On the asymptotic behaviour of random matrices in a multivariate statistical model

Roy Cerqueti* & Mauro Costantini

University "La Sapienza", Rome, Italy

Abstract

This paper aims to provide a nonparametric analysis of the integrated processes of an integer order, via a theoretical solution of a generalized eigenvalue problem. To this end, we introduce a mean operator for the process, by using weights belonging to a Sobolev Space.

Keywords: Generalized Eigenvalues Problem, Sobolev Spaces, Asymptotic Convergence

*Corresponding author.

Department of Mathematics, Faculty of Economics, University of Rome "La Sapienza", Via del Castro Laurenziano, 9 - 00161 - Rome, Italy.

E-mail: roy.cerqueti@uniroma1.it

1 Introduction

In this paper we develop a nonparametric model to analyze a p-variate process Y_t that is integrated of order d. More generally than Bierens (1997), whose model describe p-variate integrated processes of order 1, we take into account the α-th differences of Y_t ($\alpha = 1, \ldots, d$), that are opportunely weighted, and construct a pair of random matrices, related to the stationary and nonstationary part of the process, referring to the following definition:

> *Given $p \in \mathbf{N}$, d integer, a discrete time p-variate integrated process of order d, $Y_t \sim I(d)$, is defined by the following property: $\Delta^k Y_t$ is a nonstationary process, for $k = 0, 1, \ldots, d-1$ and $\Delta^d Y_t$ is a stationary process.*

Then we derive their asymptotic behaviors, using Andersen et al. (1982), and we solve a generalized eigenvalue problem.

The novelties of our model are basically two. First, we propose a nonparametric analysis of each integrated process of an integer order. Theoretical results covering cases of order 1 and 2, that are principally linked with economic phenomena, are obtained (see Bierens (1997) for the case of order 1). Second, the Sobolev Spaces theory is introduced (see Ladyzhenskaya and Uraltseva, 1968), in order to reduce the number of used weight functions.

The paper is organized as follows. Section 2 presents the data generating process. In Section 3 the random matrices are defined, and their asymptotic behavior is studied. Section 4 provides the solution of the generalized eigenvalue problem.

2 Data generating process

In this section we provide a description of the data generating process. First of all, we recall the basic definition of integrated processes of an integer order d.

Definition 2.1 *A discrete time p-variate integrated process of order d, $Y_t \sim I(d)$, is described by the following difference equation:*

$$Y_t = \Delta^{-d}\epsilon_t = (1-L)^{-d}\epsilon_t, \tag{1}$$

where $p \in \mathbf{N}$, $Y_t = (Y_t^1, \ldots, Y_t^p)$, $\epsilon_t = (\epsilon_t^1, \ldots, \epsilon_t^p)$ is a zero-mean stationary process, L is the lag operator, i.e. $L\epsilon_t := \epsilon_{t-1}$, and $\Delta := 1 - L$.

It is easy to show that if $Y_t \sim I(d)$, then $Y_t - Y_0 \sim I(d)$. Therefore, we don't lose of generality assuming $Y_t \sim I(d)$, with $Y_0 = 0$. This assumption is used for the rest of the paper.

If the hypotheses of the Wold decomposition theorem are satisfied, we can write

$$\epsilon_t = \sum_{j=0}^{\infty} C_j v_{t-j} =: C(L)v_t, \quad t = 1, \ldots, n, \tag{2}$$

where v_t is a p-variate stationary white noise process and $C(L)$ is a p-squared matrix of lag polynomials in the lag operator L.

Let us now state a condition for the matrix $C(L)$ defined in (2).

