Data Science in Finance and Economics

https://www.aimspress.com/journal/dsfe

DSFE, 4(1): 132–159.
DOI:10.3934/DSFE.2024005
Received: 28 December 2023
Revised: 13 February 2024
Accepted: 27 February 2024
Published: 13 March 2024

Research article

An abelian way approach to study random extended intervals and their ARMA processes

Babel Raïssa GUEMDJO KAMDEM[1], Jules SADEFO KAMDEM[2]*and Carlos OGOUYANDJOU[3]

[1] Advanced School of Economics and Commerce, University of Douala, Cameroon
[2] Faculté d'Economie et MRE UR 209, Université de Montpellier, France
[3] Institut de Mathématiques et de Sciences Physiques, Université Abomey-Calavi, Bénin

* **Correspondence:** Email: jules.sadefo-kamdem@umontpellier.fr.

Abstract: An extended interval is a range $A = [\underline{A}, \overline{A}]$ where $\underline{A}$ may be bigger than $\overline{A}$. This is not really natural, but is what has been used as the definition of an extended interval so far. In the present work we introduce a new, natural, and very intuitive way to see an extended interval. From now on, an extended interval is a subset of the Cartesian product $\mathbb{R} \times \mathbb{Z}_2$, where $\mathbb{Z}_2 = \{0, 1\}$ is the set of directions; the direction 0 is for increasing intervals, and the direction 1 for decreasing ones. For instance, $[3, 6] \times \{1\}$ is the decreasing version of $[6, 3]$. Thereafter, we introduce on the set of extended intervals a family of metrics d_γ, depending on a function $\gamma(t)$, and show that there exists a unique metric d_γ for which $\gamma(t)dt$ is what we have called an "adapted measure". This unique metric has very good properties, is simple to compute, and has been implemented in the software *R*. Furthermore, we use this metric to define variability for random extended intervals. We further study extended interval-valued ARMA time series and prove the Wold decomposition theorem for stationary extended interval-valued times series.

Keywords: random set; random extended interval; distance; measure; time series

JEL Codes: C15, C22, C53, C58, C59

1. Introduction

Intervals analysis (see Bauch and Neumaier (1992); Moore (1966); Jaulin et al., (2001); Alefeld and Herzberger (2012)), initially developed in the 1960s to take into account in a rigorous way, different types of uncertainties (rounding errors due to finite precision calculations, measurement uncertainties, linearization errors), makes it possible to build supersets of the domain of variation of a real function. Coupled with the usual theorems of existence, for example, the Brouwer or Miranda theorems, the

interval theory also makes it possible to rigorously prove the existence of solutions for a system of equations (see Goldsztejn et al., (2005)). With interval analysis, it was now possible to model interval data.

In recent years, more precisely since the end of the 1980s, interval modeling has caught the attention of a growing number of researchers. The advantage of an interval-valued time series over a point-valued time series lies in that it contains both the trend (or level) information and volatility information (e.g., the range between the boundaries), while some informational loss is encountered when one uses a conventional point-valued data set, e.g., the closing prices of a stock collected at a specific time point within each time period, since it fails to record the valuable intraday information. Higher-frequency point-valued observations could result in hardly discriminating information from noises. A solution is to analyze the information in an interval format by collecting the maximum and minimum prices in a day, which avoids undesirable noises in the intraday data and contains more information than point-valued observations Sun et al. (2018). For instance, in their work Lu et al. (2022) proposed a modified threshold autoregressive interval-valued models with interval-valued factors to analyze and forecast interval-valued crude oil prices, and proved that oil price range information is more valuable than oil price level information in forecasting crude oil prices.

Huge progress in the field of interval-valued time series has been done by Billard and Diday (2000, 2003), who first proposed a linear regression model for the center points of 37 interval-valued data. They have been followed by other authors (Maia et al., (2008); Hsu and Wu (2008); Wang and Li (2011); González-Rivera and Lin (2013); Wang et al., (2016)). To study interval data, all those references apply point-valued techniques on the center, the left bound, or the right bound. By so doing, they may not efficiently make use of the information contained in interval data. In 2016, Han et al. (2016) developed a minimum-distance estimator to match the interval model predictor with the observed interval data as much as possible. They proposed a parsimonious autoregressive model for a vector of interval-valued time series processes with exogenous explanatory interval variables in which an interval observation is considered as a set of ordered numbers. It is shown that their model can efficiently utilize the information contained in interval data, and thus provides more efficient inferences than point-based data and models Han et al. (2015). As recent development in the field, one can refer to the work of Dai et al., (2023) where a new set-valued GARCH model was constructed. We also advise readers to look at the work of Wu et al., (2023).

Despite all these advances, the classical theory of interval modeling has some inconveniences. We can enumerate two which are addressed in another work and in the present paper, respectively.

First, the set of random intervals (or more generally random sets) is not a vector space. Indeed, the set of intervals is not an abelian group for the classical addition of intervals. So, all the useful theorems obtained through orthogonal projection such as the Wold decomposition theorem cannot be extended to interval-valued processes. Second, in time series, interval-valued data does not take into account some specifications or details of the study period, for instance in financial markets where a movement in stock prices during a given trading period is an observation of bounded intervals by maximum and minimum daily prices (see Han et al. (2016)). One can use two concepts to address each of these inconveniences. One can consider the set of random intervals as a ”pseudovector space” where vectors do not necessarily have opposites. This concept of a pseudovector space was developed in Kamdem et al., (2020) to address the first inconvenience stated above. The second inconvenience can be addressed by working with ”extended intervals” instead of classical intervals, as in the present paper.

Indeed, it may be more relevant to consider extended intervals formed by the opening and closing prices, regarding stock prices. Also, for the daily temperature in meteorology, instead of taking the max and min, it would be better in some cases to take the morning and evening temperature, as well as for the systolic and diastolic blood pressures in medicine. For this last example of blood pressure, when plotting the blood pressure of somebody as extended intervals of morning and evening records, one can easily see days where the morning blood pressure was higher than the evening one, which can indicate illness or emotional issues.

Therefore, given the constraints imposed by classical interval theory and its application on time series, our approach is based on the concept of extended or generalized intervals for which the left bound is not necessarily less than the right one. This generalization makes our modeling approach relevant for time series analysis. This generalization guarantees the completeness of interval space and consistency between interval operations. Extended intervals are also used for time series analysis in Han et al. (2012), but their approach does not highlight the advantages of generalized interval-valued variables.

Our contribution is therefore both theoretical and empirical. In other words, we have conceptualized and redefined some of the specific characteristics of the set of extended intervals. More precisely, we define on the set of extended intervals, a topology which generalizes the natural topology on the set of classical interval, unlike the topology introduced by Ortolf (1969) on generalized intervals, which restricted on classical interval is different from the natural topology.

The rest of the work is organized as follows: The main purpose of Section 2 is to fix notations, and give a novel and consistent definition of extended intervals. In Section 3 we introduce a suitable class of distances on the set of random extended intervals, which solves a disadvantage of the Hausdorff. We use this new distance to define the variance and covariance of random extended intervals and we show that they share some useful properties with point-valued random variables, (see propositions 3.3 and 3.4). Section 4 is concerned with stationary extended interval-valued time series, and ARMA model are investigated. In Section 5, we prove the Wold decomposition version of extended interval-valued time series. Section 6 is about numerical studies. In this section we present an algorithm to convert efficiently point-valued data to extended interval-valued data. We make a simulation of an I-AR(1) process and illustrate the interpretation of a plot of extended intervals on a few data on blood pressure. We also do empirical analysis and forecasting of the French CAC 40 market index from June 1st to July 26, 2019. The paper ends with a short conclusion.

2. Extended intervals

In this section, we first recall some basic concepts related to standard intervals. Next, we define what is meant by "extended interval", and we introduce the set $\mathbb{R}_{\leftarrow}$ of real numbers traveled in the reverse direction as a Cartesian product. At the end of this section, we present a novel representation of extended intervals.

Let $K_{kc}(\mathbb{R})$ be the set of nonempty compact (and convex) intervals. For $A = [a_1, a_2], B = [b_1, b_2] \in K_{kc}(\mathbb{R})$, and $\lambda \in \mathbb{R}$, we recall the operations

$$A + B = [a_1 + b_1, a_2 + b_2] \tag{2.1}$$

$$\lambda A = \begin{cases} [\lambda a_1, \lambda a_2] & \text{if } \lambda \geq 0 \\ [\lambda a_2, \lambda a_1] & \text{if } \lambda \leq 0 \end{cases}. \tag{2.2}$$

It is noteworthy that $K_{kc}(\mathbb{R})$ is closed under those operations, but it is not a vector space, since $A + (-1)A$ is not necessarily $\{0\}$, unless $A = \{0\}$. The Hausdorff distance d_H is defined for closed intervals $[a_1, a_2]$ and $[b_1, b_2]$ by

$$d_H([a_1, a_2], [b_1, b_2]) = max(|b_1 - a_1|, |b_2 - a_2|).$$

It is well-known that $(K_{kc}(\mathbb{R}), d_H)$ is a complete metric space (see Yang and Li (2005) for details). For $A \in K_{kc}(\mathbb{R})$, the support function of A is the function $s(\cdot, A) : \mathbb{R} \to \mathbb{R}$ defined by

$$s(x, A) = \sup\{ax \ ; \ a \in A\}. \tag{2.3}$$

Equivalently, if we set $A = [a_1, a_2]$,

$$s(x, A) = \max(xa_1, xa_2).$$

Keep in mind that $s(x, A)$ returns x times the left bound of A when x is negative, and x times the right bound of A when x is positive. This observation will be used to extend the support function on "extended closed intervals".

Definition 1. *An **extended interval** is a range A of real numbers between $\underline{A}$ and $\overline{A}$, with $\underline{A}, \overline{A} \in \mathbb{R} \cup \{\pm\infty\}$, traveled through from $\underline{A}$ to $\overline{A}$.*

The difference with standard intervals is that, for extended intervals, we do not impose that $\underline{A} \leq \overline{A}$, but the running direction is important. We say that A is an **increasing extended interval** or a **proper interval** when $\underline{A} < \overline{A}$, a **decreasing extended interval** or an **improper interval** when $\underline{A} > \overline{A}$, and a **degenerate interval** when $\underline{A} = \overline{A}$. When $\underline{A}$ and $\overline{A}$ are in A, we say that A is an extended closed interval and denote it by $A = \lfloor \underline{A}, \overline{A} \rfloor$. We also have extended open intervals $\rfloor \underline{A}, \overline{A} \lfloor$, $\mathbb{R} = \rfloor -\infty, \infty \lfloor$ and $\mathbb{R}_{\leftarrow} := \rfloor \infty, -\infty \lfloor$.

Every non-degenerate extended interval A represents the classical interval from $\min(\underline{A}, \overline{A})$ to $\max(\underline{A}, \overline{A})$ in the increasing direction (for an increasing extended interval) or in the decreasing direction (for a decreasing extended interval). We call $\underline{A}$ the left bound and $\overline{A}$ the right bound of the extended interval A.

