Sequential linear regression with online standardized data

Kévin Duarte, Jean-Marie Monnez, Eliane Albuisson

Sequential linear regression with online standardized data

Kévin Duarte[1,2,3*], Jean-Marie Monnez[1,2,3,4], Eliane Albuisson[1,5,6]

1 Université de Lorraine, Institut Elie Cartan de Lorraine, UMR 7502, Vandoeuvre-lès-Nancy, F-54506, France
2 Project-Team BIGS, INRIA, Villers-lès-Nancy, F-54600, France
3 INSERM U1116, Centre d'Investigations Cliniques-Plurithématique 1433, Université de Lorraine, Nancy, France
4 Université de Lorraine, Institut Universitaire de Technologie Nancy-Charlemagne, Nancy, F-54052, France
5 BIOBASE, Pôle S^2R, CHRU de Nancy, Vandoeuvre-lès-Nancy, France
6 Faculté de Médecine, InSciDenS, Vandoeuvre-lès-Nancy, France

* k.duarte@chru-nancy.fr

Abstract

The present study addresses the problem of sequential least square multidimensional linear regression, particularly in the case of a data stream, using a stochastic approximation process. To avoid the phenomenon of numerical explosion which can be encountered and to reduce the computing time in order to take into account a maximum of arriving data, we propose using a process with online standardized data instead of raw data and the use of several observations per step or all observations until the current step. Herein, we define and study the almost sure convergence of three processes with online standardized data: a classical process with a variable step-size and use of a varying number of observations per step, an averaged process with a constant step-size and use of a varying number of observations per step, and a process with a variable or constant step-size and use of all observations until the current step. Their convergence is obtained under more general assumptions than classical ones. These processes are compared to classical processes on 11 datasets for a fixed total number of observations used and thereafter for a fixed processing time. Analyses indicate that the third-defined process typically yields the best results.

1 Introduction

In the present analysis, A' denotes the transposed matrix of A while the abbreviation "a.s." signifies almost surely.

Let $R = \left(R^1, ..., R^p\right)$ and $S = \left(S^1, ..., S^q\right)$ be random vectors in $\mathbb{R}^p$ and $\mathbb{R}^q$ respectively. Considering the least square multidimensional linear regression of S with respect to R: the (p, q) matrix θ and the $(q, 1)$ matrix η are estimated such that $E\left[\left\|S - \theta' R - \eta\right\|^2\right]$ is minimal.

Denote the covariance matrices

$$\begin{aligned} B &= Covar\left[R\right] = E\left[\left(R - E\left[R\right]\right)\left(R - E\left[R\right]\right)'\right], \\ F &= Covar\left[R, S\right] = E\left[\left(R - E\left[R\right]\right)\left(S - E\left[S\right]\right)'\right]. \end{aligned}$$

If we assume B is positive definite, i.e. there is no affine relation between the components of R, then

$$\theta = B^{-1}F, \eta = E[S] - \theta' E[R].$$

Note that, R_1 denoting the random vector in $\mathbb{R}^{p+1}$ such that $R_1' = \begin{pmatrix} R' & 1 \end{pmatrix}$, θ_1 the $(p+1, q)$ matrix such that $\theta_1' = \begin{pmatrix} \theta' & \eta \end{pmatrix}$, $B_1 = E[R_1 R_1']$ and $F_1 = E[R_1 S']$, we obtain $\theta_1 = B_1^{-1} F_1$.

In order to estimate θ (or θ_1), a stochastic approximation process (X_n) in $\mathbb{R}^{p \times q}$ (or $\mathbb{R}^{(p+1) \times q}$) is recursively defined such that

$$X_{n+1} = X_n - a_n (B_n X_n - F_n),$$

where (a_n) is a sequence of positive real numbers, eventually constant, called step-sizes (or gains). Matrices B_n and F_n have the same dimensions as B and F, respectively. The convergence of (X_n) towards θ is studied under appropriate definitions and assumptions on B_n and F_n.

Suppose that $((R_{1n}, S_n), n \geq 1)$ is an i.i.d. sample of (R_1, S). In the case where $q = 1$, $B_n = R_{1n} R_{1n}'$ and $F_n = R_{1n} S_n'$, several studies have been devoted to this stochastic gradient process (see for example Monnez [1], Ljung [2] and references hereafter). In order to accelerate general stochastic approximation procedures, Polyak [3] and Polyak and Juditsky [4] introduced the averaging technique. In the case of linear regression, Györfi and Walk [5] studied an averaged stochastic approximation process with a constant step-size. With the same type of process, Bach and Moulines [6] proved that the optimal convergence rate is achieved without strong convexity assumption on the loss function.

However, this type of process may be subject to the risk of numerical explosion when components of R or S exhibit great variances and may have very high values. For datasets used as test sets by Bach and Moulines [6], all sample points whose norm of R is fivefold greater than the average norm are removed. Moreover, generally only one observation of (R, S) is introduced at each step of the process. This may be not convenient for a large amount of data generated by a data stream for example.

Two modifications of this type of process are thus proposed in this article.

The first change in order to avoid numerical explosion is the use of standardized, i.e. of zero mean and unit variance, components of R and S. In fact, the expectation and the variance of the components are usually unknown and will be estimated online.

The parameter θ can be computed from the standardized components as follows. Let σ^j the standard deviation of R^j for $j = 1, ..., p$ and σ_1^k the standard deviation of S^k for $k = 1, ..., q$. Define the following matrices

$$\Gamma = \begin{pmatrix} \frac{1}{\sigma^1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma^p} \end{pmatrix}, \Gamma^1 = \begin{pmatrix} \frac{1}{\sigma_1^1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma_1^q} \end{pmatrix}.$$

Let $S_c = \Gamma^1 (S - E[S])$ and $R_c = \Gamma (R - E[R])$. The least square linear regression of S_c with respect to R_c is achieved by estimating the (p, q) matrix θ_c such that $E\left[\left\| S_c - \theta_c' R_c \right\|^2\right]$ is minimal. Then $\theta_c = \Gamma^{-1} \left(B^{-1} F\right) \Gamma^1 \Leftrightarrow \theta = B^{-1} F = \Gamma \theta_c \left(\Gamma^1\right)^{-1}$.

The second change is to use, at each step of the process, several observations of (R, S) or an estimation of B and F computed recursively from all observations until the current step without storing them.

More precisely, the convergence of three processes with online standardized data is studied in sections 2, 3, 4 respectively.

First, in section 2, a process with a variable step-size a_n and use of several online standardized observations at each step is studied; note that the number of observations at each step may vary with n.

Secondly, in section 3, an averaged process with a constant step-size and use of a varying number of online standardized observations at each step is studied.

Thirdly, in section 4, a process with a constant or variable step-size and use of all online standardized observations until the current step to estimate B and F is studied.

These three processes are tested on several datasets when $q = 1$, S being a continuous or binary variable, and compared to existing processes in section 5. Note that when S is a binary variable, linear regression is equivalent to a linear discriminant analysis. It appears that the third-defined process most often yields the best results for the same number of observations used or for the same duration of computing time used.

These processes belong to the family of stochastic gradient processes and are adapted to data streams. Batch gradient and stochastic gradient methods are presented and compared in [7] and reviewed in [8], including noise reduction methods, like dynamic sample sizes methods, stochastic variance reduced gradient (also studied in [9]), second-order methods, ADAGRAD [10] and other methods. This work makes the following contributions to the variance reduction methods:

- In [9], the authors proposed a modification of the classical stochastic gradient algorithm to reduce directly the gradient of the function to be optimized in order to obtain a faster convergence. It is proposed in this article to reduce this gradient by an online standardization of the data.
- Gradient clipping [11] is another method to avoid a numerical explosion. The idea is to limit the norm of the gradient to a maximum number called threshold. This number must be chosen, a bad choice of threshold can affect the computing speed. Moreover it is then necessary to compare the norm of the gradient to this threshold at each step. In our approach the limitation of the gradient is implicitly obtained by online standardization of the data.
- If the expectation and the variance of the components of R and S were known, standardization of these variables could be made directly and convergence of the processes obtained using existing theorems. But these moments are unknown in the case of a data stream and are estimated online in this study. Thus the assumptions of the theorems of almost sure (a.s.) convergence of the processes studied in sections 2 and 3 and the corresponding proofs are more general than the classical ones in the linear regression case [1–5].
- The process defined in section 4 is not a classical batch method. Indeed in this type of method (gradient descent), the whole set of data is known a priori and is used at each step of the process. In the present study, new data are supposed to arrive at each step, as in a data stream, and are added to the preceding set of data, thus reducing by averaging the variance. This process can be considered as a dynamic batch method.
- A suitable choice of step-size is often crucial for obtaining good performance of a stochastic gradient process. If the step-size is too small, the convergence will be slower. Conversely, if the step-size is too large, a numerical explosion may occur during the first iterations. Following [6], a very simple choice of the step-size is proposed for the methods with a constant step-size.
- Another objective is to reduce computing time in order to take into account a maximum of data in the case of a data stream. It appears in the experiments that

the use of all observations until the current step without storing them, several observations being introduced at each step, increases at best in general the convergence speed of the process. Moreover this can reduce the influence of outliers.

As a whole the major contributions of this work are to reduce gradient variance by online standardization of the data or use of a "dynamic" batch process, to avoid numerical explosions, to reduce computing time and consequently to better adapt the stochastic approximation processes used to the case of a data stream.

2 Convergence of a process with a variable step-size

Let $(B_n, n \geq 1)$ and $(F_n, n \geq 1)$ be two sequences of random matrices in $\mathbb{R}^{p\times p}$ and $\mathbb{R}^{p\times q}$ respectively. In this section, the convergence of the process $(X_n, n \geq 1)$ in $\mathbb{R}^{p\times q}$ recursively defined by

$$X_{n+1} = X_n - a_n \left(B_n X_n - F_n\right)$$

and its application to sequential linear regression are studied.

2.1 Theorem

Let X_1 be a random variable in $\mathbb{R}^{p\times q}$ independent from the sequence of random variables $\left((B_n, F_n), n \geq 1\right)$ in $\mathbb{R}^{p\times p} \times \mathbb{R}^{p\times q}$.

Denote T_n the σ-field generated by X_1 and $(B_1, F_1), ..., (B_{n-1}, F_{n-1})$.
$X_1, X_2, ..., X_n$ are T_n-measurable.

Let (a_n) be a sequence of positive numbers.

Make the following assumptions:

(H1a) There exists a positive definite symmetrical matrix B such that a.s.

1) $\sum_{n=1}^{\infty} a_n \left\| E\left[B_n | T_n\right] - B \right\| < \infty$

2) $\sum_{n=1}^{\infty} a_n^2 E\left[\left\| B_n - B \right\|^2 | T_n\right] < \infty.$

(H2a) There exists a matrix F such that a.s.

1) $\sum_{n=1}^{\infty} a_n \left\| E\left[F_n | T_n\right] - F \right\| < \infty$

2) $\sum_{n=1}^{\infty} a_n^2 E\left[\left\| F_n - F \right\|^2 | T_n\right] < \infty.$

(H3a) $\sum_{n=1}^{\infty} a_n = \infty, \sum_{n=1}^{\infty} a_n^2 < \infty.$

Theorem 1 *Suppose H1a, H2a and H3a hold. Then X_n converges to $\theta = B^{-1}F$ a.s.*

State the Robbins-Siegmund lemma [12] used in the proof.