Assumption 2.1 *The process ϵ_t can be written as in (2), where v_t are i.i.d. zero-mean p-variate gaussian variables with variance equals to the identity matrix of order p, I_p, and there exist $C_1(L)$ and $C_2(L)$ p-squared matrices of lag polynomials in the lag operator L such that all the roots of $\det C_1(L)$ are outside the complex unit circle and $C(L) = C_1(L)^{-1}C_2(L)$.*

The lag polynomial $C(L) - C(1)$ attains value zero at $L = 1$ with algebraic multiplicity equals to d. Thus, there exists a lag polynomial

$$D(L) = \sum_{k=0}^{\infty} D_k L^k$$

such that $C(L) - C(1) = (1 - L)^d D(L)$. Therefore, we can write

$$\epsilon_t = C(L)v_t = C(1)v_t + [C(L) - C(1)]v_t = C(1)v_t + D(L)(1 - L)^d v_t. \quad (3)$$

Let us define $w_t := D(L)v_t$. Then, substituting w_t into (3), we get

$$\epsilon_t = C(1)v_t + (1 - L)^d w_t. \quad (4)$$

(4) implies that, given $Y_t \sim I(d)$, we can write recursively

$$\Delta^{d-1}Y_t = \Delta^{d-1}Y_{t-1} + \epsilon_t = \Delta^{d-1}Y_{t-1} + C(1)v_t +$$

$$+ (1 - L)^d w_t = \Delta^{d-1}Y_0 + (1 - L)^{d-1} w_t - w_0 + C(1) \sum_{j=1}^{t} v_j, \quad (5)$$

where $\text{rank}(C(1)) = p - r < p$.

Remark 2.1 *By Assumption 2.1, we have that $C(L)v_t$ and $D(L)v_t$ are well-defined stationary processes.*

Assumption 2.2 *Let us consider R_r the matrix of the eigenvectors of $C(1)C(1)^T$ corresponding to the r zero eigenvalues. Then the matrix $R_r^T D(1)D(1)^T R_r$ is nonsingular.*

Assumption 2.2 implies that Y_t cannot be integrated of order $\bar{d}$, with $\bar{d} > d$. In fact, if there exists $\bar{d} > d$ such that $Y_t \sim I(\bar{d})$, then the lag polynomial $D(L)$ admits a unit root with algebraic multiplicity $\bar{d} - d$, and so $D(1)$ is singular. Therefore $R_r^T D(1)D(1)^T R_r$ is singular, and Assumption 2.2 does not hold.

3 The weighted random matrices

This section starts by considering a transformation of the data generating process via a weighted mean operator, in order to define a pair of random matrices related to the stationary and nonstationary part of the process.

We introduce a weight function, representing the scale factor of Y_t. It can be formalized by defining the adjusted process z as follows.

$$z_n := \frac{1}{n}\sum_{t=1}^{n} Y_t \cdot G_n(t), \qquad n \in \mathbf{N} \text{ and } G : [0, +\infty) \to \mathbf{R}. \tag{6}$$

The nonstationary part of the process is

$$M_n^{NS} := \sum_{j=0}^{d-1} \Delta^j z_n. \tag{7}$$

A straightforward computation gives:

$$\Delta^j z_n = \frac{1}{n}\sum_{k=0}^{j} \binom{j}{k} \Big[\sum_{t=1}^{n} \Delta^k G_n(t) \cdot \Delta^{j-k} Y_t\Big], \qquad n \in \mathbf{N}. \tag{8}$$

By arranging the terms of $\Delta^j z_t$ with respect to the differences $\Delta^\alpha Y_t$, $\alpha \in \{0, \dots, j\}$, equation (7) can be rewritten as

$$M_n^{NS} = \frac{1}{n}\sum_{j=0}^{d-1}\sum_{t=1}^{n} \Delta^j Y_t \cdot \Big[\sum_{k=0}^{d-j-1} \binom{k+j}{k} \Delta^k G_n(t)\Big]. \tag{9}$$

Under some hypotheses on the asymptotic behavior and on the functional structure of the G_n's, the convergence of M_n is obtained. The following result holds.

Theorem 3.1 *Assume that the following conditions hold.*

- G_n *belongs to the Sobolev Space* $(H^{1,d-1}(0, +\infty), || \cdot ||_{1,d-1})$*, for each* $n \in \mathbf{N}$.