2.1. A new way to see bounded extended intervals

A bounded extended interval can be seen as a subset of the product set

$$\mathbb{R}_{\rightleftarrows} := \mathbb{R} \times \mathbb{Z}_2 = \mathbb{R} \times \{0, 1\} =: \mathbb{R} \times \{+, -\}.$$

An element of $\mathbb{R}_{\rightleftarrows}$ is then a pair (x, α) where $x \in \mathbb{R}$, and the direction $\alpha \in \{0, 1\}$. In this structure, we have two kinds of degenerate extended intervals, namely $\{a\}^+ := \{a\} \times \{0\}$ and $\{a\}^- := \{a\} \times \{1\}$. A decreasing extended interval (when $\underline{A} > \overline{A}$) is written as $\rfloor \underline{A}, \overline{A} \lfloor := [\overline{A}, \underline{A}] \times \{1\}$, and an increasing interval (when $\underline{A} < \overline{A}$) as $\rfloor \underline{A}, \overline{A} \lfloor := [\underline{A}, \overline{A}] \times \{0\}$.

Thus, $\mathbb{R}_{\rightleftarrows}$ is the set of real numbers $\mathbb{R}$ endowed with two directions represented by the elements of the Abelian group $\mathbb{Z}_2$. The direction 0 (or +) means you move on the real line from the left to the right, and the direction 1 (or −) means you move from the right to the left. Further, the product $[2, 4] \times \{0, 1\}$ is the subset of $\mathbb{R}_{\rightleftarrows}$ in which one can move either from 2 to 4 or from 4 to 2. Equivalently, $[2, 4] \times \{0, 1\} = ([2, 4] \times \{0\}) \cup ([2, 4] \times \{1\})$.

We denote $[a,b]\times\{0\}$ by $[a,b]^+$, or just $[a,b]$, and $[a,b]\times\{1\}$ by $[a,b]^-$. Also, we denote $(x,0)$ by x^+ or just x, and $(x,1)$ by x^-. For instance, $3\in[2,4]$ and $3\notin[2,4]^-$, while $3^-\notin[2,4]$ and $3^-\in[2,4]^-$.

Practically, talking about the French CAC40 index, if we say that we got 4922^- today, this will mean that we got a value of 4922 and the index was decreasing when we got this value. This is an example of how this new structure of extended intervals can be very useful in the context of the trading market, and more.

The best choice of topology on the second member $\{0,1\}$ of $\mathbb{R}_{\rightleftarrows}$ is the discrete topology: every subset is open. So, if we also endow $\mathbb{R}$ with its natural topology, the only compact and convex subset for the product topology in $\mathbb{R}_{\rightleftarrows}$, are the closed extended intervals $\lfloor\underline{A},\overline{A}\rfloor$.

We need now to clarify how to compute the intersection of extended intervals with our notations. First observe that $A\subseteq B$ means that $\underline{B}\le\underline{A}\le\overline{A}\le\overline{B}$ or $\underline{B}\ge\underline{A}\ge\overline{A}\ge\overline{B}$. For instance, $\lfloor1,2\rfloor\not\subseteq\lfloor3,1\rfloor$. In fact, the elements of $\lfloor1,2\rfloor$ are $1^+, 1.2^+, 1.5^+$, and so on, and do not belong to $\lfloor3,1\rfloor=[1,3]^-$. The only obstruction for the inclusion to hold in this example is the difference in the running direction between both intervals.

Proposition 2.1. *Let A and B be two compact extended intervals. If A and B are running in opposite directions, then $A\cap B=\emptyset$. Otherwise, the intersection $A\cap B$ is the biggest extended interval C such that $C\subseteq A$ and $C\subseteq B$. This is naturally extended to general subsets A and B.*

Example 2.1. $\lfloor0,1\rfloor\cap\lfloor1,2\rfloor=\{1\}$, $\lfloor1,0\rfloor\cap\lfloor2,1\rfloor=\{1\}_{\leftarrow}$, $\lfloor0,1\rfloor\cap\lfloor2,1\rfloor=\emptyset$, $\lfloor2,1\rfloor\cap\lfloor3,1\rfloor=\lfloor2,1\rfloor$, $\lfloor3,1\rfloor\cap\lfloor4,2\rfloor=\lfloor3,2\rfloor$, $\mathbb{R}\cap\mathbb{R}_{\leftarrow}=\emptyset$.

Now that union and intersection are well defined for subsets of $\mathbb{R}_{\rightleftarrows}$, one can define topologies on the latter.

Definition 2. *The natural topology of $\mathbb{R}_{\rightleftarrows}$ is the topology generated by the set of extended open intervals.*

The topology induced on $\mathbb{R}$ by $\mathbb{R}_{\rightleftarrows}$ coincides with the natural topology of $\mathbb{R}$. We denote by $\mathcal{K}(\mathbb{R})$ the set of all compact extended intervals, except decreasing degenerate extended intervals. That means that all degenerate intervals in $\mathcal{K}(\mathbb{R})$ are increasing. We extend the Hausdorff distance on $\mathcal{K}(\mathbb{R})$ as

$$d_H(A,B)=max(|\underline{A}-\underline{B}|,|\overline{A}-\overline{B}|). \tag{2.4}$$

Example 2.2. *In $\mathcal{K}(\mathbb{R})$, the extended closed intervals $\lfloor\underline{A},\overline{A}\rfloor$ and $\lfloor\overline{A},\underline{A}\rfloor$ are different, unless $\underline{A}=\overline{A}$, and $d_H(\lfloor\underline{A},\overline{A}\rfloor,\lfloor\overline{A},\underline{A}\rfloor)=|\overline{A}-\underline{A}|$. This distance can be viewed as the effort needed to turn $\lfloor\underline{A},\overline{A}\rfloor$ into $\lfloor\overline{A},\underline{A}\rfloor$.*

It is simple to see that each extended interval $A\in\mathcal{K}(\mathbb{R})$ is uniquely defined by the restriction of its support function on $\{-1,1\}$. Moreover, the map $(\mathcal{K}(\mathbb{R}),d_H)\to\left(\mathbb{R}^{\{-1,1\}},d_{max}\right)$ is an isometry. (To be precise, d_{max} here is the maximum distance given by $d_{max}(f,g)=\max(|g(-1)-f(-1)|,|g(1)-f(1)|)$.) Thus, the following result is a consequence of the completeness of $\left(\mathbb{R}^{\{-1,1\}},d_{max}\right)$.

Theorem 2.1. *$(\mathcal{K}(\mathbb{R}),d_H)$ is a complete metric space.*

We endow $\mathcal{K}(\mathbb{R})$ with the topology induced by the Hausdorff distance d_H. We extend multiplication (2.2) on extended intervals in such a way that multiplication of an increasing extended interval by a negative number gives a decreasing extended interval and vice versa. This ensure the consistency of the extensions on $\mathcal{K}(\mathbb{R})$ of the internal composition laws (2.1)–(2.2):

$$\lambda A=\lfloor\lambda\underline{A},\lambda\overline{A}\rfloor,\quad A+B=\lfloor\underline{A}+\underline{B},\overline{A}+\overline{B}\rfloor,\qquad A-B=\lfloor\underline{A}-\underline{B},\overline{A}-\overline{B}\rfloor,\qquad\forall\lambda\in\mathbb{R}. \tag{2.5}$$

The operator $-$ can be seen as an extension of the difference of Hukuhara defines for standard intervals by $A - B = [\min(\underline{A} - \underline{B}, \overline{A} - \overline{B}), \max(\underline{A} - \underline{B}, \overline{A} - \overline{B})]$. It is simple to see that $(\mathcal{K}(\mathbb{R}), +, \cdot)$ is a vector space and $0 := [0, 0]$ is the zero vector.

For extended closed intervals A and B the support function reads

$$s_A(u) = \begin{cases} \sup\{ux;\ x \in A\} & \text{if } \underline{A} \leq \overline{A} \\ \inf\{ux;\ x \in A\} & \text{if } \overline{A} < \underline{A} \end{cases}. \tag{2.6}$$

For instance, $s_A(-1) = -\underline{A}$ and $s_A(1) = \overline{A}$. Hence, the support function from the vector space of extended closed intervals to the vector space $\mathbb{R}^{\{-1,1\}}$ of maps from $\{-1, 1\}$ to $\mathbb{R}$, is linear. That is, for all compact extended intervals A, B,

$$\begin{aligned} s_{A+B} &= s_A + s_B \\ s_{\lambda A} &= \lambda s_A, \qquad \forall \lambda \in \mathbb{R} \\ s_{A-B} &= s_A - s_B. \end{aligned}$$

For any extended interval A, we call the vector of s_A the column vector $S_A = (-s_A(-1), s_A(1))'$.

3. Extended interval-valued random variables

Let $(\Omega, \mathscr{A}, P)$ be a probability space. For any $A \in \mathcal{K}(\mathbb{R})$, we set

$$hits(A) = \{B \in \mathcal{K}(\mathbb{R}); A \cap B \neq \emptyset\}$$

as set of compact extended intervals that hit A. We endow the set $\mathcal{K}(\mathbb{R})$ of compact extended intervals with the $\sigma-$algebra $\mathfrak{B}(\mathcal{K}(\mathbb{R}))$ generated by $\{hits(A);\ A \in \mathcal{K}(\mathbb{R})\}$. For simplicity, we denote $X^{-1}(hits(A)) := \{\omega \in \Omega;\ X(\omega) \cap A \neq \emptyset\}$ by $X^{-1}(A)$ and call it the inverse image of A by X. This inverse image $X^{-1}(A)$ is the collection of $\omega \in \Omega$ such that $X(\omega)$ hits A. The following three definitions are equivalent to the ones given in Han et al. (2012).