Lemma 2 *Let (Ω, A, P) be a probability space and (T_n) a non-decreasing sequence of sub-σ-fields of A. Suppose for all n, z_n, α_n, β_n and γ_n are four integrable non-negative T_n-measurable random variables defined on (Ω, A, P) such that:*

$$E\left[z_{n+1} | T_n\right] \quad \leq z_n \left(1 + \alpha_n\right) + \beta_n - \gamma_n \quad a.s.$$

Then, in the set $\left\{\sum_{n=1}^{\infty}\alpha_n<\infty,\sum_{n=1}^{\infty}\beta_n<\infty\right\}$, (z_n) *converges to a finite random variable and* $\sum_{n=1}^{\infty}\gamma_n<\infty$ *a.s.*

Proof of Theorem 1. The Frobenius norm $\|A\|$ for a matrix A is used. Recall that, if $\|A\|_2$ denotes the spectral norm of A, $\|AB\|\leq\|A\|_2\|B\|$.

$$\begin{aligned}X_{n+1}-\theta &= X_n-\theta-a_n\left(B_nX_n-F_n\right)\\ &= \left(I-a_nB\right)\left(X_n-\theta\right)-a_n\left(\left(B_n-B\right)X_n-\left(F_n-F\right)\right)\end{aligned}$$

Denote
$Z_n=\left(B_n-B\right)X_n-\left(F_n-F\right)=\left(B_n-B\right)\left(X_n-\theta\right)+\left(B_n-B\right)\theta-\left(F_n-F\right)$ and $X_n^1=X_n-\theta$. Then:

$$\begin{aligned}X_{n+1}^1 &= \left(I-a_nB\right)X_n^1-a_nZ_n\\ \left\|X_{n+1}^1\right\|^2 &= \left\|\left(I-a_nB\right)X_n^1\right\|^2-2a_n\langle\left(I-a_nB\right)X_n^1,Z_n\rangle+a_n^2\left\|Z_n\right\|^2.\end{aligned}$$

Denote λ the smallest eigenvalue of B. As $a_n\longrightarrow 0$, we have for n sufficiently large

$$\left\|I-a_nB\right\|_2=1-a_n\lambda<1.$$

Then, taking the conditional expectation with respect to T_n yields almost surely:

$$\begin{aligned}E\left[\left\|X_{n+1}^1\right\|^2|T_n\right] &\leq \left(1-a_n\lambda\right)^2\left\|X_n^1\right\|^2+2a_n|\langle\left(I-a_nB\right)X_n^1,E\left[Z_n|T_n\right]\rangle|+\\ &\quad a_n^2E\left[\left\|Z_n\right\|^2|T_n\right],\\ E\left[Z_n|T_n\right] &= \left(E\left[B_n|T_n\right]-B\right)X_n^1+\left(E\left[B_n|T_n\right]-B\right)\theta-\left(E\left[F_n|T_n\right]-F\right).\end{aligned}$$

Denoting

$$\begin{aligned}\beta_n &= \left\|E\left[B_n|T_n\right]-B\right\|,\delta_n=\left\|E\left[F_n|T_n\right]-F\right\|,\\ b_n &= E\left[\left\|B_n-B\right\|^2|T_n\right],d_n=E\left[\left\|F_n-F\right\|^2|T_n\right],\end{aligned}$$

we obtain, as $\left\|X_n^1\right\|\leq 1+\left\|X_n^1\right\|^2$:

$$\begin{aligned}\left|\langle\left(I-a_nB\right)X_n^1,E\left[Z_n|T_n\right]\rangle\right| &\leq \left\|X_n^1\right\|\left\|E\left[Z_n|T_n\right]\right\|\\ &\leq \left\|X_n^1\right\|^2\left(\beta_n\left(1+\|\theta\|\right)+\delta_n\right)+\beta_n\|\theta\|+\delta_n,\\ E\left[\left\|Z_n\right\|^2|T_n\right] &\leq 3b_n\left\|X_n^1\right\|^2+3b_n\|\theta\|^2+3d_n,\\ E\left[\left\|X_{n+1}^1\right\|^2|T_n\right] &\leq \left(1+a_n^2\lambda^2+2\left(1+\|\theta\|\right)a_n\beta_n+2a_n\delta_n+3a_n^2b_n\right)\left\|X_n^1\right\|^2+\\ &\quad 2\|\theta\|a_n\beta_n+2a_n\delta_n+3\|\theta\|^2a_n^2b_n+3a_n^2d_n-2a_n\lambda\left\|X_n^1\right\|^2.\end{aligned}$$

Applying Robbins-Siegmund lemma under assumptions H1a, H2a and H3a implies that there exists a non-negative random variable T such that a.s.

$$\left\|X_n^1\right\|\longrightarrow T,\sum_{n=1}^{\infty}a_n\left\|X_n^1\right\|^2<\infty.$$

As $\sum_{n=1}^{\infty}a_n=\infty$, $T=0$ a.s. ■

A particular case with the following assumptions is now studied.

(H1a') There exist a positive definite symmetrical matrix B and a positive real number b such that a.s.

1) for all n, $E\left[B_n|T_n\right] = B$

2) $\sup_n E\left[\left\|B_n - B\right\|^2 |T_n\right] < b$.

(H2a') There exist a matrix F and a positive real number d such that a.s.

1) for all n, $E\left[F_n|T_n\right] = F$

2) $\sup_n E\left[\left\|F_n - F\right\|^2 |T_n\right] < d$.

(H3a') Denoting λ the smallest eigenvalue of B,

$\left(a_n = \frac{a}{n^\alpha}, a > 0, \frac{1}{2} < \alpha < 1\right)$ or $\left(a_n = \frac{a}{n}, a > \frac{1}{2\lambda}\right)$.

Theorem 3 *Suppose H1a', H2a' and H3a' hold. Then X_n converges to θ almost surely and in quadratic mean. Moreover $\overline{\lim} \frac{1}{a_n} E\left[\left\|X_n - \theta\right\|^2\right] < \infty$.*

Proof of Theorem 3. In the proof of theorem 1, take $\beta_n = 0$, $\delta_n = 0$, $b_n < b$, $d_n < d$; then a.s.:

$$E\left[\left\|X_{n+1}^1\right\|^2 |T_n\right] \leq \left(1 + \lambda^2 a_n^2 + 3ba_n^2\right) \left\|X_n^1\right\|^2 + 3\left(b\left\|\theta\right\|^2 + d\right) a_n^2 - 2a_n\lambda \left\|X_n^1\right\|^2.$$

Taking the mathematical expectation yields:

$$E\left[\left\|X_{n+1}^1\right\|^2\right] \leq \left(1 + \left(\lambda^2 + 3b\right) a_n^2\right) E\left[\left\|X_n^1\right\|^2\right] + 3\left(b\left\|\theta\right\|^2 + d\right) a_n^2 - 2a_n\lambda E\left[\left\|X_n^1\right\|^2\right].$$

By Robbins-Siegmund lemma:

$$\exists t \geq 0 : E\left[\left\|X_n^1\right\|^2\right] \longrightarrow t; \sum_{n=1}^{\infty} a_n E\left[\left\|X_n^1\right\|^2\right] < \infty.$$

As $\sum_{n=1}^{\infty} a_n = \infty$, $t = 0$. Therefore, there exist $N \in \mathbb{N}$ and $f > 0$ such that for $n > N$:

$$E\left[\left\|X_{n+1}^1\right\|^2\right] \leq \left(1 - 2a_n\lambda\right) E\left[\left\|X_n^1\right\|^2\right] + fa_n^2.$$

Applying a lemma of Schmetterer [13] for $a_n = \frac{a}{n^\alpha}$ with $\frac{1}{2} < \alpha < 1$ yields:

$$\overline{\lim}\, n^\alpha E\left[\left\|X_n^1\right\|^2\right] < \infty.$$

Applying a lemma of Venter [14] for $a_n = \frac{a}{n}$ with $a > \frac{1}{2\lambda}$ yields:

$$\overline{\lim}\, nE\left[\left\|X_n^1\right\|^2\right] < \infty \ \blacksquare$$

2.2 Application to linear regression with online standardized data

Let $(R_1, S_1), ..., (R_n, S_n), ...$ be an i.i.d. sample of a random vector (R, S) in $\mathbb{R}^p \times \mathbb{R}^q$. Let Γ (respectively Γ^1) be the diagonal matrix of order p (respectively q) of the inverses of the standard deviations of the components of R (respectively S).

Define the correlation matrices

$$B = \Gamma E\left[\left(R - E\left[R\right]\right)\left(R - E\left[R\right]\right)'\right]\Gamma,$$
$$F = \Gamma E\left[\left(R - E\left[R\right]\right)\left(S - E\left[S\right]\right)'\right]\Gamma^1.$$

Suppose that B^{-1} exists. Let $\theta = B^{-1}F$.

Denote $\bar{R}_n$ (respectively $\bar{S}_n$) the mean of the n-sample $(R_1, R_2, ..., R_n)$ of R (respectively $(S_1, S_2, ..., S_n)$ of S).

Denote $\left(V_n^j\right)^2$ the variance of the n-sample $\left(R_1^j, R_2^j, ..., R_n^j\right)$ of the j^{th} component R^j of R, and $\left(V_n^{1k}\right)^2$ the variance of the n-sample $\left(S_1^k, S_2^k, ..., S_n^k\right)$ of the k^{th} component S^k of S.

Denote Γ_n (respectively Γ_n^1) the diagonal matrix of order p (respectively q) whose element (j,j) (respectively (k,k)) is the inverse of $\sqrt{\frac{n}{n-1}}V_n^j$ (respectively $\sqrt{\frac{n}{n-1}}V_n^{1k}$).

Let $(m_n, n \geq 1)$ be a sequence of integers. Denote $M_n = \sum_{k=1}^{n} m_k$ for $n \geq 1$, $M_0 = 0$ and $I_n = \{M_{n-1} + 1, ..., M_n\}$.

Define

$$\begin{aligned} B_n &= \Gamma_{M_{n-1}} \tfrac{1}{m_n} \textstyle\sum_{j \in I_n} \left(R_j - \bar{R}_{M_{n-1}}\right)\left(R_j - \bar{R}_{M_{n-1}}\right)' \Gamma_{M_{n-1}}, \\ F_n &= \Gamma_{M_{n-1}} \tfrac{1}{m_n} \textstyle\sum_{j \in I_n} \left(R_j - \bar{R}_{M_{n-1}}\right)\left(S_j - \bar{S}_{M_{n-1}}\right)' \Gamma^1_{M_{n-1}}. \end{aligned}$$

Define recursively the process $(X_n, n \geq 1)$ in $\mathbb{R}^{p \times q}$ by

$$X_{n+1} = X_n - a_n\left(B_n X_n - F_n\right).$$

Corollary 4 *Suppose there is no affine relation between the components of R and the moments of order 4 of (R,S) exist. Suppose moreover that assumption H3a" holds:*

(H3a") $a_n > 0, \sum_{n=1}^{\infty} \frac{a_n}{\sqrt{n}} < \infty, \sum_{n=1}^{\infty} a_n^2 < \infty$.

Then X_n converges to θ a.s.

This process was tested on several datasets and some results are given in section 5 (process S11 for $m_n = 1$ and S12 for $m_n = 10$).

The following lemma is first proved.

Lemma 5 *Suppose the moments of order 4 of R exist and $a_n > 0$, $\sum_{n=1}^{\infty} \frac{a_n}{\sqrt{n}} < \infty$. Then $\sum_{n=1}^{\infty} a_n \left\|\bar{R}_{M_{n-1}} - E\left[R\right]\right\| < \infty$ and $\sum_{n=1}^{\infty} a_n \left\|\Gamma_{M_{n-1}} - \Gamma\right\| < \infty$ a.s.*

Proof of Lemma 5. The usual Euclidean norm for vectors and the spectral norm for matrices are used in the proof.

Step 1:

Denote $Var\left[R\right] = E\left[\left\|R - E\left[R\right]\right\|^2\right] = \sum_{j=1}^{p} Var\left[R^j\right]$.

$$E\left[\left\|\bar{R}_{M_{n-1}} - E\left[R\right]\right\|^2\right] = \sum_{j=1}^{p} Var\left[\bar{R}^j_{M_{n-1}}\right] = \sum_{j=1}^{p} \frac{Var\left[R^j\right]}{M_{n-1}} \leq \frac{Var\left[R\right]}{n-1}.$$

Then:

$$\sum_{n=1}^{\infty} a_n E\left[\left\|\bar{R}_{M_{n-1}} - E\left[R\right]\right\|\right] \quad \leq \quad \sqrt{Var\left[R\right]} \sum_{n=1}^{\infty} \frac{a_n}{\sqrt{n-1}} < \infty \text{ by H3a".}$$

It follows that $\sum_{n=1}^{\infty} a_n \left\|\bar{R}_{M_{n-1}} - E\left[R\right]\right\| < \infty$ a.s.

Likewise $\sum_{n=1}^{\infty} a_n \left\|\bar{S}_{M_{n-1}} - E\left[S\right]\right\| < \infty$ a.s.