- *It results*

$$\lim_{n\to+\infty} n^2 4^{n(d-2)}||G_n||_{1,d-2} = 0.$$

- *There exists $F_n : [0,+\infty) \to \mathbf{R}$ with support (ζ_n, ξ_n) such that*

  - *$(G_n - F_n)$ belongs to the Sobolev Space $(H^{1,1}(0,+\infty), ||\cdot||_{1,1})$, for each $n \in \mathbf{N}$;*

  - *it results*

$$\lim_{n\to+\infty} 4^{n(d-1)}||G_n(x) - F_n(x)||_{1,1} = 0.$$

Then we have

$$\lim_{n\to+\infty} ||M_n^{NS} - \frac{1}{n}\sum_{t=1}^{n} F_n(t)\Delta^{d-1}Y_t|| = 0.$$

Proof. In order to prove the result, it is sufficient to show that $\forall\, \epsilon > 0$, $\exists\, n_\epsilon \in \mathbf{N}$ such that

$$n > n_\epsilon \Rightarrow ||M_n^{NS} - \frac{1}{n}\sum_{t=1}^{n} F_n(t)\Delta^{d-1}Y_t|| < \epsilon. \tag{10}$$

We stress that, $\forall\, \epsilon_1 > 0$, $\exists\, n^1_{\epsilon_1}, n^2_{\epsilon_1} \in \mathbf{N}$ such that,

$$\Big|\Big|\frac{1}{n}\sum_{t=1}^{n} \Delta^k G_n(t) - \int_{\mathbf{R}^+} G_n^{(k)}(x)\mathrm{d}x\Big|\Big| < \epsilon_1 \qquad \text{for } n > n^1_{\epsilon_1} \tag{11}$$

and

$$\Big|\Big|\frac{1}{n}\sum_{t=1}^{n} |G_n(t) - F_n(t)| - \int_{\mathbf{R}^+} |G_n(x) - F_n(x)|\mathrm{d}x\Big|\Big| < \epsilon_1 \qquad \text{for } n > n^2_{\epsilon_1}. \tag{12}$$

Then, by (11) and (12), for ϵ_1 small enough and $n > \max\{n^1_{\epsilon_1}, n^2_{\epsilon_1}\}$, we have

$$\Big|\Big|M_n^{NS} - \frac{1}{n}\sum_{t=1}^{n} F_n(t)\Delta^{d-1}Y_t\Big|\Big| =$$

$$= \Big|\Big|\frac{1}{n}\sum_{j=0}^{d-1}\sum_{t=1}^{n} \Delta^j Y_t \cdot \Big[\sum_{k=0}^{d-j-1}\binom{k+j}{k}\Delta^k G_n(t)\Big] - \frac{1}{n}\sum_{t=1}^{n} F_n(t)\Delta^{d-1}Y_t\Big|\Big| \leq$$

$$\leq \sum_{j=0}^{d-2} \Big|\Big| \sum_{t=1}^{n} \Delta^j Y_t \Big|\Big| \cdot \Big[\sum_{k=0}^{d-j-2} \binom{k+j}{k} \Big|\Big| \frac{1}{n} \sum_{t=1}^{n} \Delta^k G_n(t) \Big|\Big| \Big] +$$

$$+ \Big|\Big| \frac{1}{n} \sum_{t=1}^{n} [G_n(t) - F_n(t)] \Delta^{d-1} Y_t \Big|\Big| \sim$$

$$\sim \sum_{j=0}^{d-2} \sum_{t=1}^{n} ||\Delta^j Y_t|| \cdot \Big[\sum_{k=0}^{d-j-2} \binom{k+j}{k} \int_{\mathbf{R}^+} |G_n^{(k)}(x)| \mathrm{d}x \Big] +$$

$$+ ||\Delta^{d-1} Y_t|| \cdot \int_{\mathbf{R}^+} |G_n(x) - F_n(x)| \mathrm{d}x. \tag{13}$$