Definition 3. *A random extended interval on a probability space $(\Omega, \mathscr{A}, P)$ is a map $X : \Omega \to \mathcal{K}(\mathbb{R})$ such that, for any $A \in \mathcal{K}(\mathbb{R})$, $X^{-1}(A) \in \mathscr{A}$.*

So, a random extended interval is a measurable map $X : \Omega \to \mathcal{K}(\mathbb{R})$ from the underlying probability space to $\mathcal{K}(\mathbb{R})$, endowed with the $\sigma-$algebra $\mathfrak{B}(\mathcal{K}(\mathbb{R}))$. We denote by $\mathcal{U}[\Omega, \mathcal{K}(\mathbb{R})]$ the set of random extended intervals. $\mathcal{U}[\Omega, \mathcal{K}(\mathbb{R})]$ inherits from the vector space structure of $\mathcal{K}(\mathbb{R})$. The distribution of $X \in \mathcal{U}[\Omega, \mathcal{K}(\mathbb{R})]$ is the map $P_X : \mathfrak{B}(\mathcal{K}(\mathbb{R})) \to [0, 1]$ defined on $O \in \mathfrak{B}(\mathcal{K}(\mathbb{R}))$ by

$$P_X(O) := P(X \in O).$$

Definition 4. *A map $f : \Omega \to \mathbb{R}$ is called a **selection map** for a random extended interval X when $f(\omega) \in X(\omega)$ for almost every $\omega \in \Omega$.*

Selection maps for $X = \lfloor \underline{X}, \overline{X} \rfloor$ are then maps leaving between $\underline{X}$ and $\overline{X}$. For instance, $\underline{X}$ and $\overline{X}$ are selection maps for X. The expectation of X is the set of expectations of measurable selection maps for X. More precisely:

Definition 5. *The **expectation** of a random extended interval X on a probability space* $(\Omega, \mathscr{A}, P)$ *is the extended interval*

$$E[X] = \lfloor E[\underline{X}], E[\overline{X}] \rfloor. \tag{3.7}$$

Proposition 3.2. *For any* $X, Y \in \mathcal{U}[\Omega, \mathcal{K}(\mathbb{R})]$ *and* $\lambda \in \mathbb{R}$, $E[X + \lambda Y] = E[X] + \lambda E[Y]$.

We denote by $\mathcal{S}_X = \{f \in L^1(\Omega)$ such that f is a selection map for $X\}$ the set of integrable selection maps for X and $\mathcal{S}_X(\mathscr{A}_0) = \{f \in L^1(\Omega, \mathscr{A}_0)$ such that f is a selection map for $X\}$ the set of $(\Omega, \mathscr{A}_0)-$integrable selection maps for X, where $\mathscr{A}_0$ a sub$-\sigma-$field of $\mathscr{A}$. The expectation of X is the classical interval $\{E[f]\ ;\ f \in S_X\}$ together with the running direction coming from X.

3.1. The distance D_γ

To quantify the variability of X, that is the dispersion of X around its expectation, we need a suitable distance measure on random extended intervals. The first distance that could come to mind is the Hausdorff distance. But, a disadvantage of the Hausdorff distance is for instance that $d_H([0, 2], [5, 6]) = 5 = d_H([0, 2], [5, 7])$, while intuitively the distance between $[0, 2]$ and $[5, 6]$ should be less than the distance between $[0, 2]$ and $[5, 7]$.

In Bertoluzza et al. (1995), the authors defined the squared distance $d^2_\gamma(A, B)$ between two standard intervals as follow: For any interval $A = [\underline{A}, \overline{A}]$, we consider the one-to-one map $\nabla_A : [0, 1] \to A$, $t \mapsto t\underline{A} + (1 - t)\overline{A}$. Then, the squared distance $d^2_\gamma(A, B)$ is given by

$$d^2_\gamma(A, B) = \int_0^1 (\nabla_A(t) - \nabla_B(t))^2\, \gamma(t)dt = \int_0^1 \left(t(\underline{A} - \underline{B}) + (1 - t)(\overline{A} - \overline{B})\right)^2 \gamma(t)dt, \tag{3.8}$$

where $\gamma(t)dt$ is a Borel measure on $[0, 1]$ such that:

$$\gamma(t) \geq 0 \text{ for every } t \in [0, 1]; \tag{3.9a}$$

$$\int_0^1 \gamma(t)dt = 1; \tag{3.9b}$$

$$\gamma(t) = \gamma(1 - t); \tag{3.9c}$$

$$\gamma(0) > 0 \tag{3.9d}$$

We extend d_γ on extended intervals with the same formula (3.8) and assumptions (3.9a)-(3.9d). If $d^2_\gamma(A, B) = 0$, then $\nabla_A(t) = \nabla_B(t)$ for almost every $t \in [0, 1]$, which implies that $\underline{A} = \underline{B}$ and $\overline{A} = \overline{B}$; thus $A = B$. For triangle inequality, we first write

$$(\nabla_A(t) - \nabla_C(t))^2 = (\nabla_A(t) - \nabla_B(t))^2 + (\nabla_B(t) - \nabla_C(t))^2 + 2(\nabla_A(t) - \nabla_B(t))(\nabla_B(t) - \nabla_C(t)).$$

Hence,

$$d^2_\gamma(A, C) = d^2_\gamma(A, B) + d^2_\gamma(B, C) + 2\int_0^1 (\nabla_A(t) - \nabla_B(t))(\nabla_B(t) - \nabla_C(t))\gamma(t)dt. \tag{3.10}$$

From here, using Hölder's inequality, one gets the triangle inequality. Thus, d_γ is a distance on the set $\mathcal{K}(\mathbb{R})$ of extended intervals. The two extended intervals $A = \lfloor \underline{A}, \overline{A} \rfloor$ and $\tilde{A} = \lfloor \overline{A}, \underline{A} \rfloor$ represent the same standard interval but are different in $\mathcal{K}(\mathbb{R})$, and $d_\gamma(A, \tilde{A}) = |\underline{A} - \overline{A}|cst$ (with $cst = \left(\int_0^1 (2t - 1)^2\gamma(t)dt\right)^{1/2} \neq 0$) vanishes if and only if $\underline{A} = \overline{A}$. This distance can be seen as the effort needed to turn $\tilde{A}$ into A.

Conditions (3.9a)–(3.9b) are required if we want the distance d_γ on degenerate intervals $[a, a]$ and $[b, b]$ to give the usual distance $|b - a|$. On other hand, the distance d_γ is suitable for intervals since it does not share some disadvantages of the Hausdorff distance, see Bertoluzza et al. (1995) for more details.

The norm of an interval A is the distance between A and 0: $\|A\| = d_\gamma(A, 0)$. Condition (3.9c) means that there is no preferable position between left and right bounds. More precisely, this condition implies that $\|\|\lfloor a, 0\rfloor\|\| = \|\|\lfloor 0, a\rfloor\|\| = |a|\left(\int_0^1 t^2\gamma(t)dt\right)^{1/2}$. The previous observation justifies the following definition.

Definition 6. *We say that $\gamma(t)dt$ is an adapted measure if, in addition to conditions* (3.9a)–(3.9d) *one has*

$$\int_0^1 t^2\gamma(t)dt = 1 \tag{3.9f}$$

Example 3.3. *One can check that, with*

$$\gamma(t) = t(1-t)\left(480 - \frac{10240}{3\pi}\sqrt{t(1-t)}\right) + 1,$$

$\gamma(t)dt$ is an adapted measure. We will refer to this as the standard adapted measure. It has been used in the software R Core Team (2021) to check Lemma 3.1.

Generally, for any $c \in (0, \infty)$, the formula

$$\gamma_c(t) = t(1-t)\left(a + b\sqrt{t(1-t)}\right) + c,$$

defines an adapted measure for $a = -30c + 510$ and $b = \frac{512(c-21)}{3\pi}$.

This d_γ distance can be related to the D_K distance measure developed by Körner and Näther (2002) as follows:

$$\begin{aligned} d_\gamma^2(A, B) &= (s_A(-1) - s_B(-1))^2K(-1,-1) + (s_A(1) - s_B(1))^2K(1,1) \\ &\quad - 2(s_A(-1) - s_B(-1))(s_A(1) - s_B(1))K(-1,1) \\ &= \begin{pmatrix} -s_A(-1) + s_B(-1) \\ s_A(1) - s_B(1) \end{pmatrix}' \begin{pmatrix} K(-1,-1) & K(-1,1) \\ K(1,-1) & K(1,1) \end{pmatrix} \begin{pmatrix} -s_A(-1) + s_B(-1) \\ s_A(1) - s_B(1) \end{pmatrix} \\ d_\gamma^2(A, B) &= S'_{A-B}\mathcal{K}_\gamma S_{A-B} \end{aligned} \tag{3.11}$$

where the kernel $\mathcal{K}_\gamma = (K(i, j))_{i,j=-1,1}$ introduced by Han et al. (2012) is given by

$$\begin{cases} K(-1,-1) = \int_0^1 t^2\gamma(t)dt \\ K(1,1) = \int_0^1 (1-t)^2\gamma(t)dt \\ K(-1,1) = K(1,-1) = \int_0^1 t(1-t)\gamma(t)dt \end{cases}. \tag{3.12}$$

We will often denote $\langle S_{A-B}, S_{A-B}\rangle_\gamma := d_\gamma^2(A, B)$. As observed before by Han et al. (2012), the kernel $\mathcal{K}_\gamma$ is symmetric positive definite and defines an inner product on $\mathcal{K}(\mathbb{R})$. We use some properties of this inner product in order to perform the proofs of Lemma 3.2 and Theorem 3.2. The following lemma shows that there exists a unique distance d_γ with $\gamma(t)dt$ an adapted measure. This lemma is also useful for numerical simulations.

Lemma 3.1. *All adapted measures induce the same metric given by*

$$\mathcal{K}_\gamma = \begin{pmatrix} 1 & -1/2 \\ -1/2 & 1 \end{pmatrix} \quad \textit{and} \quad d_\gamma^2(A,B) = (\underline{A}-\underline{B})^2 + (\overline{A}-\overline{B})^2 - (\underline{A}-\underline{B})(\overline{A}-\overline{B}).$$

Proof. If $\gamma(t)dt$ is an adapted measure, then $K(1,1) = K(-1,-1) = \int_0^1 t^2\gamma(t)dt = 1$. Using conditions (3.9a)-(3.9d), one shows that $K(-1,1) = K(1,-1) = -1/2$.

Let X and Y be two random intervals. For any $\omega \in \Omega$, $X(\omega)$ and $Y(\omega)$ are two extended intervals and one can compute the distance $d_\gamma(X(\omega), Y(\omega))$. We defined a new distance on random extended intervals by taking the square root of the mean of the squared distance $d_\gamma^2(X(\omega), Y(\omega))$ in $(\Omega, \mathscr{A}, P)$.

Definition 7. *The D_γ distance is defined for two random extended intervals X, Y by*

$$D_\gamma(X,Y) = \left(E[d_\gamma^2(X,Y)]\right)^{1/2} = \sqrt{\int_\Omega \int_0^1 \left(\nabla_{X(\omega)}(t) - \nabla_{Y(\omega)}(t)\right)^2 \gamma(t)dt\, dP(\omega)},$$

provided the integral converges.

We denote by $\mathcal{L}^2[\Omega, \mathcal{K}(\mathbb{R})]$ the set of random extended intervals X such that $E\|X\|_\gamma^2 := E(d_\gamma^2(X,0)) = D_\gamma^2(X,0) < \infty$.

Lemma 3.2. $\mathcal{L}^2[\Omega, \mathcal{K}(\mathbb{R})]$ *is a vector space under laws* (2.5).