Step 2:

$$\begin{aligned}
\left\|\Gamma_{M_{n-1}} - \Gamma\right\| &= \max_{j=1,\ldots,p} \left| \frac{1}{\sqrt{\frac{M_{n-1}}{M_{n-1}-1}} V^j_{M_{n-1}}} - \frac{1}{\sqrt{Var\left[R^j\right]}} \right| \\
&\leq \sum_{j=1}^{p} \frac{\left| \sqrt{\frac{M_{n-1}}{M_{n-1}-1}} V^j_{M_{n-1}} - \sqrt{Var\left[R^j\right]} \right|}{\sqrt{\frac{M_{n-1}}{M_{n-1}-1}} V^j_{M_{n-1}} \sqrt{Var\left[R^j\right]}} \\
&= \sum_{j=1}^{p} \frac{\left| \frac{M_{n-1}}{M_{n-1}-1} \left(V^j_{M_{n-1}}\right)^2 - Var\left[R^j\right] \right|}{\sqrt{\frac{M_{n-1}}{M_{n-1}-1}} V^j_{M_{n-1}} \sqrt{Var\left[R^j\right]} \left(\sqrt{\frac{M_{n-1}}{M_{n-1}-1}} V^j_{M_{n-1}} + \sqrt{Var\left[R^j\right]} \right)}.
\end{aligned}$$

Denote μ_4^j the centered moment of order 4 of R^j. We have :

$$\begin{aligned}
E\left[\left| \frac{M_{n-1}}{M_{n-1}-1} \left(V^j_{M_{n-1}}\right)^2 - Var\left[R^j\right] \right|\right] &\leq \sqrt{Var\left[\frac{M_{n-1}}{M_{n-1}-1} \left(V^j_{M_{n-1}}\right)^2\right]} \\
&= O\left(\sqrt{\frac{\mu_4^j - \left(Var\left[R^j\right]\right)^2}{M_{n-1}}} \right).
\end{aligned}$$

Then by H3a", as $M_{n-1} \geq n-1$:

$$\begin{aligned}
&\sum_{n=1}^{\infty} a_n \sum_{j=1}^{p} E\left[\left| \frac{M_{n-1}}{M_{n-1}-1} \left(V^j_{M_{n-1}}\right)^2 - Var\left[R^j\right] \right|\right] < \infty \\
\Rightarrow &\sum_{n=1}^{\infty} a_n \sum_{j=1}^{p} \left| \frac{M_{n-1}}{M_{n-1}-1} \left(V^j_{M_{n-1}}\right)^2 - Var[R^j] \right| < \infty \text{ a.s.}
\end{aligned}$$

As $\left(V^j_{M_{n-1}}\right)^2 \longrightarrow Var\left[R^j\right]$ a.s., $j = 1, ..., p$, this implies :

$$\sum_{n=1}^{\infty} a_n \left\|\Gamma_{M_{n-1}} - \Gamma\right\| < \infty \text{ a.s.} \blacksquare$$

Proof of Corollary 4.

Step 1: prove that assumption H1a1 of theorem 1 is verified.

Denote $R^c = R - E\left[R\right]$, $R_j^c = R_j - E\left[R\right]$, $\bar{R}_j^c = \bar{R}_j - E\left[R\right]$.

$$\begin{aligned}
B_n &= \Gamma_{M_{n-1}} \frac{1}{m_n} \sum_{j \in I_n} \left(R_j^c - \bar{R}^c_{M_{n-1}}\right) \left(R_j^c - \bar{R}^c_{M_{n-1}}\right)' \Gamma_{M_{n-1}} \\
&= \Gamma_{M_{n-1}} \frac{1}{m_n} \sum_{j \in I_n} \left(R_j^c R_j^{c\prime} - \bar{R}^c_{M_{n-1}} R_j^{c\prime} - R_j^c \left(\bar{R}^c_{M_{n-1}}\right)' + \bar{R}^c_{M_{n-1}} \left(\bar{R}^c_{M_{n-1}}\right)' \right) \Gamma_{M_{n-1}}. \\
B &= \Gamma E\left[R^c R^{c\prime}\right] \Gamma.
\end{aligned}$$

As $\Gamma_{M_{n-1}}$ and $\bar{R}_{M_{n-1}}$ are T_n-measurable and R_j^c, $j \in I_n$, is independent of T_n, with $E\left[R_j^c\right] = 0$:

$$\begin{aligned} E\left[B_n|T_n\right] - B &= \Gamma_{M_{n-1}}\left(E\left[R^c R^{c\prime}\right] + \bar{R}^c_{M_{n-1}}\left(\bar{R}^c_{M_{n-1}}\right)'\right)\Gamma_{M_{n-1}} - \Gamma E\left[R^c R^{c\prime}\right]\Gamma \\ &= \left(\Gamma_{M_{n-1}} - \Gamma\right) E\left[R^c R^{c\prime}\right]\Gamma_{M_{n-1}} + \Gamma E\left[R^c R^{c\prime}\right]\left(\Gamma_{M_{n-1}} - \Gamma\right) \\ &\quad + \Gamma_{M_{n-1}}\bar{R}^c_{M_{n-1}}\left(\bar{R}^c_{M_{n-1}}\right)'\Gamma_{M_{n-1}} \text{ a.s.} \end{aligned}$$

As $\Gamma_{M_{n-1}}$ and $\bar{R}^c_{M_{n-1}}$ converge respectively to Γ and 0 a.s. and by lemma 5, $\sum_{n=1}^{\infty} a_n \left\|\Gamma_{M_{n-1}} - \Gamma\right\| < \infty$ and $\sum_{n=1}^{\infty} a_n \left\|\bar{R}^c_{M_{n-1}}\right\| < \infty$ a.s., it follows that $\sum_{n=1}^{\infty} a_n \left\|E\left[B_n|T_n\right] - B\right\| < \infty$ a.s.

Step 2: prove that assumption H1a2 of theorem 1 is verified.

$$\begin{aligned} \left\|B_n - B\right\|^2 &\leq 2\left\|\Gamma_{M_{n-1}}\frac{1}{m_n}\sum_{j\in I_n}\left(R_j^c - \bar{R}^c_{M_{n-1}}\right)\left(R_j^c - \bar{R}^c_{M_{n-1}}\right)'\Gamma_{M_{n-1}}\right\|^2 \\ &\quad + 2\left\|\Gamma E\left[R^c R^{c\prime}\right]\Gamma\right\|^2 \\ &\leq 2\left\|\Gamma_{M_{n-1}}\right\|^4 \frac{1}{m_n}\sum_{j\in I_n}\left\|R_j^c - \bar{R}^c_{M_{n-1}}\right\|^4 + 2\left\|\Gamma E\left[R^c R^{c\prime}\right]\Gamma\right\|^2 \\ &\leq 2\left\|\Gamma_{M_{n-1}}\right\|^4 \frac{1}{m_n}\sum_{j\in I_n} 2^3\left(\left\|R_j^c\right\|^4 + \left\|\bar{R}^c_{M_{n-1}}\right\|^4\right) + 2\left\|\Gamma E\left[R^c R^{c\prime}\right]\Gamma\right\|^2. \end{aligned}$$

$$E\left[\left\|B_n - B\right\|^2 |T_n\right] \leq 2^4\left\|\Gamma_{M_{n-1}}\right\|^4\left(E\left[\left\|R^c\right\|^4\right] + \left\|\bar{R}^c_{M_{n-1}}\right\|^4\right) + 2\left\|\Gamma E\left[R^c R^{c\prime}\right]\Gamma\right\|^2 \text{ a.s.}$$

As $\Gamma_{M_{n-1}}$ and $\bar{R}^c_{M_{n-1}}$ converge respectively to Γ and 0 a.s., and $\sum_{n=1}^{\infty} a_n^2 < \infty$, it follows that $\sum_{n=1}^{\infty} a_n^2 E\left[\left\|B_n - B\right\|^2 |T_n\right] < \infty$ a.s.

Step 3: the proofs of the verification of assumptions H2a1 and H2a2 of theorem 1 are similar to the previous ones, B_n and B being respectively replaced by

$$\begin{aligned} F_n &= \Gamma_{M_{n-1}}\frac{1}{m_n}\sum_{j\in I_n}\left(R_j^c - \bar{R}^c_{M_{n-1}}\right)\left(S_j^c - \bar{S}^c_{M_{n-1}}\right)'\Gamma^1_{M_{n-1}}, \\ F &= \Gamma E\left[R^c S^{c\prime}\right]\Gamma^1 \blacksquare \end{aligned}$$

3 Convergence of an averaged process with a constant step-size

In this section, the process $(X_n, n \geq 1)$ with a constant step-size a and the averaged process $(Y_n, n \geq 1)$ in $\mathbb{R}^{p\times q}$ are recursively defined by

$$\begin{aligned} X_{n+1} &= X_n - a\left(B_n X_n - F_n\right) \\ Y_{n+1} &= \frac{1}{n+1}\sum_{j=1}^{n+1} X_j = Y_n - \frac{1}{n+1}\left(Y_n - X_{n+1}\right). \end{aligned}$$

The a.s. convergence of $(Y_n, n \geq 1)$ and its application to sequential linear regression are studied.

3.1 Lemma

Lemma 6 *Let three real sequences (u_n), (v_n) and (a_n), with $u_n > 0$ and $a_n > 0$ for all n, and a real positive number λ such that, for $n \geq 1$,*

$$u_{n+1} \quad \leq \quad (1 - a_n\lambda)\, u_n + a_n v_n.$$

Suppose:
1) $v_n \longrightarrow 0$
2) $\left(a_n = a < \frac{1}{\lambda}\right)$ *or* $\left(a_n \longrightarrow 0,\ \sum_{n=1}^{\infty} a_n = \infty\right)$.
Under assumptions 1 and 2, $u_n \longrightarrow 0$.

Proof of Lemma 6. In the case a_n depending on n, as $a_n \longrightarrow 0$, we can suppose without loss of generality that $1 - a_n\lambda > 0$ for $n \geq 1$. We have:

$$u_{n+1} \quad \leq \quad \prod_{i=1}^{n} (1 - a_i\lambda)\, u_1 + \sum_{i=1}^{n} a_i \prod_{l=i+1}^{n} (1 - a_l\lambda)\, v_i, \text{ with } \prod_{n+1}^{n} = 1.$$

Now, for $n_1 \leq n_2 \leq n$ and $0 < c_i < 1$ with $c_i = a_i\lambda$ for all i, we have:

$$\begin{aligned}
\sum_{i=n_1}^{n_2} c_i \prod_{l=i+1}^{n} (1 - c_l) &= \sum_{i=n_1}^{n_2} (1 - (1 - c_i)) \prod_{l=i+1}^{n} (1 - c_l) \\
&= \sum_{i=n_1}^{n_2} \left(\prod_{l=i+1}^{n} (1 - c_l) - \prod_{l=i}^{n} (1 - c_l) \right) \\
&= \prod_{l=n_2+1}^{n} (1 - c_l) - \prod_{l=n_1}^{n} (1 - c_l) \leq \prod_{l=n_2+1}^{n} (1 - c_l) \leq 1.
\end{aligned}$$

Let $\epsilon > 0$. There exists N such that for $i > N$, $|v_i| < \frac{\epsilon}{3}\lambda$. Then for $n \geq N$, applying the previous inequality with $c_i = a_i\lambda$, $n_1 = 1$, $n_2 = N$, yields:

$$\begin{aligned}
u_{n+1} &\leq \prod_{i=1}^{n} (1 - a_i\lambda)\, u_1 + \sum_{i=1}^{N} a_i\lambda \prod_{l=i+1}^{n} (1 - a_l\lambda) \frac{|v_i|}{\lambda} + \frac{\epsilon}{3} \sum_{i=N+1}^{n} a_i\lambda \prod_{l=i+1}^{n} (1 - a_l\lambda) \\
&\leq \prod_{i=1}^{n} (1 - a_i\lambda)\, u_1 + \frac{1}{\lambda} \max_{1 \leq i \leq N} |v_i| \prod_{l=N+1}^{n} (1 - a_l\lambda) + \frac{\epsilon}{3}.
\end{aligned}$$

In the case a_n depending on n, $ln\,(1 - a_i\lambda) \sim -a_i\lambda$ as $a_i \longrightarrow 0\,(i \longrightarrow \infty)$; then, as $\sum_{n=1}^{\infty} a_n = \infty$, $\prod_{l=N+1}^{n} (1 - a_l\lambda) \longrightarrow 0\,(n \longrightarrow \infty)$.

In the case $a_n = a$, $\prod_{l=N+1}^{n} (1 - a\lambda) = (1 - a\lambda)^{n-N} \longrightarrow 0\,(n \longrightarrow \infty)$ as $0 < 1 - a\lambda < 1$.

Thus there exists N_1 such that $u_{n+1} < \epsilon$ for $n > N_1$ ■

3.2 Theorem

Make the following assumptions

(H1b) There exist a positive definite symmetrical matrix B in $\mathbb{R}^{p\times p}$ and a positive real number b such that a.s.

