont positives ou nulles, H3b si celles de B_n le sont. Monnez et Skiredj (2018) établissent un théorème de convergence p.s. de (X_n) sous H1a mais sous des hypothèses H3a,c plus restrictives.

Nous donnons un deuxième théorème en remplaçant H3a par :

H3 (a') $\sum_1^\infty a_n \|B_n - B\| < \infty$ p.s.

Théorème 2 *Sous les hypothèses H1a, H2a,b,c, H3a',b, H4a, H5a, X_n converge p.s. vers V_1 ou $-V_1$, $\sum_1^\infty a_n |\lambda_1 - \langle BX_n, X_n\rangle_n| < \infty$ p.s. et Λ_n converge p.s. vers λ_1.*

Ce théorème pourra être utilisé dans des cas où la suite de matrices aléatoires (B_n) converge p.s. vers B. On pourra alors tenir compte à chaque étape de l'information contenue dans les données jusqu'à l'étape courante, comme on le verra dans l'application à l'ACP en ligne.

Pour estimer les vecteurs propres de B associés aux valeurs propres par ordre décroissant jusqu'au rang r, on définit les processus $(Y_n^i), (T_n^i), (X_n^i), (\Lambda_n^i)$ pour $i = 2, ..., r$:

$$\begin{aligned} Y_{n+1}^i &= (I + a_n B_n) X_n^i \\ T_{n+1}^i &= Y_{n+1}^i - \sum_{j<i} \left\langle Y_{n+1}^i, X_{n+1}^j \right\rangle_n X_{n+1}^j, \; X_{n+1}^i = \frac{T_{n+1}^i}{\left\| T_{n+1}^i \right\|_n} \qquad (4) \\ \Lambda_{n+1}^i &= (1 - a_n) \Lambda_n^i + a_n \left\langle B_n X_n^i, X_n^i \right\rangle_n . \qquad (5) \end{aligned}$$

Pour $i = 1, ..., r$, $\left(X_{n+1}^1, ..., X_{n+1}^r\right)$ est obtenu à partir de $\left(Y_{n+1}^1, ..., Y_{n+1}^r\right)$ en faisant une orthonormalisation au sens de Gram-Schmidt par rapport à la métrique Q_n.

Soit $(e_1, ..., e_p)$ une base de $\mathbb{R}^p$. Pour $i \leq p$, on note ${}^i\Lambda\mathbb{R}^p$ la puissance extérieure d'ordre i de $\mathbb{R}^p$, engendrée par les C_p^i produits extérieurs $e_{j_1} \wedge e_{j_2} \wedge ... \wedge e_{j_i}$, $j_1 < j_2 < ... < j_i \in \{1, ..., p\}$. On note ${}^{i1}B$ l'endomorphisme dans ${}^i\Lambda\mathbb{R}^p$ tel que :

$$ {}^{i1}B(x_1 \wedge ... \wedge x_i) = \sum_{j=1}^{i} x_1 \wedge ... \wedge Bx_j \wedge ... \wedge x_i.$$

Pour $1 \leqslant j_1 < ... < j_i \leqslant p$, $V_{j_1} \wedge ... \wedge V_{j_i}$ est vecteur propre de ${}^{i1}B$ associé à la valeur propre $\sum_{l=1}^{i} \lambda_{j_l}$. $V_1 \wedge ... \wedge V_i$ est associé à la plus grande valeur propre $\sum_{j=1}^{i} \lambda_j$.

On suppose :

H2 (d) Pour $i = 1, ..., r$, $\|{}^{i1}B\| = \sum_{j=1}^{i} \lambda_j$.

H5 (b) Pour $i = 1, ..., r$, X_1^i est absolument continu, indépendant de $B_1, ..., B_n, ...$

En appliquant le théorème 1, nous démontrons pour $i = 1, ..., r$ la convergence p.s. du processus $({}^iX_n) = (X_n^1 \wedge ... \wedge X_n^i)$ dans ${}^i\Lambda\mathbb{R}^p$ vers $V^1 \wedge ... \wedge V^i$ et en déduisons le théorème suivant :

Théorème 3 *Sous les mêmes hypothèses que dans le théorème 1 ou le théorème 2, sauf H2c et H5a remplacées respectivement par H2d et H5b, pour $i = 1, ..., r$, X_n^i converge p.s. vers V_i* ou $-V_i$, $\sum_1^\infty a_n \left|\lambda_i - \langle BX_n^i, X_n^i\rangle_n\right| < \infty$ *p.s. et Λ_n^i converge p.s. vers λ_i.*

2 Application à l'ACP en ligne

On peut trouver une présentation de différentes méthodes d'ACP en ligne dans Cardot et Degras (2018).

Soit M une métrique dans $\mathbb{R}^p$. On définit dans le dual de $\mathbb{R}^p$ la métrique $Q = M^{-1}$. Dans l'ACP d'un vecteur aléatoire Z dans $\mathbb{R}^p$, la $l^{ième}$ composante principale est une combinaison linéaire des composantes centrées de Z, $U_l = (\theta_l)' (Z - E[Z])$, non corrélée à $U_1, ..., U_{l-1}$, de variance maximale sous la contrainte $(\theta_l)' M^{-1}\theta_l = 1$. θ_l est vecteur propre de la matrice M^{-1}-symétrique $B = MCovar[Z] = ME\left[(Z - E[Z])(Z - E[Z])'\right] = M(E[ZZ'] - E[Z]E[Z'])$ associé à la $l^{ième}$ plus grande valeur propre λ_l. Dans le cas d'un flux de données, les moments de Z et la métrique M lorsqu'elle dépend de caractéristiques de Z sont a priori inconnus.