Proof. It is enough to show that $\mathcal{L}^2[\Omega, \mathcal{K}(\mathbb{R})]$ is a sub-vector space of $\mathcal{U}[\Omega, \mathcal{K}(\mathbb{R})]$. Let $X, Y \in \mathcal{L}^2[\Omega, \mathcal{K}(\mathbb{R})]$ and $\lambda \in \mathbb{R}$. Then, $D_\gamma(\lambda X, 0) = |\lambda| D_\gamma(X,0)$ and

$$\begin{aligned} D_\gamma^2(X+Y,0) &= E\left[S'_{X+Y}\mathcal{K}_\gamma S_{X+Y}\right] \\ &= E\left[(S_X + S_Y)'\mathcal{K}_\gamma(S_X + S_Y)\right] \\ &= D_\gamma^2(X,0) + D_\gamma^2(Y,0) + 2E\left[S'_X\mathcal{K}_\gamma S_Y\right] \\ &\leq 2D_\gamma^2(X,0) + 2D_\gamma^2(Y,0). \end{aligned}$$

The last inequality comes from the fact that, using Cauchy-Schwarz inequality,

$$2S'_X\mathcal{K}_\gamma S_Y = 2\langle S_X, S_Y\rangle_\gamma \leq 2\sqrt{\langle S_X, S_X\rangle_\gamma}\sqrt{\langle S_Y, S_Y\rangle_\gamma} \leq \langle S_X, S_X\rangle_\gamma + \langle S_Y, S_Y\rangle_\gamma$$

It is simple to see that for any $X, Y \in \mathcal{L}^2[\Omega, \mathcal{K}(\mathbb{R})]$, $0 \leq D_\gamma(X,Y) < \infty$ and the triangle inequality for D_γ follows from the one of $d\gamma$. However, D_γ is not a metric on $\mathcal{L}^2[\Omega, \mathcal{K}(\mathbb{R})]$ since $D_\gamma(X,Y) = 0$ does not imply the strict equality $X = Y$, but that they are equal almost everywhere. We denote by $L^2[\Omega, \mathcal{K}(\mathbb{R})]$ the quotient set of $\mathcal{L}^2[\Omega, \mathcal{K}(\mathbb{R})]$ under the equivalent relation "being equal almost everywhere". Then, D_γ is a metric on $L^2[\Omega, \mathcal{K}(\mathbb{R})]$. We will keep denoting any class in $L^2[\Omega, \mathcal{K}(\mathbb{R})]$ by a representative $X \in \mathcal{L}^2[\Omega, \mathcal{K}(\mathbb{R})]$.

Theorem 3.2. $\left(\mathcal{K}(\mathbb{R}), d_\gamma\right)$ *and* $\left(L^2[\Omega, \mathcal{K}(\mathbb{R})], D_\gamma\right)$ *are complete metric spaces.*

Proof. Assume that $(A_n = \lfloor \underline{A}_n, \overline{A}_n \rfloor)_n$ is a $d_\gamma-$ Cauchy sequence in $\mathcal{K}(\mathbb{R})$. Then, $(\underline{A}_n, \overline{A}_n)'_n$ is a Cauchy sequence in $\mathbb{R}^2$ and so converges, say, to $(\underline{A}, \overline{A})'$. In fact, that $d_\gamma(A_p, A_q) = S'_{A_p - A_q} \mathcal{K}_\gamma S_{A_p - A_q}$ goes to 0 as p, q go to infinity implies that $S_{A_p - A_q} = (-\underline{A}_p + \underline{A}_q, \overline{A}_p - \overline{A}_q)'$ goes to 0 as p, q go to infinity. Also, $(A_n)_n$ converges to $A = \lfloor \underline{A}, \overline{A} \rfloor$ since $d_\gamma(A_n, A) = S'_{A_n - A} \mathcal{K}_\gamma S_{A_n - A}$. Hence, $\left(\mathcal{K}(\mathbb{R}), d_\gamma\right)$ is a complete metric space. Now, assume that $(X_n = \lfloor \underline{X}_n, \overline{X}_n \rfloor)_n$ is a $D_\gamma-$Cauchy sequence in $L^2[\Omega, \mathcal{K}(\mathbb{R})]$. Then, from Fatou's lemma and Definition 7,

$$E[\liminf_{p,q\to\infty} d_\gamma^2(X_p(\omega), X_q(\omega))] \leq \liminf_{p,q\to\infty} E[d_\gamma^2(X_p(\omega), X_q(\omega))] = 0.$$

Hence, $E[\liminf_{p,q\to\infty} d_\gamma^2(X_p(\omega), X_q(\omega))] = 0$, which implies that for almost every $\omega \in \Omega$, $\liminf_{p,q\to\infty} d_\gamma^2(X_p(\omega), X_q(\omega)) = 0$. Hence there exists a subsequence $(X_{n_k}(\omega))$ which is a Cauchy sequence in the complete metric space $\left(\mathcal{K}(\mathbb{R}), d_\gamma\right)$. So, for almost every ω, $(X_{n_k}(\omega))_k$ d_γ-converges to $X(\omega) = \lfloor \underline{X}(\omega), \overline{X}(\omega) \rfloor$; setting $X(\omega)$ to be 0 for the remaining ω, one obtains an random extended interval X. As $\lim_{k\to\infty} d_\gamma^2(X_{n_k}, X) = 0$, we also have that $\lim_{k\to\infty} d_\gamma^2(X_n, X_{n_k}) = d_\gamma^2(X_n, X)$ for any n. Using Fatou's lemma again,

$$\lim_{n\to\infty} E[d_\gamma^2(X_n, X)] = \lim_{n\to\infty} E[\liminf_{k\to\infty} d_\gamma^2(X_n, X_{n_k})] \leq \lim_{n\to\infty} \liminf_{k\to\infty} E[d_\gamma^2(X_n, X_{n_k})] = 0,$$

since $\lim_{p,q\to\infty} E[d_\gamma^2(X_p(\omega), X_q(\omega))] = 0$ implies that $\lim_{n,k\to\infty} E[d_\gamma^2(X_n, X_{n_k})] = 0$.

Remark 3.1. *It is clear that the space $\mathcal{K}(\mathbb{R})$ of compact extended intervals can be identified as a $2-$dimensional vector space, and the metric d_γ can be written as*

$$d_\gamma(A, B) = \|US_Y - US_X\|,$$

where U is a matrix such that $K_\gamma = U'U$. Thus, $\left(L^2[\Omega, \mathcal{K}(\mathbb{R})], D_\gamma\right)$ is identified as the $2-$dimensional random vector space on $(\Omega, \mathscr{A}, P)$ with $D_\gamma(X, Y) = E(\|US_Y - US_X\|^2)^{1/2}$, and the previous result follows from the completeness of the $2-$dimensional random vector space.

Definition 8. *We say that a sequence (X_n) of random extended intervals converges to X in probability under the metric d_γ when $(d_\gamma^2(X_n, X))$ converges to 0 in probability, that is*

$$\forall \varepsilon > 0, \qquad \lim_{n\to\infty} P(d_\gamma^2(X_n, X) \geq \varepsilon) = 0.$$

Theorem 3.3. *A sequence (X_n) such that $\sup_n E\|X_n\| < \infty$, converges to X in $(L^2[\Omega, \mathcal{K}(\mathbb{R})], D_\gamma)$ if and only if (X_n) converges to X in probability under the metric d_γ.*

Proof. Let us assume that (X_n) converges to X, that is $(D_\gamma^2(X_n, X) = E[d_\gamma^2(X_n, X)])$ converges to 0. That means that $(d_\gamma(X_n, X))$ converges to 0 in norm L^2 in $(\Omega, \mathscr{A}, P)$, which implies that $(d_\gamma^2(X_n, X))$ converges to 0 in probability. Conversely, assume that (X_n) converges to X in probability under the metric d_γ. So, the inequality $|d_\gamma(X_n, 0) - d_\gamma(X, 0)| \leq d_\gamma(X_n, X)$ implies that $(\|X_n\|)$ converges to $\|X\|$ in probability. By Fatou's Lemma,

$$E\|X\| \leq \liminf_{n\to\infty} E\|X_n\| \leq \sup_n E\|X_n\| < \infty.$$

The inequality

$$d_\gamma^2(X_n, X) \leq 2\|X_n\|^2 + 2\|X\|^2$$

implies that $(d_\gamma(X_n, X))$ is uniformly integrable. Finally, the dominated convergence theorem implies that $(D_\gamma(X_n, X))$ converges to 0.

Corollary 3.1. *Let (X_n) be a sequence of random extended intervals such that $\sup_n E\|X_n\| < \infty$ and (λ_n) a family of nonnegative real numbers such that $\sum \lambda_n^2 < \infty$. Then, $(S_n = \sum_{i=0}^n \lambda_i X_i)$ converges in probability under the metric d_γ.*

Definition 9 (Han et al. (2012))**.** *The covariance of two random extended intervals X, Y is the real*

$$\begin{aligned} Cov(X, Y) &:= E\langle S_{X-E[X]}, S_{Y-E[Y]}\rangle_\gamma \\ &= \int_\Omega \int_0^1 \left(\nabla_{X(\omega)}(t) - \nabla_{E[X]}(t)\right)\left(\nabla_{Y(\omega)}(t) - \nabla_{E[Y]}(t)\right)\gamma(t)dt\, dP(\omega). \end{aligned} \tag{3.13}$$

The variance of X is the real

$$Var(X) = Cov(X, X) = E\langle S_{X-E[X]}, S_{X-E[X]}\rangle_\gamma = D_\gamma^2(X, E[X]). \tag{3.14}$$

The next proposition is the extended interval version of Theorem 4.1 of Yang and Li (2005).

Proposition 3.3. *For all random extended intervals X, Y, Z, the following hold:*

① $Var(C) = 0$*, for every constant interval C;*

② $Var(X + Y) = Var(X) + 2Cov(X, Y) + Var(Y)$*;*

③ $Cov(X, Y) = Cov(Y, X)$*;*

④ $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$*;*

⑤ $Cov(\lambda X, Y) = \lambda Cov(X, Y)$*;*

⑥ $Var(\lambda X) = \lambda^2 Var(X)$*, for every $\lambda \in \mathbb{R}$;*

⑦ $P(d_\gamma(X, E[X]) \geq \varepsilon) \leq Var(X)/\varepsilon^2$ *for every $\varepsilon > 0$ (Chebyshev inequality).*

Proof. For any constant extended interval C, one has $E[C] = C$ and $Var(C) = 0$ follows. Using the linearity of S and the form (3.11) of the metric d_γ, one proves items ②-⑥. The Chebyshev inequality follows from the fact that $P(d_\gamma(X, E[X]) \geq \varepsilon) \leq E[d_\gamma(X, E[X])^2]/\varepsilon^2$.

In the particular case of adapted measures, we have the following results, which are very useful in numerical simulations.

