Analyse en composantes principales d'un flux de données d'espérance variable dans le temps

Jean-Marie Monnez

Institut Elie Cartan UMR 7502 – Laboratoire de Mathématiques
Nancy-Université, CNRS, INRIA
BP 239 – F 54506 – Vandoeuvre-lès-Nancy Cedex
jean-marie.monnez@iecn.u-nancy.fr

Résumé. On considère un flux de données représenté par une suite de vecteurs de données. On suppose que chaque vecteur de données est une réalisation d'un vecteur aléatoire dont l'espérance mathématique varie dans le temps selon un modèle linéaire pour chacune des composantes. On utilise des processus d'approximation stochastique pour estimer en ligne les paramètres des modèles linéaires et en même temps les facteurs de l'ACP du vecteur aléatoire.

1 Modèle de flux et plan d'étude

Soit un flux de données, représenté par une suite de vecteurs $(z_1, \ldots, z_n, \ldots)$ dans $\mathbb{R}^p$.

1.1 Modèle d'étude et formulation

On suppose que :

- pour tout n, z_n est la réalisation d'un vecteur aléatoire Z_n, d'espérance mathématique variable dans le temps ;
- les vecteurs aléatoires Z_n sont mutuellement indépendants ;
- pour tout n, on a la décomposition $Z_n = \theta_n + R_n$, $\theta_n = (\theta_n^1 \ldots \theta_n^p)'$ étant un vecteur de $\mathbb{R}^p$, la loi du vecteur aléatoire R_n ne dépendant pas de n, $E[R_n] = 0, Covar[R_n] = \Sigma$ (matrice de covariance de R_n) ; ceci revient à supposer que les $R_n = Z_n - \theta_n$ constituent un échantillon i.i.d. d'un vecteur aléatoire R dans $\mathbb{R}^p$ tel que $E[R] = 0, Covar[R] = \Sigma$; on a alors $E[Z_n] = \theta_n, Covar[Z_n] = \Sigma$; $r_n = z_n - E[Z_n]$ représente la donnée z_n centrée ;
- pour $i = 1, \ldots, p$, il existe un vecteur β^i de $\mathbb{R}^{n_i}$ inconnu et , pour tout n, un vecteur U_n^i de $\mathbb{R}^{n_i}$ connu au temps n tels que $\theta_n^i = \langle \beta^i, U_n^i \rangle$, $\langle ., . \rangle$ dé signant le produit scalaire euclidien usuel dans $\mathbb{R}^{n_i}$; U_n^i peut être un vecteur de fonctions connues du temps (θ_n^i est alors une combinaison linéaire de fonctions connues du temps) ou un vecteur de valeurs de variables explicatives contrôlées ; si l'on note Z_n^i, respectivement R_n^i, la $i^{ème}$ composante de Z_n, respectivement R_n, on a alors le modèle de régression linéaire

$$Z_n^i = \langle \beta^i, U_n^i \rangle + R_n^i, \quad i = 1, \ldots, p.$$