1) $\lim_{n\to\infty}\left(E\left[B_n|T_n\right]-B\right)=0$

2) $\sum_{n=1}^{\infty}\frac{1}{n}\left(E\left[\left\|E\left[B_n|T_n\right]-B\right\|^2\right]\right)^{\frac{1}{2}}<\infty$

3) $\sup_n E\left[\left\|B_n-B\right\|^2|T_n\right]\leq b.$

(H2b) There exist a matrix F in $\mathbb{R}^{p\times q}$ and a positive real number d such that a.s.

1) $\lim_{n\to\infty}\left(E\left[F_n|T_n\right]-F\right)=0$

2) $\sup_n E\left[\left\|F_n-F\right\|^2|T_n\right]\leq d.$

(H3b) λ and λ_{max} being respectively the smallest and the largest eigenvalue of B, $0<a<\min\left(\frac{1}{\lambda_{max}},\frac{2\lambda}{\lambda^2+b}\right)$.

Theorem 7 *Suppose H1b, H2b and H3b hold. Then Y_n converges to $\theta=B^{-1}F$ a.s.*

Remark 1 *Györfi and Walk [5] proved that Y_n converges to θ a.s. and in quadratic mean under the assumptions $E\left[B_n|T_n\right]=B$, $E\left[F_n|T_n\right]=F$, H1b2 and H2b2. Theorem 7 is an extension of their a.s. convergence result when $E\left[B_n|T_n\right]\longrightarrow B$ and $E\left[F_n|T_n\right]\longrightarrow F$ a.s.*

Remark 2 *Define $R_1=\begin{pmatrix}R\\1\end{pmatrix}$, $B=E\left[R_1R_1'\right]$, $F=E\left[R_1S'\right]$. If $\left(\left(R_{1n},S_n\right),n\geq 1\right)$ is an i.i.d. sample of (R_1,S) whose moments of order 4 exist, assumptions H1b and H2b are verified for $B_n=R_{1n}R_{1n}'$ and $F_n=R_{1n}S_n'$ as $E\left[R_{1n}R_{1n}'|T_n\right]=E\left[R_1R_1'\right]=B$ and $E\left[R_{1n}S_n'|T_n\right]=F$.*

Proof of Theorem 7. Denote

$$\begin{aligned}Z_n &= (B_n-B)(X_n-\theta)+(B_n-B)\theta-(F_n-F),\\ X_n^1 &= X_n-\theta,\\ Y_n^1 &= Y_n-\theta=\frac{1}{n}\sum_{j=1}^{n}X_j^1.\end{aligned}$$

Step 1: give a sufficient condition to have $Y_n^1\longrightarrow 0$ a.s.

We have (cf. proof of theorem 1):

$$\begin{aligned}X_{n+1}^1 &= \left(I-aB\right)X_n^1-aZ_n,\\ Y_{n+1}^1 &= \frac{1}{n+1}X_1^1+\frac{1}{n+1}\sum_{j=2}^{n+1}X_j^1\\ &= \frac{1}{n+1}X_1^1+\frac{1}{n+1}\sum_{j=2}^{n+1}\left(I-aB\right)X_{j-1}^1-a\frac{1}{n+1}\sum_{j=2}^{n+1}Z_{j-1}\\ &= \frac{1}{n+1}X_1^1+\frac{n}{n+1}\left(I-aB\right)Y_n^1-a\frac{1}{n+1}\sum_{j=1}^{n}Z_j.\end{aligned}$$

Take now the Frobenius norm of Y_{n+1}^1:

$$\left\|Y_{n+1}^1\right\| \leq \left\|\left(I-aB\right)Y_n^1\right\|+a\left\|\frac{1}{n+1}\sum_{j=1}^{n}Z_j-\frac{1}{n+1}\frac{1}{a}X_1^1\right\|.$$

Under H3b, all the eigenvalues of $I - aB$ are positive and the spectral norm of $I - aB$ is equal to $1 - a\lambda$. Then :

$$\left\|Y_{n+1}^1\right\| \quad \leq \quad (1 - a\lambda) \left\|Y_n^1\right\| + a \left\| \frac{1}{n+1} \sum_{j=1}^{n} Z_j - \frac{1}{n+1} \frac{1}{a} X_1^1 \right\|.$$

By lemma 6, it suffices to prove $\frac{1}{n} \sum_{j=1}^{n} Z_j \longrightarrow 0$ a.s. to conclude $Y_n^1 \longrightarrow 0$ a.s.

Step 2: prove that assumptions H1b and H2b imply respectively $\frac{1}{n} \sum_{j=1}^{n} B_j \longrightarrow B$ and $\frac{1}{n} \sum_{j=1}^{n} F_j \longrightarrow F$ a.s.

The proof is only given for (B_n), the other one being similar.

Assumption H1b3 implies $\sup_n E\left[\|B_n - B\|^2\right] < \infty$. It follows that, for each element B_n^{kl} and B^{kl} of B_n and B respectively, $\sum_{n=1}^{\infty} \frac{Var\left[B_n^{kl} - B^{kl}\right]}{n^2} < \infty$. Therefore:

$$\frac{1}{n} \sum_{j=1}^{n} \left(B_j^{kl} - B^{kl} - E\left[B_j^{kl} - B^{kl} | T_j\right]\right) \longrightarrow 0 \text{ a.s.}$$

As $E\left[B_j^{kl} - B^{kl} | T_j\right] \longrightarrow 0$ a.s. by H1b1, we have for each (k, l)

$$\frac{1}{n} \sum_{j=1}^{n} \left(B_j^{kl} - B^{kl}\right) \longrightarrow 0 \text{ a.s.}$$

Then $\frac{1}{n} \sum_{j=1}^{n} (B_j - B) \longrightarrow 0$ a.s.

Step 3: prove now that $\frac{1}{n} \sum_{j=1}^{n} (B_j - B) X_j^1 \longrightarrow 0$ a.s.

Denote $\beta_n = \|E[B_n | T_n] - B\|$ and $\gamma_n = \|E[F_n | T_n] - F\|$. $\beta_n \longrightarrow 0$ and $\gamma_n \longrightarrow 0$ a.s. under H1b1 and H2b1. Then: $\forall \delta > 0$, $\forall \varepsilon > 0$, $\exists N(\delta, \varepsilon)$: $\forall n \geq N(\delta, \varepsilon)$,

$$P\left(\{sup_{j>n}(\beta_j) \leq \delta\} \bigcap \{sup_{j>n}(\gamma_j) \leq \delta\}\right) > 1 - \varepsilon.$$

As $a < \frac{2\lambda}{\lambda^2 + b}$, choose η such that:

$$0 < \eta < \frac{1}{b} \left(\frac{2\lambda}{a} - \left(\lambda^2 + b\right) \right) \Leftrightarrow \lambda > \frac{a}{2} \left(\lambda^2 + b + \eta b\right).$$

Choose δ such that

$$0 < \delta < \frac{1}{(1 - a\lambda)(\|\theta\| + 2)} \Big(\lambda - \frac{a}{2} \big(\lambda^2 + b + \eta b\big) \Big).$$

Let ε be fixed. Denote $N_0 = N(\delta, \varepsilon)$ and, for $n > N_0$,

$$\begin{aligned} G_n &= \left(\left\{ \sup_{N_0 < j \leq n} (\beta_j) \leq \delta \right\} \bigcap \left\{ \sup_{N_0 < j \leq n} (\gamma_j) \leq \delta \right\} \right), \\ G &= \left(\left\{ \sup_{j > N_0} (\beta_j) \leq \delta \right\} \bigcap \left\{ \sup_{j > N_0} (\gamma_j) \leq \delta \right\} \right) = \bigcap_{n > N_0} G_n. \end{aligned}$$

Remark that G_n is T_n-measurable and, I_G denoting the indicator of G,

$$G \subset G_{n+1} \subset G_n \Leftrightarrow I_G \leq I_{G_{n+1}} \leq I_{G_n}.$$

Step 3a: prove that $\sup_n E\left[\left\|X_n^1\right\|^2 I_{G_n}\right] < \infty$.

$$\begin{aligned}\left\|X_{n+1}^1\right\|^2 I_{G_{n+1}} &\leq \left\|X_{n+1}^1\right\|^2 I_{G_n} = \left\|(I-aB)\, X_n^1 I_{G_n} - aZ_n I_{G_n}\right\|^2 \\ &\leq \left\|(I-aB)\, X_n^1 I_{G_n}\right\|^2 - 2a\left\langle (I-aB)\, X_n^1 I_{G_n}, Z_n I_{G_n}\right\rangle + a^2 \left\|Z_n I_{G_n}\right\|^2.\end{aligned}$$

As the spectral norm $\|I-aB\| = 1-a\lambda$, taking the conditional expectation with respect to T_n yields a.s.

$$\begin{aligned}E\left[\left\|X_{n+1}^1\right\|^2 I_{G_{n+1}} | T_n\right] &\leq (1-a\lambda)^2 \left\|X_n^1 I_{G_n}\right\|^2 - 2a\left\langle (I-aB)\, X_n^1 I_{G_n}, E\left[Z_n|T_n\right] I_{G_n}\right\rangle \\ &\quad + a^2 E\left[\left\|Z_n I_{G_n}\right\|^2 |T_n\right].\end{aligned}$$

Now:

$$\begin{aligned}\left\|E\left[Z_n|T_n\right] I_{G_n}\right\| &= \left\| \left(E\left[B_n|T_n\right] - B\right) X_n^1 I_{G_n} + \left(E\left[B_n|T_n\right] - B\right)\theta I_{G_n} \right. \\ &\quad \left. - \left(E\left[F_n|T_n\right] - F\right) I_{G_n}\right\| \\ &\leq \delta \left\|X_n^1 I_{G_n}\right\| + \delta\left(\|\theta\| + 1\right) \\ E\left[\left\|Z_n I_{G_n}\right\|^2 |T_n\right] &\leq (1+\eta)\, E\left[\left\|\left(B_n - B\right) X_n^1 I_{G_n}\right\|^2 |T_n\right] \\ &\quad + \left(1 + \frac{1}{\eta}\right) E\left[\left\|\left(B_n - B\right)\theta I_{G_n} - \left(F_n - F\right) I_{G_n}\right\|^2 |T_n\right] \\ &\leq (1+\eta)\, b \left\|X_n^1 I_{G_n}\right\|^2 + 2\left(1 + \frac{1}{\eta}\right)\left(b\|\theta\|^2 + d\right).\end{aligned}$$

Therefore:

$$\begin{aligned}E\left[\left\|X_{n+1}^1\right\|^2 I_{G_{n+1}} | T_n\right] &\leq \left((1-a\lambda)^2 + 2a\left(1-a\lambda\right)\delta + a^2\left(1+\eta\right) b\right)\left\|X_n^1 I_{G_n}\right\|^2 \\ &+ 2a\left(1-a\lambda\right)\delta\left(\|\theta\|+1\right)\left\|X_n^1 I_{G_n}\right\| \\ &+ 2a^2\left(1 + \frac{1}{\eta}\right)\left(b\|\theta\|^2 + d\right).\end{aligned}$$

As $\left\|X_n^1 I_{G_n}\right\| \leq 1 + \left\|X_n^1 I_{G_n}\right\|^2$, taking mathematical expectation yields:

$$\begin{aligned}E\left[\left\|X_{n+1}^1\right\|^2 I_{G_{n+1}}\right] &\leq \rho E\left[\left\|X_n^1 I_{G_n}\right\|^2\right] + e, \\ \rho &= (1-a\lambda)^2 + 2a\left(1-a\lambda\right)\delta\left(\|\theta\|+2\right) + a^2\left(1+\eta\right) b, \\ e &= 2a\left(1-a\lambda\right)\delta\left(\|\theta\|+1\right) + 2a^2\left(1+\frac{1}{\eta}\right)\left(b\|\theta\|^2 + d\right).\end{aligned}$$

As $\rho = 1 + 2a\left(\left(1-a\lambda\right)\left(\|\theta\|+2\right)\delta - \lambda + \frac{a}{2}\left(\lambda^2 + b + \eta b\right)\right) < 1$ by the choice of δ, this implies $g = \sup_n E\left[\left\|X_n^1\right\|^2 I_{G_n}\right] < \infty$.