Soit $(Z_{11}, ..., Z_{1m_1}, ..., Z_{n1}, ..., Z_{nm_n}, ...)$ un échantillon i.i.d. de Z. On suppose que $Z_{n1}, ..., Z_{nm_n}$ sont observés à l'étape n. 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 M_{n-1} une matrice symétrique

définie positive T_n-mesurable, estimateur de M fonction des Z_i utilisés jusqu'à l'étape $n-1$, et $\overline{Z}_{n-1}$ la moyenne de ces Z_i. Par exemple, si M est la métrique diagonale des inverses des variances des composantes de Z, M_{n-1} est la métrique diagonale des inverses des estimateurs de ces variances, calculés à partir de tous les Z_i utilisés jusqu'à l'étape $n-1$. On définit la métrique T_n-mesurable $Q_n = M_{n-1}^{-1}$. Soit la matrice (p, m_n) Z_n^c qui a pour colonnes $Z_{n1} - \overline{Z}_{n-1}, ..., Z_{nm_n} - \overline{Z}_{n-1}$.

Un premier processus est obtenu en prenant :

$$B_n = M_{n-1} Z_n^c \left(Z_n^c\right)'. \tag{6}$$

On utilise pour déterminer M_{n-1} un processus récursif sans qu'il soit nécessaire de stocker les données. On n'a pas $E\left[B_n \mid T_n\right] = B$, mais les hypothèses H4b et H6a ci-dessous impliquent $E\left[B_n \mid T_n\right] \longrightarrow B$ p.s.

H4 (b) $M_n \longrightarrow M, \sum_1^\infty a_n \left\|M_{n-1} - M\right\| < \infty \; p.s.$

H6 (a) $\|Z\|$ est bornée p.s.

Théorème 4 *Sous les hypothèses H1b, H2b, H4b, H5b, H6a, on a les conclusions du théorème 3.*

Soit maintenant les matrices (p, m_n) Z_n^{1c}, qui a pour colonnes $Z_{n1} - \overline{Z}_n, ..., Z_{nm_n} - \overline{Z}_n$, et $Z_n = (Z_{n1} ... Z_{nm_n})$. Un deuxième processus est obtenu en prenant :

$$B_n = M_n \frac{1}{\sum_1^n m_i} \sum_{i=1}^n Z_i^{1c} \left(Z_i^{1c}\right)' = M_n \left(\frac{1}{\sum_1^n m_i} \sum_{i=1}^n Z_i \left(Z_i\right)' - \overline{Z}_n \overline{Z}_n' \right). \tag{7}$$

.

On constate qu'il n'est pas nécessaire de stocker les données des étapes précédentes. On utilise l'information contenue dans les données précédentes, comme sous une autre forme dans l'algorithme History PCA de Yang et al (2018). On n'a pas $E\left[B_n \mid T_n\right] = B$, mais les hypothèses H6b ci-dessous et H4b impliquent $B_n \longrightarrow B$ p.s.

H6 (b) Les moments d'ordre 4 de Z existent.

Théorème 5 *Sous les hypothèses H1a, H2b, H4b, H5b, H6b, on a les conclusions du théorème 3.*

Ces processus peuvent être utilisés par exemple pour estimer en ligne les composantes générales d'une analyse canonique généralisée (ACG, gCCA) : le vecteur aléatoire Z est

composé de q sous-vecteurs $Z^1, ..., Z^q$; les composantes générales de l'ACG peuvent être considérées comme les composantes principales d'une ACP avec une métrique M diagonale par blocs, le $k^{ième}$ bloc diagonal étant l'inverse de la matrice de covariance de Z^k. On définit un processus d'approximation stochastique de chacune de ces inverses, on en déduit un processus d'approximation (M_n) de M qui doit vérifier l'hypothèse H4b, puis un processus d'estimation en ligne des composantes principales comme précédemment.

3 Conclusion

La convergence presque sûre du processus de Oja a été établie sous des hypothèses plus générales sur la suite (B_n), deux théorèmes utilisables respectivement dans des cas où l'on utilise plusieurs observations à chaque étape d'un processus ou toutes les observations jusqu'à l'étape courante ont été établis et appliqués à l'analyse en composantes principales d'un vecteur aléatoire Z, permettant l'estimation en ligne de moments de Z et de la métrique. Ces résultats étendent ou complètent ceux donnés par Benzécri (1969), Oja et Karhunen (1985), Monnez (1994), Duflo (1997) et Brandière (1998) pour l'ACP, Monnez et Skiredj (2018).

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. et Degras, D. (2018), Online principal component analysis in high dimension: which algorithm to choose?, *International Statistical Review,* 86, 1, pp. 29-50.

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

Monnez, J.M. (1994), Convergence d'un processus d'approximation stochastique en analyse factorielle, *Pub. Inst. Stat. Univ. Paris,* 38, 1, pp. 37-56.

Monnez, J.M. et Skiredj, A. (2018), Convergence of a normed eigenvector stochastic approximation process and application to online principal component analysis of a data stream, *hal-01844419.*

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.

Yang, P., Hsieh, C.J., Wang, J.L. (2018), History PCA: a new algorithm for streaming PCA, *arXiv:1802.05447v1.*