Proposition 3.4. *If $\gamma(t)dt$ is an adapted measure, a, b are random variables, and X is a random extended interval, then*

① $Var(\lfloor a, 0\rfloor) = Var(\lfloor 0, a\rfloor) = Var(a)$*;*

② $Var(\lfloor a, a\rfloor) = Var(a)$*;*

③ $Cov(\lfloor a, 0\rfloor, \lfloor 0, b\rfloor) = -\frac{1}{2}Cov(a, b)$;

④ $Var(X) = Var(\underline{X}) - Cov(\underline{X}, \overline{X}) + Var(\overline{X})$;

⑤ $Cov(X, Y) = Cov(\underline{X}, \underline{Y}) + Cov(\overline{X}, \overline{Y}) - \frac{1}{2}Cov(\underline{X}, \overline{Y}) - \frac{1}{2}Cov(\underline{Y}, \overline{X})$;

⑥ $E\|X\|^2 = E[\underline{X}^2] + E[\overline{X}^2] - E[\underline{X}\overline{X}]$.

Item ⑤ of the above proposition is similar to the one obtained for classical intervals in Example 4.1 of Yang and Li (2005), but the two last terms $-\frac{1}{2}Cov(\underline{X}, \overline{Y}) - \frac{1}{2}Cov(\underline{Y}, \overline{X})$ are not present in the formula of Yang and Li. This difference can be explained by the fact that, for our distance d_γ, there is no preference between the left and the right bound, which is not the case for the distance d_p used by Yang and Li (2005). From the formula of Yang, if the left bounds of X, Y are independent and their right bounds are also independent then $Cov(X, Y) = 0$, which is not the case for our formula ⑤ above.

Let $\mathcal{L}^2[\Omega, \mathcal{K}(\mathbb{R})]_0 = \{X \in \mathcal{U}[\Omega, \mathcal{K}(\mathbb{R})], E[X] = 0, \text{ and } E[\|X\|_\gamma^2] < \infty\}$, that is, the sub-vector space of $\mathcal{L}^2[\Omega, \mathcal{K}(\mathbb{R})]$ made by random extended interval with mean zero. For a random extended interval $X \in \mathcal{L}^2[\Omega, \mathcal{K}(\mathbb{R})]_0$, $Cov(X, X) = 0$ means that $X = E[X] = 0$ almost everywhere. Hence, formula (3.13) cannot define a scalar product on $\mathcal{L}^2[\Omega, \mathcal{K}(\mathbb{R})]_0$. We denote by $L^2[\Omega, \mathcal{K}(\mathbb{R})]_0$ the set of classes of zero mean random extended intervals equal almost everywhere. We will keep denoting any class in $L^2[\Omega, \mathcal{K}(\mathbb{R})]_0$ by a representative $X \in \mathcal{L}^2[\Omega, \mathcal{K}(\mathbb{R})]_0$. $L^2[\Omega, \mathcal{K}(\mathbb{R})]_0$ inherits from the structure of the vector space of $\mathcal{L}^2[\Omega, \mathcal{K}(\mathbb{R})]_0$, and for $X, Y \in L^2[\Omega, \mathcal{K}(\mathbb{R})]_0$, the formula (3.13) reads

$$Cov(X, Y) = E\langle S_X, S_Y\rangle_\gamma = \int_\Omega \int_0^1 \nabla_{X(\omega)} \nabla_{Y(\omega)} \gamma(t) dt dP(\omega) \tag{3.15}$$

and is a scalar product on $L^2[\Omega, \mathcal{K}(\mathbb{R})]_0$.

Theorem 3.4. *$(L^2[\Omega, \mathcal{K}(\mathbb{R})]_0, Cov)$ is a Hilbert space.*

Proof. From what is written above, Cov is a scalar product on $L^2[\Omega, \mathcal{K}(\mathbb{R})]_0$. For the completeness, use the fact that $\langle , \rangle_\gamma$ defined a scalar product on $\mathbb{R}^2$.

Example 3.4. *Take $\Omega = \mathbb{R}$, $\mathscr{A}$ the Borel topology, and $P = dx$ the Borel measure. Let us consider the random extended interval*

$$X = \lfloor f(\omega), g(\omega)\rfloor, \tag{3.16}$$

the left and right bounds respectively,

$$f(\omega) = (1/\sqrt{2\pi}) \exp(-0.5\omega^2)$$
$$g(\omega) = 0.3 \exp(-0.3\omega).$$

We may write $X \rightsquigarrow \mathcal{NE}(0, 1, 0.3)$ to say that the left bound of X follows the standard normal distribution and its right bound follows the exponential distribution with parameter 0.3*. The density functions of those random variables have been plotted on Figure 1.*

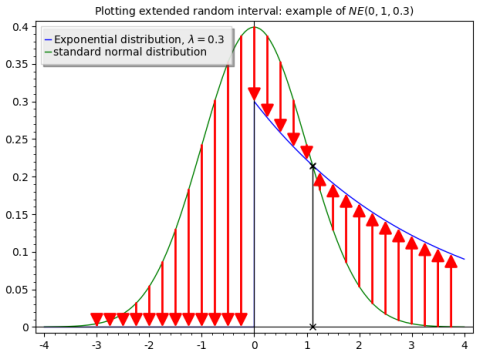


Figure 1. We represent extended intervals with arrows. An arrow pointing up for increasing extended intervals, and down for decreasing extended intervals.

4. Stationary extended interval time series

Let $(X_t)_{t\in\mathbb{Z}}$ be an extended interval time series; that is, for any integer t, X_t is an random extended interval. We denote by A_t the expectation of X_t and by $C_t(j) = Cov(X_t, X_{t-j})$ the auto-covariance function.

Definition 10. *We say that an extended interval time series (X_t) is stationary when neither A_t nor $C_t(j)$ depends on t. In this case, we just denote them A and $C(j)$, respectively.*

For any $n \in \mathbb{Z}^+$, the auto-covariance matrix is given by

$$\mathbf{C}_n = (C(i-j))_{1\le i,j\le n} = \begin{pmatrix} C(0) & C(1) & \cdots & C(n-1) \\ C(1) & C(0) & \cdots & C(n-2) \\ \vdots & \vdots & \vdots & \vdots \\ C(n-1) & C(n-2) & \cdots & C(0) \end{pmatrix}. \tag{4.17}$$

The proof of the following theorem is similar to the one of Theorem 4 in Wang et al., (2016).

Theorem 4.5. *The auto-covariance function of any stationary process satisfies:*

① $C(k) = C(-k)$ *for all* $k \in \mathbb{Z}$;

② $|C(k)| \le C(0)$ *for all* $k \in \mathbb{Z}$;

③ *the auto-covariance matrix* $\mathbf{C}_n$ *is positive semi-definite;*

④ *if* $C(0) > 0$ *and* $(C(k))$ *converges to* 0 *then* $\mathbf{C}_n$ *is positives definite.*

Let $X_1, \ldots, X_T$ be a sample of a stationary extended interval time series (X_t) with expectation A. An unbiased estimator of A is given by

$$mX = \frac{X_1 + \cdots + X_T}{T} \tag{4.18}$$

and the sample-covariance is given by

$$\widehat{C}(k) = \frac{1}{T}\sum_{i=1}^{T-|k|}\int_0^1 (\nabla_{X_{i+|k|}}(t) - \nabla_{mX}(t))(\nabla_{X_i}(t) - \nabla_{mX}(t))\gamma(t)dt. \tag{4.19}$$

Theorem 4.6. *Let (X_t) be a stationary extended interval-valued time series with expectation A and auto-covariance function $C(k)$ such that $(C(k))$ converges to 0. Then, mX is a consistent estimator of A; that is, for any $\varepsilon > 0$, $\lim_{T\to\infty} P(d_\gamma(mX, A) \geq \varepsilon) = 0$.*

Proof. One has

$$Var(mX) = D_\gamma^2(mX, A) = E\langle S_{mX-A}, S_{mX-A}\rangle_\gamma = \frac{1}{T^2}\sum_{i,j=1}^{T} E\langle S_{X_i-A}, S_{X_j-A}\rangle_\gamma$$
$$= \frac{1}{T^2}\sum_{i,j=1}^{T} C(i-j) = \frac{1}{T^2}\sum_{i-j=-T}^{T}(T-|i-j|)C(i-j) = \frac{1}{T}\sum_{k=-T}^{T}\left(1-\frac{k}{T}\right)C(k).$$

So, $Var(mX)$ goes to 0 as T goes to infinity since $(C(k))$ converges to 0. By the Chebyshev inequality, $\forall\varepsilon > 0$, $P(d_\gamma(m, A) \geq \varepsilon) \leq Var(mX)/\varepsilon^2$ goes to 0 as T goes to infinity.

As usual, $\widehat{C}(k)$ is not an unbiased estimator of $C(k)$ (unless $mX = A$), but:

Theorem 4.7. *If $(C(k))$ converges to 0 as k goes to infinity, then for any k, $\widehat{C}(k)$ is an asymptotically unbiased estimator of $C(k)$, that is $\lim_{T\to\infty} E[\widehat{C}(k)] = C(k)$.*

Proof.

$$\widehat{C}(k) = \frac{1}{T}\sum_{i=1}^{T-|k|}\int_0^1 (\nabla_{X_{i+|k|}}(t) - \nabla_{mX}(t))(\nabla_{X_i}(t) - \nabla_{mX}(t))\gamma(t)dt$$
$$= \frac{1}{T}\sum_{i=1}^{T-|k|}\int_0^1 (\nabla_{X_{i+|k|}}(t) - \nabla_A(t))(\nabla_{X_i}(t) - \nabla_A(t))\gamma(t)dt + \frac{1}{T}\sum_{i=1}^{T-|k|}\int_0^1 (\nabla_{mX}(t) - \nabla_A(t))^2\gamma(t)dt$$
$$- \frac{1}{T}\sum_{i=1}^{T-|k|}\int_0^1 (\nabla_{mX}(t) - \nabla_A(t))(\nabla_{X_{i+|k|}}(t) + \nabla_{X_i}(t) - 2\nabla_A(t))\gamma(t)dt$$

Hence,

$$\lim_{T\to\infty} E[\widehat{C}(k)] = \lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{T-|k|} E[C(k)] + \lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{T-|k|} Var(mX)$$
$$- \lim_{T\to\infty}\frac{1}{T}\sum_{i=1}^{T-|k|}(Cov(mX, X_{i+|k|}) + Cov(mX, X_i))$$
$$= C(k) - \lim_{T\to\infty}\frac{1}{T^2}\sum_{i=1}^{T-|k|}\sum_{j=1}^{T}\left(Cov(X_j, X_{i+|k|}) + Cov(X_j, X_i)\right)$$
$$= C(k) - \lim_{T\to\infty}\frac{1}{T^2}\sum_{j-i=-T}^{T}(T-|j-i|)\left(C(j-i-|k|) + C(j-i)\right)$$
$$= C(k) - \lim_{T\to\infty}\frac{1}{T}\sum_{l=-T}^{T}\left(1-\frac{|l|}{T}\right)(C(l-|k|) + C(l)) = C(k)$$

4.1. Extended Interval-valued AutoRegressive Moving-Average process

Let (X_t) be an extended interval-valued stationary time series with expectation A, and auto-covariance function $C(k)$. To capture the dynamics of (X_t) one can assume that it follows an **interval autoregressive moving-average (I-ARMA)** process of order (p, q), that is

$$X_t = K + \sum_{i=1}^{p} \theta_i X_{t-i} + \varepsilon_t + \sum_{i=1}^{q} \phi_i \varepsilon_{t-i}, \tag{4.20}$$

Where K is a constant extended interval, ϕ_i and θ_i are the parameters of the model, $(\varepsilon_t) \rightsquigarrow IID(\{0\}, \sigma^2)$, and for each t, ε_t is uncorrelated with the past of X_t. This model was introduced and studied by Han et al. (2012). They call such a model an Autoregressive Conditional Interval Model, and they proposed a D_K-distance based estimation method to estimate the parameters. Our interest in this method is to do forecasting and we propose a different estimation method, based on the Yule-Walker equation.