Step 3b: conclusion.

$$\begin{aligned}E\left[\left\|\left(B_n - B\right) X_n^1 I_{G_n}\right\|^2\right] &= E\left[E\left[\left\|\left(B_n - B\right) X_n^1 I_{G_n}\right\|^2 |T_n\right]\right] \\ &\leq E\left[E\left[\left\|B_n - B\right\|^2 |T_n\right]\left\|X_n^1 I_{G_n}\right\|^2\right] \\ &\leq bg.\end{aligned}$$

Then: $\sum_{n=1}^{\infty} \frac{E\left[\left\|(B_n - B) X_n^1 I_{G_n}\right\|^2\right]}{n^2} < \infty$. Therefore a.s.:

$$\frac{1}{n}\sum_{j=1}^{n}\left((B_j - B) X_j^1 I_{G_j} - E\left[(B_j - B) X_j^1 I_{G_j}|T_j\right]\right) \longrightarrow 0.$$

Now:

$$\begin{aligned}\sum_{n=1}^{\infty}\frac{1}{n}E\left[\left\|(E[B_n|T_n] - B) X_n^1 I_{G_n}\right\|\right] &\leq \sum_{n=1}^{\infty}\frac{1}{n}E\left[\|E[B_n|T_n] - B\|\left\|X_n^1 I_{G_n}\right\|\right] \\ &\leq \sum_{n=1}^{\infty}\frac{1}{n}\left(E\left[\|E[B_n|T_n] - B\|^2\right]\right)^{\frac{1}{2}}\left(E\left[\left\|X_n^1 I_{G_n}\right\|^2\right]\right)^{\frac{1}{2}} \\ &\leq g^{\frac{1}{2}}\sum_{n=1}^{\infty}\frac{1}{n}\left(E\left[\|E[B_n|T_n] - B\|^2\right]\right)^{\frac{1}{2}} < \infty \text{ by H1b2.}\end{aligned}$$

Then:

$$\sum_{n=1}^{\infty}\frac{1}{n}\left\|(E[B_n|T_n] - B) X_n^1 I_{G_n}\right\| < \infty \text{ a.s.}$$

This implies by the Kronecker lemma:

$$\frac{1}{n}\sum_{j=1}^{n}(E[B_j|T_j] - B) X_j^1 I_{G_j} \longrightarrow 0 \text{ a.s.}$$

Therefore:

$$\frac{1}{n}\sum_{j=1}^{n}(B_j - B) X_j^1 I_{G_j} \longrightarrow 0 \text{ a.s.}$$

In G, $I_{G_j} = 1$ for all j, therefore $\frac{1}{n}\sum_{j=1}^{n}(B_j - B) X_j^1 \longrightarrow 0$ a.s. Then:

$P\left(\frac{1}{n}\sum_{j=1}^{n}(B_j - B) X_j^1 \longrightarrow 0\right) \geq P(G) > 1 - \varepsilon$. This is true for every $\varepsilon > 0$. Thus:

$$\frac{1}{n}\sum_{j=1}^{n}(B_j - B) X_j^1 \longrightarrow 0 \text{ a.s.}$$

Therefore by step 2 and step 1, we conclude that $\frac{1}{n}\sum_{j=1}^{n} Z_j \longrightarrow 0$ and $Y_n^1 \longrightarrow 0$ a.s. ■

3.3 Application to linear regression with online standardized data

Define as in section 2:

$$\begin{aligned}B_n &= \Gamma_{M_{n-1}}\frac{1}{m_n}\sum_{j\in I_n}\left(R_j - \bar{R}_{M_{n-1}}\right)\left(R_j - \bar{R}_{M_{n-1}}\right)'\Gamma_{M_{n-1}}, \\ F_n &= \Gamma_{M_{n-1}}\frac{1}{m_n}\sum_{j\in I_n}\left(R_j - \bar{R}_{M_{n-1}}\right)\left(S_j - \bar{S}_{M_{n-1}}\right)'\Gamma^1_{M_{n-1}}.\end{aligned}$$

Denote $U = (R - E[R])(R - E[R])'$, $B = \Gamma E[U]\Gamma$ the correlation matrix of R, λ and λ_{max} respectively the smallest and the largest eigenvalue of B, $b_1 = E\left[\|\Gamma U\Gamma - B\|^2\right]$, $F = \Gamma E\left[(R - E[R])(S - E[S])'\right]\Gamma$[1].

Corollary 8 *Suppose there is no affine relation between the components of R and the moments of order 4 of (R, S) exist. Suppose H3b1 holds:*

(H3b1) $0 < a < min\left(\frac{1}{\lambda_{max}}, \frac{2\lambda}{\lambda^2 + b_1}\right)$.

Then Y_n converges to $\theta = B^{-1}F$ a.s.

This process was tested on several datasets and some results are given in section 5 (process S21 for $m_n = 1$ and S22 for $m_n = 10$).

Proof of Corollary 8.

Step 1: introduction.

Using the decomposition of $E[B_n|T_n] - B$ established in the proof of corollary 4, as $\bar{R}_{M_{n-1}} \longrightarrow E[R]$ and $\Gamma_{M_{n-1}} \longrightarrow \Gamma$ a.s., it is obvious that $E[B_n|T_n] - B \longrightarrow 0$ a.s. Likewise $E[F_n|T_n] - F \longrightarrow 0$ a.s. Thus assumptions H1b1 and H2b1 are verified.

Suppose that Y_n does not converge to θ almost surely.

Then there exists a set of probability $\varepsilon_1 > 0$ in which Y_n does not converge to θ.

Denote $\sigma^j = \sqrt{Var[R^j]}$, $j = 1, ..., p$.

As $\bar{R}_{M_{n-1}} - E[R] \longrightarrow 0$, $\sqrt{\frac{M_{n-1}}{M_{n-1}-1}}V^j_{M_{n-1}} - \sigma^j \longrightarrow 0$, $j = 1, ..., p$ and $\Gamma_{M_{n-1}} - \Gamma \longrightarrow 0$ almost surely, there exists a set G of probability greater than $1 - \frac{\varepsilon_1}{2}$ in which these sequences of random variables converge uniformly to θ.

Step 2: prove that $\sum_{n=1}^{\infty} \frac{1}{n}\left(E\left[\left\|\Gamma_{M_{n-1}} - \Gamma\right\| I_G\right]\right)^{\frac{1}{2}} < \infty$.

By step 2 of the proof of lemma 5, we have for $n > N$:

$$\left\|\Gamma_{M_{n-1}} - \Gamma\right\| I_G \leq \sum_{j=1}^{p} \frac{\left|\frac{M_{n-1}}{M_{n-1}-1}\left(V^j_{M_{n-1}}\right)^2 - \left(\sigma^j\right)^2\right|}{\sqrt{\frac{M_{n-1}}{M_{n-1}-1}}V^j_{M_{n-1}}\sigma^j\left(\sqrt{\frac{M_{n-1}}{M_{n-1}-1}}V^j_{M_{n-1}} + \sigma^j\right)} I_G.$$

As in G, $\sqrt{\frac{M_{n-1}}{M_{n-1}-1}}V^j_{M_{n-1}}$ converges uniformly to σ^j for $j = 1, ..., p$, there exists $c > 0$ such that

$$\left\|\Gamma_{M_{n-1}} - \Gamma\right\| I_G \leq c\sum_{j=1}^{p}\left|\frac{M_{n-1}}{M_{n-1}-1}\left(V^j_{M_{n-1}}\right)^2 - \left(\sigma^j\right)^2\right|.$$

Then there exists $d > 0$ such that

$$E\left[\left\|\Gamma_{M_{n-1}} - \Gamma\right\| I_G\right] \leq \frac{d}{\sqrt{M_{n-1}}} \leq \frac{d}{\sqrt{n-1}}.$$

Therefore $\sum_{n=1}^{\infty} \frac{1}{n}\left(E\left[\left\|\Gamma_{M_{n-1}} - \Gamma\right\| I_G\right]\right)^{\frac{1}{2}} < \infty$.

Step 3: prove that assumption H1b2 is verified in G.

Using the decomposition of $E[B_n|T_n] - B$ given in step 1 of the proof of corollary 4, with $R^c = R - E[R]$ and $\bar{R}^c_{M_{n-1}} = \bar{R}_{M_{n-1}} - E[R]$ yields a.s.:

$$\begin{aligned}(E[B_n|T_n] - B) I_G = {} & \left(\left(\Gamma_{M_{n-1}} - \Gamma\right) E\left[R^c R^{c\prime}\right]\Gamma_{M_{n-1}} + \Gamma E\left[R^c R^{c\prime}\right]\left(\Gamma_{M_{n-1}} - \Gamma\right)\right. \\ & \left. + \Gamma_{M_{n-1}}\bar{R}^c_{M_{n-1}}\left(\bar{R}^c_{M_{n-1}}\right)'\Gamma_{M_{n-1}}\right) I_G.\end{aligned}$$

As in G, $\Gamma_{M_{n-1}} - \Gamma$ and $\bar{R}^c_{M_{n-1}}$ converge uniformly to 0, $E[B_n|T_n] - B$ converges uniformly to 0. Moreover there exists $c_1 > 0$ such that

$$\|E[B_n|T_n] - B\| I_G \leq c_1 \left(\left\|\Gamma_{M_{n-1}} - \Gamma\right\| I_G + \left\|\bar{R}^c_{M_{n-1}}\right\| \right) \text{ a.s.}$$

By the proof of lemma 5: $E\left[\left\|\bar{R}^c_{M_{n-1}}\right\|\right] \leq \left(\frac{Var[R]}{n-1}\right)^{\frac{1}{2}}$; then

$\sum_{n=1}^{\infty} \frac{1}{n} \left(E\left[\left\|\bar{R}^c_{M_{n-1}}\right\|\right]\right)^{\frac{1}{2}} < \infty$.

By step 2: $\sum_{n=1}^{\infty} \frac{1}{n} \left(E\left[\left\|\Gamma_{M_{n-1}} - \Gamma\right\| I_G\right]\right)^{\frac{1}{2}} < \infty$.

Then: $\sum_{n=1}^{\infty} \frac{1}{n} \left(E\left[\|E[B_n|T_n] - B\| I_G\right]\right)^{\frac{1}{2}} < \infty$.

As $E[B_n|T_n] - B$ converges uniformly to 0 on G, we obtain:

$$\sum_{n=1}^{\infty} \frac{1}{n} \left(E\left[\|E[B_n|T_n] - B\|^2 I_G\right]\right)^{\frac{1}{2}} < \infty.$$

Thus assumption H1b2 is verified in G.

Step 4: prove that assumption H1b3 is verified in G.

Denote $R^c = R - E[R]$, $R^c_j = R_j - E[R]$, $\bar{R}^c_j = \bar{R}_j - E[R]$. Consider the decomposition:

$$\begin{aligned} B_n - B &= \Gamma_{M_{n-1}} \frac{1}{m_n} \sum_{j \in I_n} \left(R^c_j - \bar{R}^c_{M_{n-1}}\right)\left(R^c_j - \bar{R}^c_{M_{n-1}}\right)' \Gamma_{M_{n-1}} \\ &\quad -\Gamma E\left[R^c R^{c\prime}\right]\Gamma \\ &= \alpha_n + \beta_n \end{aligned}$$

$$\begin{aligned} \text{with } \alpha_n &= \Gamma_{M_{n-1}} \frac{1}{m_n} \sum_{j \in I_n} \left(R^c_j R^{c\prime}_j - \bar{R}^c_{M_{n-1}} R^{c\prime}_j - R^c_j \left(\bar{R}^c_{M_{n-1}}\right)' + \bar{R}^c_{M_{n-1}} \left(\bar{R}^c_{M_{n-1}}\right)' \right) \Gamma_{M_{n-1}} \\ &\quad -\Gamma \frac{1}{m_n} \sum_{j \in I_n} R^c_j R^{c\prime}_j \Gamma \\ &= \left(\Gamma_{M_{n-1}} - \Gamma\right) \left(\frac{1}{m_n} \sum_{j \in I_n} R^c_j R^{c\prime}_j \right) \Gamma_{M_{n-1}} + \Gamma \left(\frac{1}{m_n} \sum_{j \in I_n} R^c_j R^{c\prime}_j \right) \left(\Gamma_{M_{n-1}} - \Gamma\right) \\ &\quad -\Gamma_{M_{n-1}} \bar{R}^c_{M_{n-1}} \frac{1}{m_n} \sum_{j \in I_n} R^{c\prime}_j \Gamma_{M_{n-1}} - \Gamma_{M_{n-1}} \frac{1}{m_n} \sum_{j \in I_n} R^c_j \left(\bar{R}^c_{M_{n-1}}\right)' \Gamma_{M_{n-1}} \\ &\quad +\Gamma_{M_{n-1}} \bar{R}^c_{M_{n-1}} \left(\bar{R}^c_{M_{n-1}}\right)' \Gamma_{M_{n-1}}, \\ \beta_n &= \Gamma \left(\frac{1}{m_n} \sum_{j \in I_n} R^c_j R^{c\prime}_j - E\left[R^c R^{c\prime}\right] \right) \Gamma. \end{aligned}$$

Let $\eta > 0$.

$$\begin{aligned} E\left[\|B_n - B\|^2 I_G | T_n\right] &= E\left[\|\alpha_n + \beta_n\|^2 I_G | T_n\right] \\ &\leq \left(1 + \frac{1}{\eta}\right) E\left[\|\alpha_n\|^2 I_G | T_n\right] \\ &\quad + (1+\eta) E\left[\|\beta_n\|^2 I_G | T_n\right] \text{ a.s.} \end{aligned}$$

As random variables R_j^c, $j \in I_n$, are independent of T_n, as $\Gamma_{M_{n-1}}$ and $\bar{R}^c_{M_{n-1}}$ are T_n-measurable and converge uniformly respectively to Γ and 0 on G, $E\left[\|\alpha_n\|^2 I_G|T_n\right]$ converges uniformly to 0. Then, for $\delta > 0$, there exists N_1 such that for $n > N_1$, $E\left[\|\alpha_n\|^2 I_G|T_n\right] \leq \delta$ a.s.