Since $Y_t = (Y_t^1, \dots, Y_t^p)$ is a p-variate $I(d)$ process, then $\Delta^j Y_t$ follows a gaussian law with zero mean and variance-covariance matrix with finite elements. By defining the norm

$$||\Delta^j Y_t|| = ||(\Delta^j Y_t^1, \dots, \Delta^j Y_t^p)|| := \Big[\mathbf{E}[\Delta^j Y_t^1]^2 + \ldots + \mathbf{E}[\Delta^j Y_t^p]^2 \Big],$$

then there exists a constant depending on t and j, $C(t,j) > 0$, such that

$$||\Delta^j Y_t|| < C(t,j). \tag{14}$$

Let us define

$$C_{n,j} := \max_{t=1,\dots,n} C(t,j). \tag{15}$$

Then the estimate in (13) can be refined. It results:

$$(13) \leq \sum_{j=0}^{d-2} \sum_{t=1}^{n} C(t,j) \cdot \Big[\sum_{k=0}^{d-j-2} \binom{k+j}{k} \int_{\mathbf{R}^+} |G_n^{(k)}(x)| \mathrm{d}x \Big] +$$

$$+ C(t, d-1) \int_{\mathbf{R}^+} |G_n(x) - F_n(x)| \mathrm{d}x \leq$$

$$\leq \sum_{j=0}^{d-2} C_{n,j} \frac{n(n+1)}{2} \cdot \Big[\sum_{k=0}^{d-j-2} \binom{k+j}{k} \int_{\mathbf{R}^+} |G_n^{(k)}(x)| \mathrm{d}x \Big] +$$

$$+ C_{n,d-1} \int_{\mathbf{R}^+} |G_n(x) - F_n(x)| \mathrm{d}x. \tag{16}$$

By definition of the binomial coefficient, there exists a constant C_d depending on d such that

$$\binom{k+j}{k} \leq C_d, \qquad \forall\, j = 0, \ldots, d-2;\; k = 0, \ldots, d-j-2.$$

This fact implies that

$$(16) \leq \sum_{j=0}^{d-2} C_{n,j} \cdot C_d \frac{n(n+1)}{2} \cdot \Big[\sum_{k=0}^{d-j-2} \int_{\mathbf{R}^+} |G_n^{(k)}(x)| \mathrm{d}x \Big] +$$

$$+ C_{n,d-1} \int_{\mathbf{R}^+} |G_n(x) - F_n(x)| \mathrm{d}x. \tag{17}$$

Since $G_n \in H^{1,d-2}$, then $G_n \in H^{1,h}$, for each $h = 1, \ldots, d-2$, and we can write

$$(17) = \frac{n(n+1)C_d}{2} \cdot \sum_{j=0}^{d-2} C_{n,j} ||G_n||_{1,d-j-2} + C_{n,d-1} ||G_n - F_n||_{1,1}. \tag{18}$$

By standard properties of the Sobolev Spaces, we have

$$H^{1,d-2} \subset \ldots \subset H^{1,2} \subset H^{1,1},$$

and

$$||G_n||_{1,1} \leq ||G_n||_{1,2} \leq \ldots \leq ||G_n||_{1,d-2}.$$

Such properties give a further estimate:

$$(18) \leq \frac{n(n+1)C_d ||G_n||_{1,d-2}}{2} \cdot \sum_{j=0}^{d-2} C_{n,j} + C_{n,d-1} ||G_n - F_n||_{1,1} \leq$$

$$\leq \frac{n(n+1)C_d ||G_n||_{1,d-2}}{2} \cdot (d-1) C_{n,d-2} + C_{n,d-1} ||G_n - F_n||_{1,1}. \tag{19}$$

A long but easy computation shows that, for each $\epsilon_2 > 0$, there exists $n_{\epsilon_2} \in \mathbf{N}$ such that, $\forall\, j$,

$$n > n_{\epsilon_2} \Rightarrow |C_{n,j} - (2^j - 1)^{2n-2}| \sim |C_{n,j} - 4^{jn}| < \epsilon_2. \tag{20}$$

Therefore, for each $n > \max\{n^1_{\epsilon_1}, n^2_{\epsilon_1}, n_{\epsilon_2}\}$ we have the following approximation:

$$(19) \sim \frac{(d-1)C_d}{2} \cdot n^2 4^{n(d-2)} ||G_n||_{1,d-2} + 4^{n(d-1)} ||G_n - F_n||_{1,1}.$$