By taking expectation at the both sides of (4.20) one finds

$$\lambda A = K, \tag{4.21}$$

where $\lambda = 1 - \theta_1 - \cdots - \theta_p$. So, as in the case of real random variables, the expectation μ_t of X_t does not depend on t and the new series $X'_t = X_t - \frac{1}{\lambda}K$ is a zero-mean I-ARMA process, i.e. Equation (4.20) with $K = 0$. In what follows, until the numerical study section, we assume that $K = 0$, that is, (X_t) is a zero-mean stationary process. When $p = 0$, the process (X_t) is called an extended interval-valued moving-average time series process of order q, I-MA(q), and when $q = 0$, one obtains an extended interval-valued autoregressive time series process of order p, I-AR(p).

Let L be the delay operator, thus $LX_t = X_{t-1}$. Setting $\Theta(L) = 1 - \theta_1 L - \cdots - \theta_p L^p$ and $\Phi(L) = 1 + \phi_1 L + \cdots + \phi_q L^q$, equation (4.20) can be written as

$$\Theta(L)X_t = \Phi(L)\varepsilon_t. \tag{4.22}$$

The functions Θ and Φ are called the autoregressive and moving-average polynomials, respectively.

In particular, if (X_t) is an I-MA(1) process: $X_t = \varepsilon_t + \phi\varepsilon_{t-1}$, then

$$C(1) = \phi\sigma^2. \tag{4.23}$$

In section 5 we show that any *non-deterministic* zero-mean stationary random extended interval process can be expressed as a $MA(\infty)$.

If the moving-average polynomial $\Phi = 1$, then (4.22) leads to

$$X_t = (1 - \Theta(L))X_t + \varepsilon_t, \tag{4.24}$$

which is an extended interval-valued autoregressive process of order p, I-AR(p). In this case, the existence and the uniqueness of a stationary solution is not guaranteed. However, when a stationary solution exits, using Proposition 3.3 it is simple to show that its auto-covariance function satisfies

$$C(k) - \sum_{i=1}^{p} \theta_i C(k-i) = 0, \text{ for any } 1 \le k \le p. \tag{4.25}$$

Hence, the parameters of an I-AR(p) process satisfy the Yule-Walker equation

$$\mathbf{C}_p\mathbf{\Theta} = \mathbf{c}_p, \tag{4.26}$$

where $\mathbf{c}_p = (C(1), \ldots, C(p))^T$, $\mathbf{\Theta} = (\theta_1, \ldots, \theta_p)^T$ and $\mathbf{C}_p$ is the auto-covariance matrix (4.17).

Theorem 4.8. *Any AR(1) process $X_t = \theta X_{t-1} + \varepsilon_t$, with $0 < \theta < 1$ and $\sup_t E\|\varepsilon_t\| < \infty$, possesses a unique stationary solution given by $X_t = \sum_{i=0}^{\infty} \theta^i \varepsilon_{t-i}$.*

Proof. One has

$$X_t = \theta X_{t-1} + \varepsilon_t = \theta^2 X_{t-2} + \theta\varepsilon_{t-1} + \varepsilon_t = \theta^{n+1} X_{t-n-1} + \sum_{i=0}^{n} \theta^i \varepsilon_{t-i}.$$

As $0 < \theta < 1$ one has that $\sum \theta^{2i} < \infty$. This together with $\sup_t E\|\varepsilon_t\| < \infty$ implies that $(S_n = \sum_{i=0}^{n} \theta^i \varepsilon_{t-i})$ converges in probability under the metric d_γ by Corollary 3.1. Since (X_t) is stationary, $Var(X_t) = E\|X_t\|^2$ is constant and

$$E\left\|X_t - \sum_{i=0}^{n} \theta^i \varepsilon_{t-i}\right\|^2 = E\|\theta^{n+1} X_{t-n-1}\|^2 = \theta^{2(n+1)} E\|X_{t-n-1}\|^2$$

goes to 0 as n goes to infinity. Hence, $E\left\|X_t - \sum_{i=0}^{\infty} \theta^i \varepsilon_{t-i}\right\|^2 = 0$. This implies that $X_t = \sum_{i=0}^{\infty} \theta^i \varepsilon_{t-i}$ a.e. From this solution, we have

$$Cov(X_{t+k}, X_t) = \sigma^2 \sum_{i=k}^{\infty} \theta^k \theta^{i-k} = \sigma^2 \frac{\theta^k}{1-\theta^2}.$$

Now, if (X_t) is an I-ARMA(1, 1) process: $X_t = \theta X_{t-1} + \varepsilon_t + \phi\varepsilon_{t-1}$, then

$$C(2) = \theta C(1) \qquad \text{and} \qquad C(1) = \theta C(0) + \phi\sigma^2. \tag{4.27}$$

5. Wold decomposition for extended interval-valued time series

Let $(X_t)_{t\in\mathbb{Z}}$ be a zero-mean extended interval-valued stationary process. The sets $S_t = \overline{S\,pan(\{X_k\}_{k=-\infty}^{t})}$ and $S_{-\infty} = \bigcap_{t=-\infty}^{\infty} S_t$ are Hilbert spaces of $L^2[\Omega, \mathcal{K}(\mathbb{R})]_0$. For any $j \geq 0$, the projection $P_{S_{t-j}} X_t$ of X_t on S_{t-j} is called the prediction of X_t on S_{t-j}. We shall say that an extended interval-valued process $(X_t)_{t\in\mathbb{Z}}$ is **deterministic** if for any $t \in \mathbb{Z}$, $X_t \in S_{t-1}$. $X_t - P_{S_{t-1}} X_t$ is called the error in the projection of X_t on S_{t-1} and when $P_{S_{t-1}} X_t = X_t$ one says that $(X_t)_{t\in\mathbb{Z}}$ is (perfectly) predictable. As $(L^2[\Omega, \mathcal{K}(\mathbb{R})]_0, Cov)$ is a Hilbert space, we have the following Wold decomposition for extended interval time series.

Theorem 5.9. *Let $(X_t)_{t\in\mathbb{Z}}$ be a non-deterministic extended interval-valued stationary time series process with expectation $\{0\}$ and auto-covariance function $(C(k))$. Then, X_t can be expressed as*

$$X_t = \sum_{k=0}^{\infty} \alpha_k \varepsilon_{t-k} + W_t \;\; a.s \tag{5.28}$$

where:

(i) $\alpha_k = \frac{1}{\sigma^2} Cov(X_t, \varepsilon_{t-k}), \quad \alpha_0 = 1$ *and* $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$*;*

(ii) $\{\varepsilon_t\} \rightsquigarrow WN(\{0\}, \sigma^2)$*, with* $\sigma^2 = Var(X_t - P_{S_{t-1}} X_t)$*;*

(iii) $Cov(W_t, \varepsilon_s) = 0$ *for all* $t, s \in \mathbb{Z}$*;*

(iv) $(W_t)_{t\in\mathbb{Z}}$ *is zero-mean, stationary and deterministic.*

Proof. For any $t \in \mathbb{Z}$, the application of Theorem 4 in Bierens (2012) to the regular sequence $(X_{t-k})_{k=0}^{\infty}$ gives that X_t can be expressed as

$$X_t = \sum_{k=0}^{\infty} \theta_k e_{t-k} + W_t \ \text{ a.s} \tag{5.29}$$

where $\{e_{t-k}\}_{k=0}^{\infty}$ is an uncorrelated process with $Cov(e_i, e_j) = \delta_{ij}$, $\theta_k = Cov(X_t, e_{t-k})$, and $\sum_{k=1}^{\infty} \theta_k^2 < \infty$, $W_t \in U_t^{\perp}$ with $U_t = \overline{S\,pan(\{e_k\}_{k=-\infty}^{t})} \subset S_t$. Since the process $(X_t)_{t\in\mathbb{Z}}$ is non-deterministic, the residual $\varepsilon_t = X_t - P_{S_{t-1}} X_t$ is different from 0 and $\varepsilon_t = \|\varepsilon_t\| e_t$, hence (5.28) holds with $\alpha_k = \theta_k / \|\varepsilon_{t-k}\|$, and (ε_t) is also uncorrelated. As $W_t, \varepsilon_t \in L^2[\Omega, \mathcal{K}(\mathbb{R})]_0$, and $E[W_t] = 0 = E[\varepsilon_t]$. $W_t \in U_t^{\perp}$ implies that $Cov(W_t, \varepsilon_s) = 0$ for any $s \le t$. For $s > t$, taking the scalar product of (5.29) with ε_s, one has $Cov(W_t, \varepsilon_s) = Cov(X_t, \varepsilon_s) = 0$ since $\varepsilon_s \in S_{s-1}^{\perp}$ and $X_t \in S_t \subset S_{s-1}$ for $s > t$. This proves **(iii)**. Let $X_{t,n}$ be the projection of X_t on $S_{t,n} = span(\{X_{t-j}\}_{j=1}^{n})$, and $\varepsilon_{t,n}$ the residual. Then, $X_{t,n}$ takes the form

$$X_{t,n} = \sum_{j=1}^{n} \beta_{j,n} X_{t-j},$$

where the scalars $\beta_{k,n}$ do not depend on t, since they are solutions of the system of equations

$$\sum_{j=1}^{n} \beta_{j,n} C(j-k) = C(k), \qquad k = 1, \ldots, n.$$

Hence, $E[X_{t,n}] = 0$, $E[\varepsilon_{t,n}] = 0$. Moreover,

$$\begin{aligned} Var(\varepsilon_{t,n}) = \|X_t - X_{t,n}\|^2 &= \left\| X_t - \sum_{j=1}^{n} \beta_{j,n} X_{t-j} \right\|^2 \\ &= C(0) + \sum_{i,j=1}^{n} \beta_{i,n} \beta_{j,n} C(i-j) - 2 \sum_{j=1}^{n} \beta_{j,n} C(j). \end{aligned}$$