Moreover, denoting $U = R^c R^{c\prime}$ and $U_j = R_j^c R_j^{c\prime}$, we have, as the random variables U_j form an i.i.d. sample of U:

$$\begin{aligned} E\left[\|\beta_n\|^2 |T_n\right] &= E\left[\left\|\frac{1}{m_n}\sum_{j\in I_n}\Gamma\left(U_j - E\left[U\right]\right)\Gamma\right\|^2 |T_n\right] \\ &\leq E\left[\|\Gamma\left(U - E\left[U\right]\right)\Gamma\|^2\right] = E\left[\|\Gamma U\Gamma - E\left[\Gamma U\Gamma\right]\|^2\right] = b_1 \text{ a.s.} \end{aligned}$$

Then:

$$E\left[\|B_n - B\|^2 I_G|T_n\right] \leq \left(1+\frac{1}{\eta}\right)\delta + (1+\eta)\, b_1 = b \text{ a.s.}$$

Thus assumption H1b3 is verified in G.

As $\bar{S}_{M_{n-1}} - E\left[S\right] \longrightarrow 0$ and $\Gamma^1_{M_{n-1}} - \Gamma^1 \longrightarrow 0$ almost surely, it can be proved likewise that there exist a set H of probability greater than $1 - \frac{\varepsilon_1}{2}$ and $d > 0$ such that $E\left[\|F_n - F\|^2 I_H|T_n\right] \leq d$ a.s. Thus assumption H2b2 is verified in H.

Step 5: conclusion.

As $a < min\left(\frac{1}{\lambda_{max}}, \frac{2\lambda}{\lambda^2 + b_1}\right)$, $b_1 < \frac{2\lambda}{a} - \lambda^2$.

Choose $0 < \eta < \frac{\frac{2\lambda}{a} - \lambda^2}{b_1} - 1$ and $0 < \delta < \frac{\frac{2\lambda}{a} - \lambda^2 - (1+\eta)\, b_1}{1 + \frac{1}{\eta}}$ such that

$$b = \left(1+\frac{1}{\eta}\right)\delta + (1+\eta)\, b_1 < \frac{2\lambda}{a} - \lambda^2 \Longleftrightarrow a < \frac{2\lambda}{\lambda^2 + b}.$$

Thus assumption H3b is verified.

Applying theorem 7 implies that Y_n converges to θ almost surely in $H \cap G$.

Therefore $P\left(Y_n \longrightarrow \theta\right) \geq P\left(H \cap G\right) > 1 - \varepsilon_1$.

This is in contradiction with $P\left(Y_n \nrightarrow \theta\right) = \varepsilon_1$. Thus Y_n converges to θ a.s. ■

4 Convergence of a process with a variable or constant step-size and use of all observations until the current step

In this section, the convergence of the process $(X_n, n \geq 1)$ in $\mathbb{R}^{p\times q}$ recursively defined by

$$X_{n+1} = X_n - a_n\left(B_n X_n - F_n\right)$$

and its application to sequential linear regression are studied.

4.1 Theorem

Make the following assumptions

(H1c) There exists a positive definite symmetrical matrix B such that $B_n \longrightarrow B$ a.s.

(H2c) There exists a matrix F such that $F_n \longrightarrow F$ a.s.
(H3c) λ_{max} denoting the largest eigenvalue of B,
$\left(a_n = a < \frac{1}{\lambda_{max}}\right)$ or $\left(a_n \longrightarrow 0, \sum_{n=1}^{\infty} a_n = \infty\right)$.

Theorem 9 *Suppose H1c, H2c and H3c hold. Then X_n converges to $B^{-1}F$ a.s.*

Proof of Theorem 9.
Denote $\theta = B^{-1}F$, $X_n^1 = X_n - \theta$, $Z_n = (B_n - B)\,\theta - (F_n - F)$. Then:

$$X_{n+1}^1 = (I - a_n B_n)\, X_n^1 - a_n Z_n.$$

Let ω be fixed belonging to the intersection of the convergence sets $\{B_n \longrightarrow B\}$ and $\{F_n \longrightarrow F\}$. The writing of ω is omitted in the following.

Denote $\|A\|$ the spectral norm of a matrix A and λ the smallest eigenvalue of B.

In the case a_n depending on n, as $a_n \longrightarrow 0$, we can suppose without loss of generality $a_n < \frac{1}{\lambda_{max}}$ for all n. Then all the eigenvalues of $I - a_n B$ are positive and $\|I - a_n B\| = 1 - a_n\lambda$.

Let $0 < \varepsilon < \lambda$. As $B_n - B \longrightarrow 0$, we obtain for n sufficiently large:

$$\begin{aligned} \|I - a_n B_n\| &\leq \|I - a_n B\| + a_n \|B_n - B\| \\ &\leq 1 - a_n\lambda + a_n\varepsilon \text{ , with } a_n < \frac{1}{\lambda - \varepsilon} \\ \left\|X_{n+1}^1\right\| &\leq (1 - a_n(\lambda - \varepsilon)) \left\|X_n^1\right\| + a_n \|Z_n\|. \end{aligned}$$

As $Z_n \longrightarrow 0$, applying lemma 6 yields $\left\|X_n^1\right\| \longrightarrow 0$.
Therefore $X_n \longrightarrow B^{-1}F$ a.s. ■

4.2 Application to linear regression with online standardized data

Let $(m_n, n \geq 1)$ be a sequence of integers. Denote $M_n = \sum_{k=1}^{n} m_k$ for $n \geq 1$, $M_0 = 0$ and $I_n = \{M_{n-1} + 1, ..., M_n\}$.

Define

$$\begin{aligned} B_n &= \Gamma_{M_n} \left(\frac{1}{M_n} \sum_{i=1}^{n} \sum_{j \in I_i} R_j R_j' - \bar{R}_{M_n} \bar{R}_{M_n}' \right) \Gamma_{M_n}, \\ F_n &= \Gamma_{M_n} \left(\frac{1}{M_n} \sum_{i=1}^{n} \sum_{j \in I_i} R_j S_j' - \bar{R}_{M_n} \bar{S}_{M_n}' \right) \Gamma_{M_n}^1. \end{aligned}$$

As $((R_n, S_n), n \geq 1)$ is an i.i.d. sample of (R, S), assumptions H1c and H2c are obviously verified with $B = \Gamma E\left[(R - E[R])(R - E[R])'\right]\Gamma$ and $F = \Gamma E\left[(R - E[R])(S - E[S])'\right]\Gamma^1$. Then:

Corollary 10 *Suppose there is no affine relation between the components of R and the moments of order 4 of (R, S) exist. Suppose H3c holds. Then X_n converges to $B^{-1}F$ a.s.*

Remark 3 *B is the correlation matrix of R of dimension p. Then $\lambda_{max} < Trace\,(B) = p$. In the case of a constant step-size a, it suffices to take $a \leq \dfrac{1}{p}$ to verify H3c.*

Remark 4 *In the definition of B_n and F_n, the R_j and the S_j are not directly pseudo-centered with respect to $\bar{R}_{M_n}$ and $\bar{S}_{M_n}$ respectively. Another equivalent definition of B_n and F_n can be used. It consists of replacing R_j by $R_j - m$, $\bar{R}_{M_n}$ by $\bar{R}_{M_n} - m$, S_j by $S_j - m$, $\bar{S}_{M_n}$ by $\bar{S}_{M_n} - m_1$, m and m_1 being respectively an estimation of $E[R]$ and $E[S]$ computed in a preliminary phase with a small number of observations. For example, at step n, $\sum_{j \in I_n} \Gamma_{M_n} (R_j - m) (\Gamma_{M_n} (R_j - m))'$ is computed instead of $\sum_{j \in I_n} \Gamma_{M_n} R_j (\Gamma_{M_n} R_j)'$. This limits the risk of numerical explosion.*

This process was tested on several datasets and some results are given in section 5 (with a variable step-size: process S13 for $m_n = 1$ and S14 for $m_n = 10$; with a constant step-size: process S31 for $m_n = 1$ and S32 for $m_n = 10$).

5 Experiments

The three previously-defined processes of stochastic approximation with online standardized data were compared with the classical stochastic approximation and averaged stochastic approximation (or averaged stochastic gradient descent) processes with constant step-size (denoted ASGD) studied in [5] and [6]. A description of the methods along with abbreviations and parameters used is given in Table 1.

Table 1. Description of the methods.

Method type	**Abbreviation**	**Type of data**	**Number of observations used at each step of the process**	**Use of all the observations until the current step**	**Step-size**	**Use of the averaged process**
Classic	C1	Raw data	1	No	variable	No
	C2		10			
	C3		1	Yes		
	C4		10			
ASGD	A1		1	No	constant	Yes
	A2		1			
Standardization 1	S11	Online standardized data	1	No	variable	No
	S12		10			
	S13		1	Yes		
	S14		10			
Standardization 2	S21		1	No	constant	Yes
	S22		10			
Standardization 3	S31		1	Yes		No
	S32		10			

With the variable S set at dimension 1, 11 datasets were considered, some of which are available in free access on the Internet, while others were derived from the EPHESUS study [15]: 6 in regression (continuous dependent variable) and 5 in linear discriminant analysis (binary dependent variable). All datasets used in our experiments

Table 2. Datasets used in our experiments.

Dataset name	N	p_a	p	Type of dependent variable	T^2	Number of outliers	
CADATA	20640	8	8	Continuous	1.6x10^6	122	www.dcc.fc.up.pt/∼ltorgo/Regression/DataSets.html
AILERONS	7154	40	9	Continuous	247.1	0	www.dcc.fc.up.pt/∼ltorgo/Regression/DataSets.html
ELEVATORS	8752	18	10	Continuous	7.7x10^4	0	www.dcc.fc.up.pt/∼ltorgo/Regression/DataSets.html
POLY	5000	48	12	Continuous	4.1x10^4	0	www.dcc.fc.up.pt/∼ltorgo/Regression/DataSets.html
eGFR	21382	31	15	Continuous	2.9x10^4	0	derived from EPHESUS study [15]
HEMG	21382	31	17	Continuous	6.0x10^4	0	derived from EPHESUS study [15]
QUANTUM	50000	78	14	Binary	22.5	1068	www.osmot.cs.cornell.edu/kddcup
ADULT	45222	97	95	Binary	4.7x10^{10}	20	www.cs.toronto.edu/∼delve/data/datasets.html
RINGNORM	7400	20	20	Binary	52.8	0	www.cs.toronto.edu/∼delve/data/datasets.html
TWONORM	7400	20	20	Binary	24.9	0	www.cs.toronto.edu/∼delve/data/datasets.html
HOSPHF30D	21382	32	15	Binary	8.1x10^5	0	derived from EPHESUS study [15]

N denotes the size of global sample, p_a the number of parameters available, p the number of parameters selected and T^2 the trace of $E\left[RR'\right]$. Outlier is defined as an observation whose the L2 norm is greater than five times the average norm.

are presented in detail in Table 2, along with their download links. An *a priori* selection of variables was performed on each dataset using a stepwise procedure based on Fisher's test with p-to-enter and p-to-remove fixed at 5 percent.