By the hypotheses, $\forall\, \epsilon > 0$, $\exists\, n_\epsilon \in \mathbf{N}$ such that, for $n > n_\epsilon$,

$$\frac{(d-1)C_d}{2} \cdot n^2 4^{n(d-2)} ||G_n||_{1,d-2} + 4^{n(d-1)} ||G_n - F_n||_{1,1} < \epsilon.$$

The result is proved, by choosing $n_\epsilon > \max\{n^1_{\epsilon_1}, n^2_{\epsilon_1}, n_{\epsilon_2}\}$. ■

Theorem 3.1 is a key result to define two suitable random matrices, that are related to the stationary and the nonstationary terms of the process. These random matrices are assumed to be dependent on an integer number $m \geq p$.

Given $\mu = 1, \ldots, m$, let us consider

$$M^{NS}_{\mu,n} = \frac{1}{n} \sum_{j=0}^{d-1} \sum_{t=1}^{n} \Delta^j Y_t \cdot \Big[\sum_{k=0}^{d-j-1} \binom{k+j}{k} \Delta^k G_{\mu,n}(t) \Big],$$

with $G_{\mu,n}$ (and related $F_{\mu,n}$) as the functions G_n (and F_n) described in Theorem 3.1.

We define

$$A_m := \sum_{\mu=1}^{m} a_{\mu,n} a^T_{\mu,n} \tag{21}$$

and

$$B_m := \sum_{\mu=1}^{m} b_{\mu,n} b^T_{\mu,n}, \tag{22}$$

where

$$a_{\mu,n} := \frac{M^{NS}_{\mu,n}/\sqrt{n}}{\sqrt{\int_{(\zeta_n,\xi_n)} \int_{(\zeta_n,\xi_n)} F_{\mu,n}(x) F_{\mu,n}(y) \min\{x,y\} \mathrm{d}x \mathrm{d}y}} \tag{23}$$

and

$$b_{\mu,n} := \frac{\sqrt{n} M^S_{\mu,n}}{\sqrt{\int_{(\zeta_n,\xi_n)} F_{\mu,n}(x)^2 \mathrm{d}x}}, \tag{24}$$

with

$$M^S_{\mu,n} := \frac{1}{n}\sum_{t=1}^{n} F_{\mu,n}(t)\Delta^d Y_t. \tag{25}$$

The random matrices defined above are the main tools of the nonparametric analysis, that will be developed in the next section.

4 Convergence results

In this section the generalized eigenvalue problem is solved. To this end, let us assume firstly that

$$\lim_{n\to+\infty} \zeta_n = 0 \qquad \text{and} \qquad \lim_{n\to+\infty} \xi_n = 0. \tag{26}$$

We define

$$\Psi_\mu := \frac{\int_{(0,1)} F_\mu(x)W(x)\mathrm{d}x}{\sqrt{\int_{(0,1)}\int_{(0,1)} F_\mu(x)F_\mu(y)\min\{x,y\}\mathrm{d}x\mathrm{d}y}},$$

$$\Phi_\mu := \frac{F_\mu(1)W(1) - \int_{(0,1)} f_\mu(x)W(x)\mathrm{d}x}{\int_{(0,1)} F_\mu(x)^2\mathrm{d}x},$$

where f_μ is the derivative of F_μ and

$$F_\mu := \lim_{n\to+\infty} F_{\mu,n}.$$

Moreover, we define the following p-variate standard normally distributed random vectors:

$$\Psi^*_\mu := \left(R^T_{p-r}C(1)C(1)^T R_{p-r}\right)^{\frac{1}{2}} R^T_{p-r}C(1)\Psi_\mu \sim N_{p-r}(0, I_{p-r}),$$

$$\Phi^*_\mu := \left(R^T_{p-r}C(1)C(1)^T R_{p-r}\right)^{\frac{1}{2}} R^T_{p-r}C(1)\Phi_\mu,$$

$$\Phi^{**}_\mu := (R^T_r D(1)D(1)^T R_r)^{-\frac{1}{2}} R^T_r D(1)\Phi_\mu \sim N_r(0, I_r),$$