Hence, $Var(\varepsilon_{t,n}) = \sigma_n$ does not depend on t and same for $\sigma = \|\varepsilon_t\| = \lim_{n\to\infty} \sigma_n$. Also,

$$Cov(X_{t+k}, \varepsilon_{t,n}) = C(k) - \sum_{j=1}^{n} \beta_{j,n} C(k+j),$$

which does not depend on t. Using the Cauchy-Schwarz inequality,

$$\lim_{n\to\infty} |Cov(X_{t+k}, \varepsilon_{t,n} - \varepsilon_t)| \le \sqrt{C(0)} \lim_{n\to\infty} \|\varepsilon_{t,n} - \varepsilon_t\| = 0,$$

which implies that $Cov(X_{t+k}, \varepsilon_t) = \lim_{n\to\infty} Cov(X_{t+k}, \varepsilon_{t,n})$ and does not depend on t. So,

$$\alpha_k = \frac{1}{\|\varepsilon_t\|} Cov(X_{t+k}, e_k) = \frac{1}{\|\varepsilon_t\|^2} Cov(X_{t+k}, \varepsilon_t)$$

does not depend on t. Moreover, $\alpha_0 = \frac{Cov(X_t,\varepsilon_t)}{\|\varepsilon_t\|^2} = 1$. All this completes the proof of **(i)** and **(ii)**. For $k \geq 0$,

$$\begin{aligned} Cov(W_t, W_{t-k}) &= Cov\left(X_{t-k} - \sum_{j=0}^{\infty} \alpha_j \varepsilon_{t-k-j}, X_t - \sum_{j=0}^{\infty} \alpha_j \varepsilon_{t-j}\right) \\ &= C(k) - \sum_{j=0}^{\infty} \alpha_j Cov(X_t, \varepsilon_{t-k-j}) - \sum_{j=k}^{\infty} \alpha_j Cov(X_{t-k}, \varepsilon_{t-j}) + \sigma^2 \sum_{j=0}^{\infty} \alpha_{j+k}\alpha_j \\ &= C(k) - \sigma^2 \sum_{j=0}^{\infty} \alpha_{j+k}\alpha_j, \end{aligned}$$

which does not depend on t. As $W_t \in S_t$, one can write $W_t = \sum_{k=0}^{\infty} a_k X_{t-k}$. Taking the covariance with ε_t and using the fact that $\varepsilon_t \perp Span(X_{t-1}, X_{t-2}, \ldots)$, one gets $Cov(W_t, \varepsilon_t) = a_0 Cov(X_t, \varepsilon_t) = a_0\|\varepsilon_t\|^2$. Since $Cov(W_t, \varepsilon_t) = 0$, one deduces that $a_0 = 0$, hence $W_t \in S_{t-1}$, and thus (W_t) is deterministic from the past of (X_t). This completes the proof of **(iv)**.

6. Numerical studies

Let (X_t) be an AR(1) process:

$$X_t = K + \theta X_{t-1} + \varepsilon_t. \tag{6.30}$$

Then, from the Yule-Walker equation, the parameter θ can be estimated by $\widehat{\theta} = \frac{\widehat{C}(1)}{\widehat{C}(0)}$ with

$$\begin{aligned} \widehat{C}(0) &= \frac{1}{T}\sum_{i=1}^{T} \int_0^1 (\nabla_{X_i} - \nabla_{mX})^2 \gamma(t)dt = \frac{1}{T}\sum_{i=1}^{T} d_\gamma^2(X_i, mX), \\ \widehat{C}(1) &= \frac{1}{T}\sum_{i=1}^{T-1} \int_0^1 (\nabla_{X_{i+1}} - \nabla_{mX})(\nabla_{X_i} - \nabla_{mX})\gamma(t)dt \\ &= \frac{1}{2T}\sum_{t=1}^{T-1} \left(d_\gamma^2(X_{i+1}, mX) + d_\gamma^2(X_i, mX) - d_\gamma^2(X_{i+1}, X_i)\right), \end{aligned}$$

where $\widehat{C}(1)$ and $\widehat{C}(0)$ are the sample-covariance.

More generally, if we assume that the I-AR(p) process (4.24) is stationary, then from Theorem 4.5, when $C(0) > 0$ and $(C(k))$ converges to 0, the Yule-Walker equation (4.26) is well-posed and from a large sample $X_1, \ldots, X_T$, the coefficients of the I-AR(p) process can be estimated by

$$\widehat{\boldsymbol{\Theta}} = \widehat{\mathbf{C}}_p \widehat{\mathbf{c}}_p.$$

Using (3.10) and (4.19), the sample-covariance can be written as

$$\widehat{C}(k) = \frac{1}{2T} \sum_{i=1}^{T-|k|} \left(d_\gamma^2(X_{i+k}, mX) + d_\gamma^2(X_i, mX) - d_\gamma^2(X_{i+k}, X_i) \right). \tag{6.31}$$

It is natural to assume that $\gamma(t)dt$ is an adapted measure and, in this case, the distance d_γ is given by Lemma 3.1 and is easy to numerically compute.

6.1. Using extended intervals to display data efficiently

Extended intervals can be very useful for displaying data. The plot of just one extended interval A gives much informations: (a) the range of values of the considered index during the recording; (b) the direction of variation of the considered index : decreasing when the arrow is pointing down, and increasing when the arrow is pointing up.

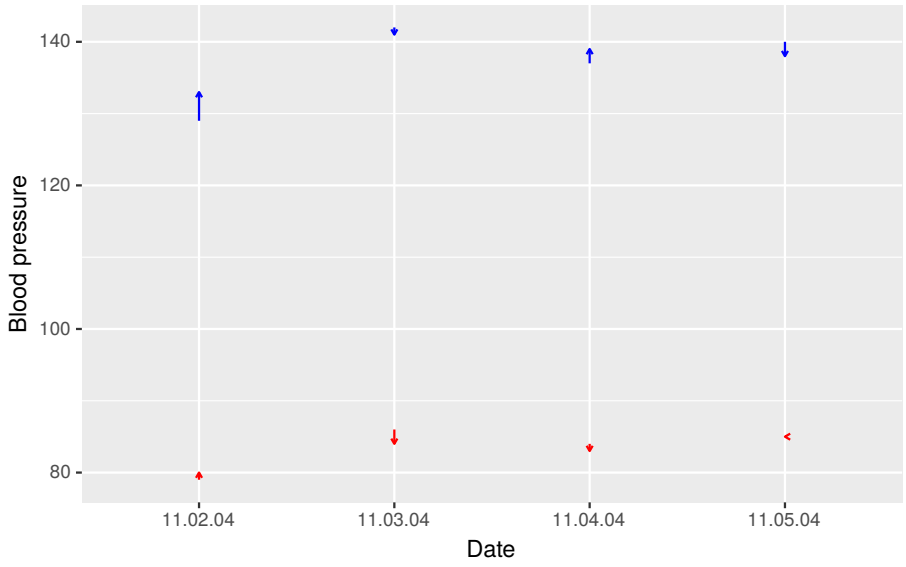


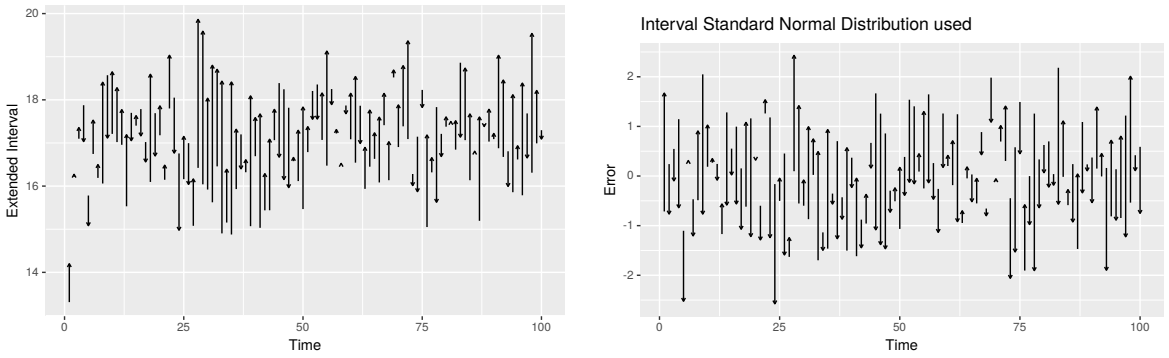
Figure 2. Systolic blood pressure in blue and diastolic blood pressure in red, of the same person, recorded over 4 days in 2004. Left bounds are the morning records and right bounds are the afternoon records.

Figure 2 displays systolic (in blue) and diastolic (in red) blood pressure of a person recorded in the morning (left bounds) and in the afternoon (right bounds), over 4 days in 2004. One sees easily that on the 11/03/04, the blood pressure recorded in the morning is higher than the one recorded in the afternoon, both for systolic and diastolic.

6.2. Simulations

Now, we plot the model (6.30) with $\theta = 0.2$, and $K = [13.31, 14.2]$, $\overline{\varepsilon_t}$ and $\underline{\varepsilon_t}$ following independent standard normal distributions. Figure 3(a) shows a sample for this model for $T = 100$, when the interval standard normal distribution used is the one plotted on Figure 3(b). One sees that most of the outputs of this sample are standard intervals (71 standard intervals versus 29 decreasing ones) while for the error

(interval standard normal distribution), they seem to be the same number (41 standard intervals versus 59 decreasing). Figure 4 displays the estimated auto-covariance function $C(k)$ and shows that it goes to 0 as k becomes large. Also, K is estimated using the formula $\widehat{K} = (1 - \widehat{\theta})mX$.



(a) (b)

Figure 3. Simulation for model (6.30) with $T = 100$

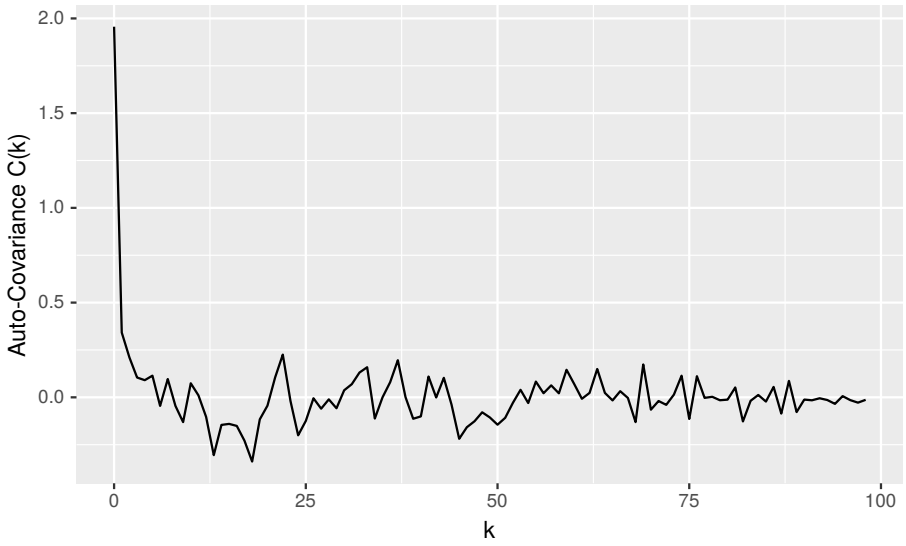


Figure 4. Auto-Covariance estimated for model (6.30) for $T = 100$

Table 1. Some estimations with R.