Let $D = \{(r_i, s_i), i = 1, 2, ..., N\}$ be the set of data in $\mathbb{R}^p \times \mathbb{R}$ and assuming that it represents the set of realizations of a random vector (R, S) uniformly distributed in D, then minimizing $E[(S - \theta' R - \eta)^2]$ is equivalent to minimizing $\frac{1}{N}\sum_{i=1}^{N}(s_i - \theta' r_i - \eta)^2$. One element of D (or several according to the process) is randomly drawn at each step to iterate the process.

To compare the methods, two different studies were performed: one by setting the total number of observations used, the other by setting the computing time.

The choice of step-size, the initialization for each method and the convergence criterion used are respectively presented and commented below.

Choice of step-size

In all methods of stochastic approximation, a suitable choice of step-size is often crucial for obtaining good performance of the process. If the step-size is too small, the convergence rate will be slower. Conversely, if the step-size is too large, a numerical explosion phenomenon may occur during the first iterations.

For the processes with a variable step-size (processes C1 to C4 and S11 to S14), we chose to use a_n of the following type:

$$a_n = \frac{c_\gamma}{(b+n)^\alpha}.$$

The constant $\alpha = \frac{2}{3}$ was fixed, as suggested by Xu [16] in the case of stochastic approximation in linear regression, and $b = 1$. The results obtained for the choice $c_\gamma = \frac{1}{p}$ are presented although the latter does not correspond to the best choice for a classical method.

For the ASGD method (A1, A2), two different constant step-sizes a as used in [6] were tested: $a = \frac{1}{T^2}$ and $a = \frac{1}{2T^2}$, T^2 denoting the trace of $E\left[RR'\right]$. Note that this

choice of constant step-size assumes knowing *a priori* the dataset and is not suitable for a data stream.

For the methods with standardization and a constant step-size a (S21, S22, S31, S32), $a = \frac{1}{p}$ was chosen since the matrix $E[RR']$ is thus the correlation matrix of R, whose trace is equal to p, such that this choice corresponds to that of [6].

Initialization of processes

All processes (X_n) were initialized by $X_1 = \underline{0}$, the null vector. For the processes with standardization, a small number of observations ($n = 1000$) were taken into account in order to calculate an initial estimate of the means and standard deviations.

Convergence criterion

The "theoretical vector" θ^1 is assigned as that obtained by the least square method in D such that $\theta^{1'} = \begin{pmatrix}\theta' & \eta\end{pmatrix}$. Let Θ^1_{n+1} be the estimator of θ^1 obtained by stochastic approximation after n iterations.

In the case of a process (X_n) with standardized data, which yields an estimation of the vector denoted θ_c in section 1 as $\theta = \Gamma\theta_c \left(\Gamma^1\right)^{-1}$ and $\eta = E[S] - \theta' E[R]$, we can define:

$$\begin{aligned} {\Theta^1_{n+1}}' &= \begin{pmatrix}\Theta'_{n+1} & H_{n+1}\end{pmatrix} \\ \text{with } \Theta_{n+1} &= \Gamma_{M_n} X_{n+1} (\Gamma^1_{M_n})^{-1} \\ H_{n+1} &= \bar{S}_{M_n} - \Theta'_{n+1} \bar{R}_{M_n}. \end{aligned}$$

To judge the convergence of the method, the cosine of the angle formed by exact θ^1 and its estimation θ^1_{n+1} was used as criterion,

$$\cos\left(\theta^1, \theta^1_{n+1}\right) = \frac{\theta^{1'}\theta^1_{n+1}}{\|\theta^1\|_2 \left\|\theta^1_{n+1}\right\|_2}.$$

Other criteria, such as $\frac{\left\|\theta^1 - \theta^1_{n+1}\right\|_2}{\|\theta^1\|_2}$ or $\frac{f(\theta^1_{n+1}) - f(\theta^1)}{f(\theta^1)}$, f being the loss function, were also tested, although the results are not presented in this article.

5.1 Study for a fixed total number of observations used

For all N observations used by the algorithm (N being the size of D) up to a maximum of $100N$ observations, the criterion value associated with each method and for each dataset was recorded. The results obtained after using $10N$ observations are provided in Table 3.

As can be seen in Table 3, a numerical explosion occured in most datasets using the classical methods with raw data and a variable step-size (C1 to C4). As noted in Table 2, these datasets had a high $T^2 = Tr(E[RR'])$. Corresponding methods S11 to S14 using the same variable step-size but with online standardized data quickly converged in most cases. However classical methods with raw data can yield good results for a suitable choice of step-size, as demonstrated by the results obtained for POLY dataset in Fig 1. The numerical explosion can arise from a too high step-size when n is small. This phenomenon can be avoided if the step-size is reduced, although if the latter is too small, the convergence rate will be slowed. Hence, the right balance must be found between step-size and convergence rate. Furthermore, the choice of this step-size generally depends on the dataset which is not known *a priori* in the case of a data stream. In conclusion, methods with standardized data appear to be more robust to the choice of step-size.

The ASGD method (A1 with constant step-size $a = \frac{1}{T^2}$ and A2 with $a = \frac{1}{2T^2}$) did not yield good results except for the RINGNORM and TWONORM datasets which

Table 3. Results after using 10N observations.

	CADATA	AILERONS	ELEVATORS	POLY	EGFR	HEMG	QUANTUM	ADULT	RINGNORM	TWONORM	HOSPHF30D	Mean rank
C1	Expl.	-0.0385	Expl.	Expl.	Expl.	Expl.	0.9252	Expl.	0.9998	1.0000	Expl.	11.6
C2	Expl.	0.0680	Expl.	Expl.	Expl.	Expl.	0.8551	Expl.	0.9976	0.9996	Expl.	12.2
C3	Expl.	0.0223	Expl.	Expl.	Expl.	Expl.	0.9262	Expl.	0.9999	1.0000	Expl.	9.9
C4	Expl.	-0.0100	Expl.	Expl.	Expl.	Expl.	0.8575	Expl.	0.9981	0.9996	Expl.	12.3
A1	-0.0013	0.4174	0.0005	0.3361	0.2786	0.2005	Expl.	0.0027	0.9998	1.0000	0.0264	9.2
A2	0.0039	0.2526	0.0004	0.1875	0.2375	0.1846	0.0000	0.0022	0.9999	1.0000	0.2047	8.8
S11	1.0000	0.9516	0.9298	1.0000	1.0000	0.9996	0.9999	0.7599	0.9999	1.0000	0.7723	5.2
S12	0.9999	0.9579	0.9311	1.0000	0.9999	0.9994	0.9991	0.6842	0.9999	1.0000	0.4566	6.1
S13	1.0000	0.9802	0.9306	1.0000	1.0000	0.9998	1.0000	0.7142	0.9999	1.0000	0.7754	3.7
S14	0.9999	0.9732	0.9303	1.0000	0.9999	0.9994	0.9991	0.6225	0.9998	1.0000	0.4551	6.9
S21	0.9993	0.6261	0.9935	Expl.	Expl.	Expl.	Expl.	Expl.	0.9998	1.0000	Expl.	10.5
S22	1.0000	0.9977	0.9900	1.0000	1.0000	0.9989	0.9999	-0.0094	0.9999	1.0000	0.9454	4.1
S31	1.0000	0.9988	0.9999	1.0000	1.0000	0.9992	0.9999	0.9907	0.9999	1.0000	0.9788	2.3
S32	1.0000	0.9991	0.9998	1.0000	1.0000	0.9992	0.9999	0.9867	0.9999	1.0000	0.9806	2.2

Expl. means numerical explosion.

were obtained by simulation (note that all methods functioned very well for these two datasets). Of note, A1 exploded for the QUANTUM dataset containing 1068 observations (2.1 %) whose L2 norm was fivefold greater than the average norm (Table 2). The corresponding method S21 with online standardized data yielded several numerical explosions with the $a = \frac{1}{p}$ step-size, however these explosions disappeared when using a smaller step-size (see Fig 1). Of note, it is assumed in corollary 8 that $0 < a < \min\left(\frac{1}{\lambda_{max}}, \frac{2\lambda}{\lambda^2 + b_1}\right)$; in the case of $a = \frac{1}{p}$, only $a < \frac{1}{\lambda_{max}}$ is certain.

Finally, for methods S31 and S32 with standardized data, the use of all observations until the current step and the very simple choice of the constant step-size $a = \frac{1}{p}$ uniformly yielded good results.

Thereafter, for each fixed number of observations used and for each dataset, the 14 methods ranging from the best (the highest cosine) to the worst (the lowest cosine) were ranked by assigning each of the latter a rank from 1 to 14 respectively, after which the mean rank in all 11 datasets was calculated for each method. A total of 100 mean rank values were calculated for a number of observations used varying from N to $100N$. The graph depicting the change in mean rank based on the number of observations used and the boxplot of the mean rank are shown in Fig 2.

Overall, for these 11 datasets, a method with standardized data, a constant step-size and use of all observations until the current step (S31, S32) represented the best method when the total number of observations used was fixed.

5.2 Study for a fixed processing time

For every second up to a maximum of 2 minutes, the criterion value associated to each dataset was recorded. The results obtained after a processing time of 1 minute are provided in Table 4.

The same conclusions can be drawn as those described in section 5.1 for the classical methods and the ASGD method. The methods with online standardized data typically faired better.

As in the previous study in section 5.1, the 14 methods were ranked from the best to the worst on the basis of the mean rank for a fixed processing time. The graph depicting the change in mean rank based on the processing time varying from 1 second to 2 minutes as well as the boxplot of the mean rank are shown in Fig 3.

Table 4. Results obtained after a fixed time of 1 minute.

	CADATA	AILERONS	ELEVATORS	POLY	EGFR	HEMG	QUANTUM	ADULT	RINGNORM	TWONORM	HOSPHF30D	Mean rank
C1	Expl.	-0.2486	Expl.	Expl.	Expl.	Expl.	0.9561	Expl.	1.0000	1.0000	Expl.	12.2
C2	Expl.	0.7719	Expl.	Expl.	Expl.	Expl.	0.9519	Expl.	1.0000	1.0000	Expl.	9.9
C3	Expl.	0.4206	Expl.	Expl.	Expl.	Expl.	0.9547	Expl.	1.0000	1.0000	Expl.	10.6
C4	Expl.	0.0504	Expl.	Expl.	Expl.	Expl.	0.9439	Expl.	1.0000	1.0000	Expl.	10.1
A1	-0.0067	0.8323	0.0022	0.9974	0.7049	0.2964	Expl.	0.0036	1.0000	1.0000	Expl.	9.0
A2	0.0131	0.8269	0.0015	0.9893	0.5100	0.2648	Expl.	0.0027	1.0000	1.0000	0.2521	8.6
S11	1.0000	0.9858	0.9305	1.0000	1.0000	1.0000	1.0000	0.6786	1.0000	1.0000	0.9686	5.8
S12	1.0000	0.9767	0.9276	1.0000	1.0000	0.9999	1.0000	0.6644	1.0000	1.0000	0.9112	5.8
S13	1.0000	0.9814	0.9299	1.0000	1.0000	0.9999	1.0000	0.4538	1.0000	1.0000	0.9329	6.1
S14	1.0000	0.9760	0.9274	1.0000	1.0000	1.0000	0.9999	0.5932	1.0000	1.0000	0.8801	6.1
S21	-0.9998	0.2424	0.6665	Expl.	Expl.	Expl.	Expl.	0.0000	1.0000	1.0000	Expl.	11.5
S22	1.0000	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000	-0.0159	1.0000	1.0000	0.9995	3.1
S31	1.0000	0.9995	1.0000	1.0000	1.0000	0.9999	1.0000	0.9533	1.0000	1.0000	0.9997	4.5
S32	1.0000	0.9999	1.0000	1.0000	1.0000	1.0000	1.0000	0.9820	1.0000	1.0000	0.9999	1.5

Expl. means numerical explosion.

As can be seen, these methods with online standardized data using more than one observation per step yielded the best results (S32, S22). One explanation may be that the total number of observations used in a fixed processing time is higher when several observations are used per step rather than one observation per step. This can be verified in Table 5 in which the total number of observations used per second for each method and for each dataset during a processing time of 2 minutes is given. Of note, the number of observations used per second in a process with standardized data and one observation per step (S11, S13, S21, S31) was found to be generally lower than in a process with raw data and one observation per step (C1, C3, A1, A2), since a method with standardization requires the recursive estimation of means and variances at each step.