and we construct the matrix $V_{r,m}$ as

$$V_{r,m} := (R_r^T D(1) D(1)^T R_r)^{\frac{1}{2}} V^*_{r,m} (R_r^T D(1) D(1)^T R_r)^{\frac{1}{2}},$$

with

$$V^*_{r,m} = \Big(\sum_{\mu=1}^m \gamma_\mu^2 \Phi_\mu^{**} \Phi_\mu^{**T}\Big) - \Big(\sum_{\mu=1}^m \gamma_\mu \Phi_\mu^{**} \Psi_\mu^{*T}\Big)\Big(\sum_{\mu=1}^m \Psi_\mu^{*} \Psi_\mu^{*T}\Big)^{-1}\Big(\sum_{\mu=1}^m \gamma_\mu \Psi_\mu^{*} \Phi_\mu^{**T}\Big),$$

where

$$\gamma_\mu = \frac{\sqrt{\int_0^1 F_\mu^2(x) \mathrm{d}x}}{\sqrt{\int_0^1 \int_0^1 F_\mu(x) F_\mu(y) \min\{x,y\} \mathrm{d}x \mathrm{d}y}}.$$

The following result summarizes the eigenvalue problem and provide a nonparametric solution for it.

Theorem 4.1 *Assume that the following hypotheses hold.*

$$\int_{(0,1)} \int_{(0,1)} F_{\mu_1}(x) F_{\mu_2}(y) \min\{x,y\} \mathrm{d}x \mathrm{d}y = 0, \qquad \mu_1 \neq \mu_2; \tag{27}$$

$$\int_{(0,1)} F_{\mu_1}(x) \int_{(0,x)} F_{\mu_2}(y) \mathrm{d}x \mathrm{d}y = 0, \qquad \mu_1 \neq \mu_2; \tag{28}$$

$$\int_{(0,1)} F_{\mu_1}(x) F_{\mu_2}(x) \mathrm{d}x = 0, \qquad \mu_1 \neq \mu_2. \tag{29}$$

If Assumptions 2.1 and 2.2 are true, then:

(I) suppose that $\hat\lambda_{1,m} \geq \ldots \geq \hat\lambda_{p,m}$ are the ordered solutions of the generalized eigenvalue problem

$$\det\Big[A_m - \lambda(B_m + n^{-2} A_m^{-1})\Big] = 0, \tag{30}$$

and $\lambda_{1,m} \geq \ldots \geq \lambda_{p-r,m}$ the ordered solutions of

$$\det\Big[\sum_{\mu=1}^m \Psi_\mu^* \Psi_\mu^{*T} - \lambda \sum_{\mu=1}^m \Phi_\mu^* \Phi_\mu^{*T}\Big] = 0. \tag{31}$$

Then we have the following convergence in distribution

$$(\hat{\lambda}_{1,m}, \ldots, \hat{\lambda}_{p,m}) \to (\lambda_{1,m}, \ldots, \lambda_{p-r,m}, 0, \ldots, 0);$$

(II) let us consider $\lambda^*_{1,m} \geq \ldots \geq \lambda^*_{r,m}$ *the ordered solutions of the generalized eigenvalue problem*

$$\det\Big[V^*_{r,m} - \lambda(R_r^T D(1)D(1)^T R_r)^{-1}\Big] = 0. \tag{32}$$

Then the following convergence in distribution holds

$$n^2(\hat{\lambda}_{p-r+1,m}, \ldots, \hat{\lambda}_{p,m}) \to (\lambda^{*2}_{1,m}, \ldots, \lambda^{*2}_{r,m}).$$

Proof. The proof is due to Lemmas 1, 2 and 4 (Bierens, 1997), and Theorem 3.1. ■

References

Anderson, S.A., H.K. Brons & S.T Jensen (1983), Distribution of eigenvalues in multivariate statistical analysis, *Annals of Statistics* **11**, 392-415.

Bierens, H.J. (1997), Nonparametric co-integration analysis, *Journal of Econometrics* **77**, 379-404.

Ladyzhenskaya, O.A. & N.N. Uraltseva (1968), *Linear and quasilinear elliptic equations*, (Academic Press, New York-London).