Of note, for the ADULT dataset with a large number of parameters selected (95), the only method yielding sufficiently adequate results after a processing time of one minute was S32, and methods S31 and S32 when $10N$ observations were used.

Table 5. Number of observations used after 2 minutes (expressed in number of observations per second).

	CADATA	AILERONS	ELEVATORS	POLY	EGFR	HEMG	QUANTUM	ADULT	RINGNORM	TWONORM	HOSPHF30D
C1	19843	33170	17133	14300	10979	9243	33021	476	31843	31677	10922
C2	166473	291558	159134	134249	104152	89485	281384	4565	262847	261881	102563
C3	17206	28985	16036	13449	10383	8878	28707	462	28123	28472	10404
C4	132088	194031	125880	106259	87844	76128	184386	4252	171711	166878	86895
A1	33622	35388	36540	35800	35280	34494	11815	15390	34898	34216	14049
A2	33317	32807	36271	35628	35314	34454	15439	16349	34401	34205	34890
S11	17174	17133	17166	16783	15648	14764	16296	1122	14067	13836	14334
S12	45717	47209	45893	43470	39937	37376	40943	4554	34799	34507	36389
S13	12062	12731	11888	12057	11211	10369	11466	620	9687	9526	10137
S14	43674	46080	43068	42123	38350	35338	39170	4512	33594	31333	32701
S21	15396	17997	16772	10265	8404	7238	9166	996	13942	13274	7672
S22	47156	47865	46318	43899	40325	37467	41320	4577	34478	31758	37418
S31	12495	12859	12775	12350	11495	10619	11608	621	9890	9694	10863
S32	44827	47035	45123	42398	38932	36288	39362	4532	33435	33385	35556

6 Conclusion

In the present study, three processes with online standardized data were defined and for which their a.s. convergence was proven.

A stochastic approximation method with standardized data appears to be advantageous compared to a method with raw data. First, it is easier to choose the

step-size. For processes S31 and S32 for example, the definition of a constant step-size only requires knowing the number of parameters p. Secondly, the standardization usually allows avoiding the phenomenon of numerical explosion often obtained in the examples given with a classical method.

The use of all observations until the current step can reduce the influence of outliers and increase the convergence rate of a process. Moreover, this approach is particularly adapted to the case of a data stream.

Finally, among all processes tested on 11 different datasets (linear regression or linear discriminant analysis), the best was a method using standardization, a constant step-size equal to $\frac{1}{p}$ and all observations until the current step, and the use of several new observations at each step improved the convergence rate.

References

1. Monnez JM. Le processus d'approximation stochastique de Robbins-Monro : résultats théoriques ; estimation séquentielle d'une espérance conditionnelle. Statistique et Analyse des Données. 1979;4(2):11-29.

2. Ljung L. Analysis of stochastic gradient algorithms for linear regression problems. IEEE Transactions on Information Theory. 1984;30(2):151-160.

3. Polyak BT. New method of stochastic approximation type. Automation and remote control. 1990;51(7):937-946.

4. Polyak BT, Juditsky AB. Acceleration of stochastic approximation by averaging. SIAM Journal on Control and Optimization. 1992;30(4):838-855.

5. Györfi L, Walk H. On the averaged stochastic approximation for linear regression. SIAM Journal on Control and Optimization. 1996;34(1):31-61.

6. Bach F, Moulines E. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). Advances in Neural Information Processing Systems. 2013;773-781.

7. Bottou L, Le Cun Y. On-line learning for very large data sets. Applied Stochastic Models in Business and Industry. 2005;21(2):137-151.

8. Bottou L, Curtis FE, Noceda J. Optimization Methods for Large-Scale Machine Learning. arXiv:1606.04838v2. 2017.

9. Johnson R, Zhang Tong. Accelerating Stochastic Gradient Descent using Predictive Variance Reduction. Advances in Neural Information Processing Systems. 2013:315-323.

10. Duchi J, Hazan E, Singer Y. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. Journal of Machine Learning Research. 2011;12:2121-2159.

11. Pascanu R, Mikolov T, Bengio Y. Understanding the exploding gradient problem. arXiv:1211.5063v1. 2012.

12. Robbins H, Siegmund D. A convergence theorem for nonnegative almost supermartingales and some applications. Optimizing Methods in Statistics, Rustagi, J.S. (ed.), Academic Press, New York. 1971;233-257.

13. Schmetterer L. Multidimensional stochastic approximation. Multivariate Analysis II, Proc. 2nd Int. Symp., Dayton, Ohio, Academic Press. 1969;443-460.

14. Venter JH. On Dvoretzky stochastic approximation theorems. The Annals of Mathematical Statistics. 1966;37:1534-1544.

15. Pitt B., Remme W., Zannad F. et al. Eplerenone, a selective aldosterone blocker, in patients with left ventricular dysfunction after myocardial infarction. New England Journal of Medicine.2003;348(14):1309-1321.

16. Xu W. Towards optimal one pass large scale learning with averaged stochastic gradient descent. arXiv:1107.2490v2. 2011.

Fig 1. Results obtained for dataset POLY using $10N$ and $100N$ observations:
A/ process C1 with variable step-size $a_n = \frac{1}{(b+n)^{\frac{2}{3}}}$ by varying b,
B/ process C1 with variable step-size $a_n = \frac{\frac{1}{p}}{(b+n)^{\frac{2}{3}}}$ by varying b,
C/ process S21 by varying constant step-size a.

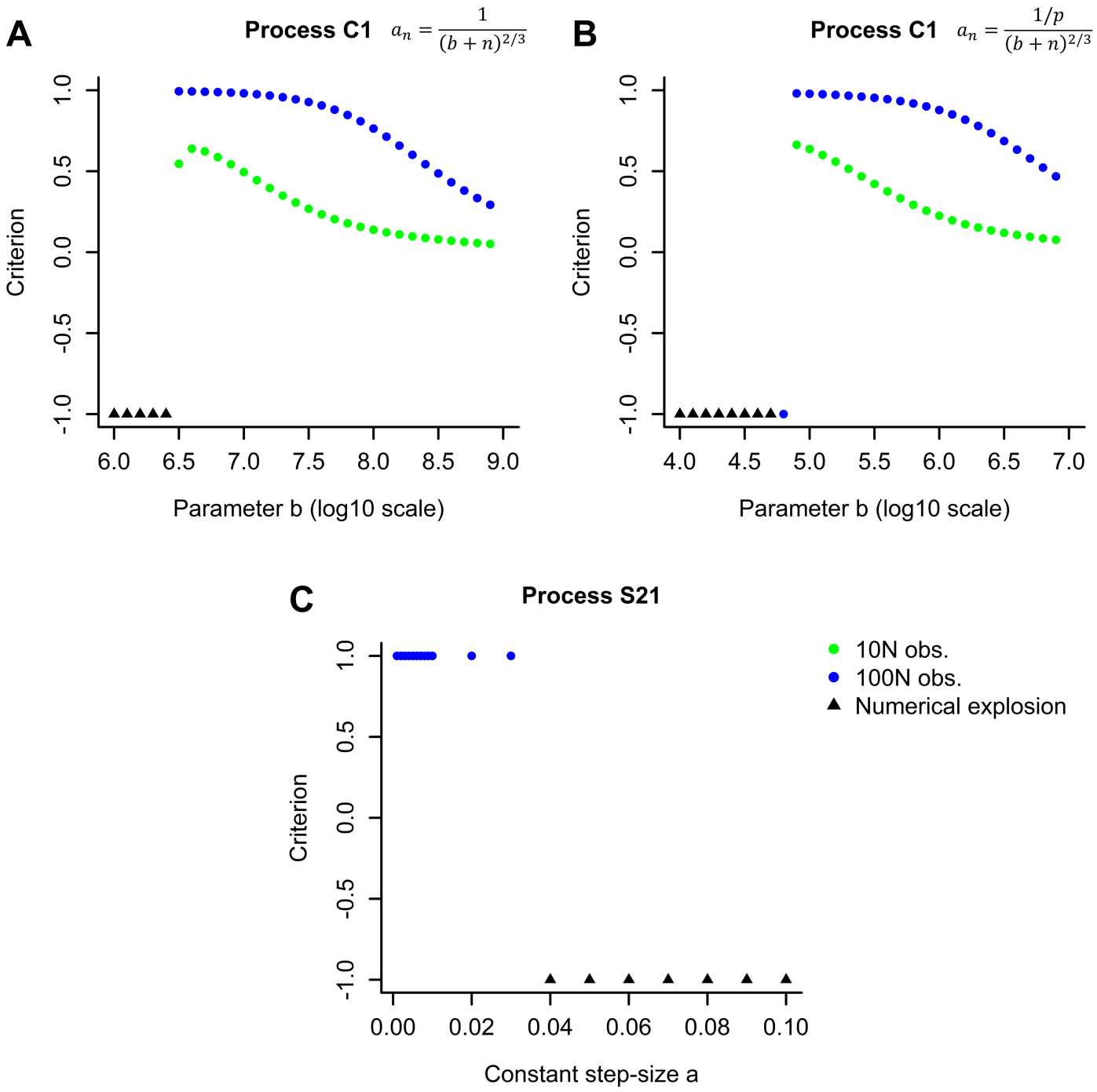

Fig 2. Results for a fixed total number of observations used: A/ change in the mean rank based on the number of observations used, B/ boxplot of the mean rank by method.

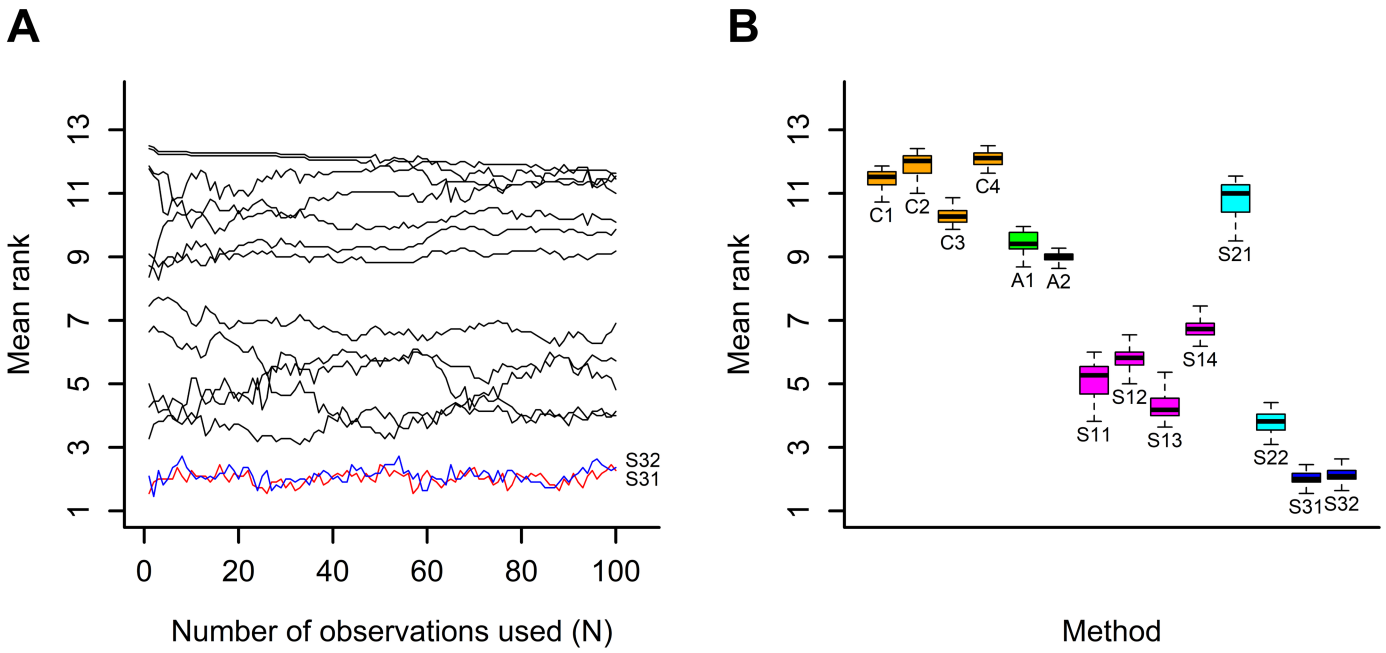

Fig 3. Results for a fixed processing time: A/ change in the mean rank based on the processing time, B/ boxplot of the mean rank by method.