T	$\widehat{K}$	$C(T-2)$	$\widehat{\theta}$	Error
100	[13.31, 14.2]	–0.02807759	0.1747072	0.02529285
500	[13.51569, 14.41001]	0.01240641	0.1892873	0.01071265

6.2.1. Forecasting with Extended intervals

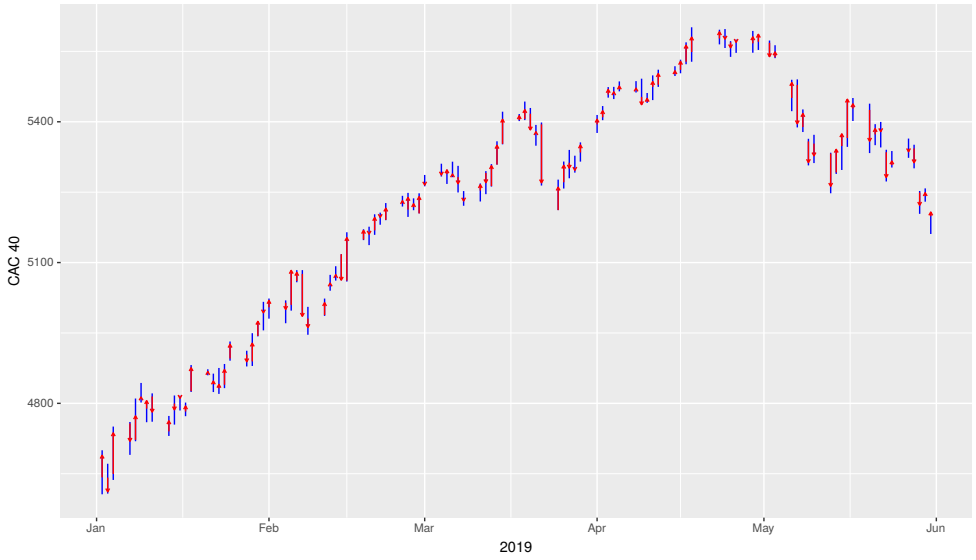


Figure 5. CAC 40 Stock Index from January 2nd to May 31st, 2019. Red arrows represent the extended intervals with left bounds the opening values (in EUR) and right bounds the closing values. The blue line segments represent the interval-valued prices composed of the lowest and highest prices of each day.

In Figure 5, we have plotted as standard min-max intervals (in blue) and open-close extended intervals (in red), the CAC 40 Stock Index from January 2nd to May 31st, 2019 (105 trading days). Extended intervals are formed by the opening values (left bounds) and the closing values (right bounds). This figure shows that most often, neither opening nor closing values are the lowest or the highest value of the index for the day. Notice that, in such an index, what is important most often is not just the opening and closing values, but also to know how it has been fluctuating along the day. For instance, the plot shows many days where the opening value and closing value are the same with fluctuation throughout the day. Now, we wish to find the I-ARMA model which best fits this data. The first step is to test stationarity. The augmented Dickey–Fuller test shows that neither the data nor its first difference are stationaries, but its second difference is stationary. So, we take the second difference data and use AIC to determine the optimal order (p, q). We define the AIC of the random interval to be the summation of the AIC of the bounds, and we assume that $p, q = 1, 2, 3, 4$. Figure 6 shows that the optimal order is $p = q = 1$. Finally, using equation (4.27), we estimated the coefficients of the I-ARMA model by $\widehat{\theta} = \frac{\widehat{C}(2)}{\widehat{C}(1)}, \widehat{\phi} = \widehat{C}(1) - \widehat{\theta C}(0)$ and we found

$$\widehat{\theta} = -0.2519991 \qquad \text{and} \qquad \widehat{\phi} = -0.5326387. \tag{6.32}$$

Figure 7 shows the forecast of the differentiated CAC 40 for the next 40 trading days. From this graph, it appears that the sense of variation of CAC 40 throughout the day has been well predicted for 25 days over 40. Also, the predicted arrow is most often on top of the real value. This prediction, for sure, can be improved by using extended intervals with non-linear estimation methods.

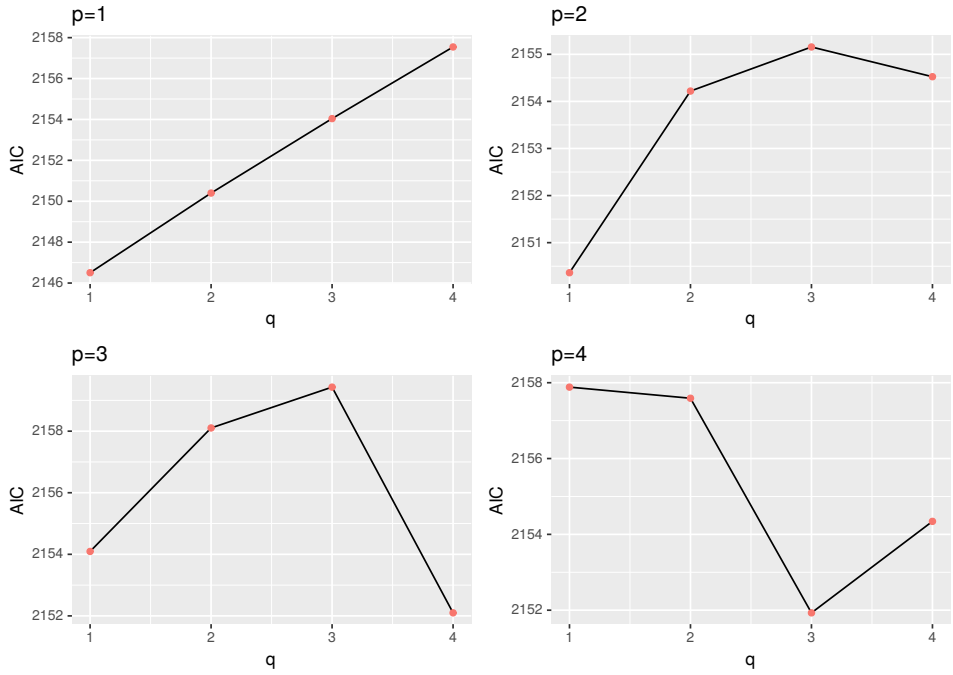


Figure 6. AIC as function of q for $p = 1, 2, 3, 4$.

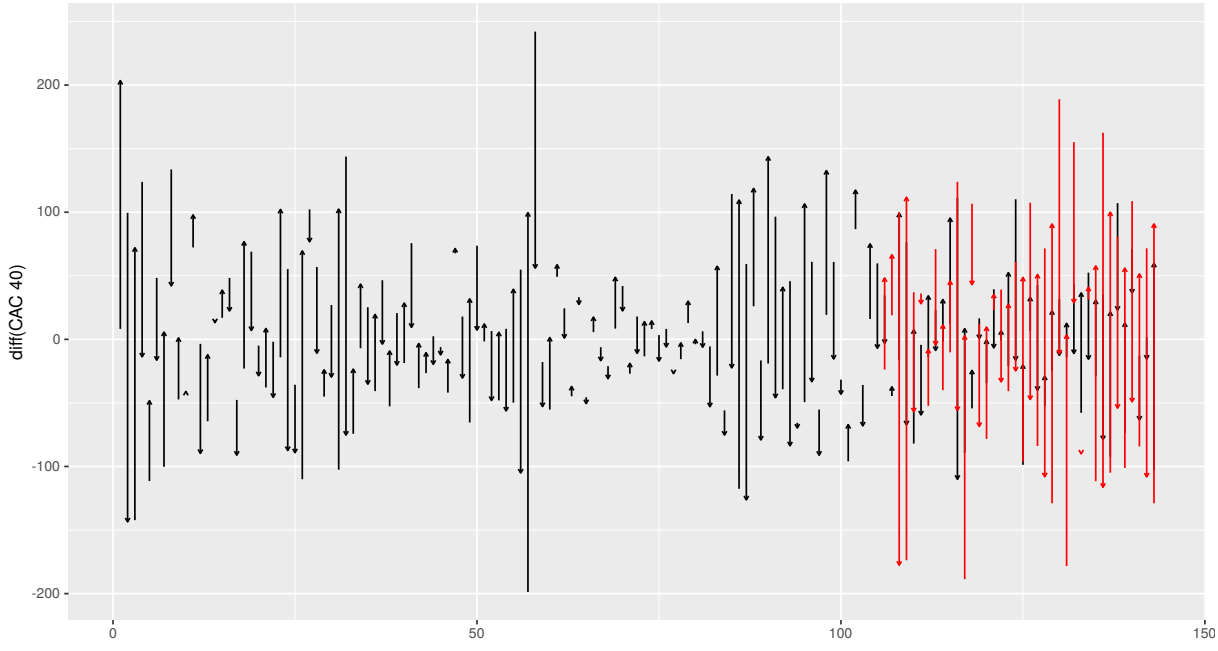


Figure 7. Forecast values from June 1st to July 26th, 2019 (red) and real values from the 2nd January to 26 July.

6.3. An algorithm to pretreat data

In this paragraph we present how data are usually pretreated and show that this process can be better performed when one wishes to use extended intervals.

Let us consider an index ID (for example the French CAC40 index) that we try to model for predicting future values. Let us assume that the values of this index are changing every minute and that we want to analyze it over one year. That will make a huge set of data to analyze if we consider every single value of the index.

What people do most often is to consider a frequency; in the case of ID, one can decide to analyze daily values. But, we have something like 1440 values every day and have to decide for the value of the day. In point-value analysis, people consider either the opening value or the closing value or the average value of the day as the value of that day. It is clear that a lot of values have been neglected and this could lead to an inconsistent analysis.

In an analysis with the standard interval, people most often consider the highest and lowest values of a day to form the interval representing the value of the index that day. (See for example, Wang and Li (2011).) By so doing, every interval contains all the values of the index that day. But, the interval can be irreasonably large and does not reflect the variations of the index during the day. One can still do better by using extended intervals.

With extended intervals, one can proceed as follows. The first value is the left bound of the first interval. If the next value is smaller (resp. bigger) then we keep looking for the next value until either the index is no longer significantly decreasing (resp. increasing), or we have passed 1440 values (the period cannot exceed 1 day). The right bound of the first interval is then the previous value recorded, and the actual value is the left bound of the second interval, and we repeat this process until the end of the data set. This process is summarized as Algorithm 1, which returns the sequence Res of extended intervals obtained and the corresponding sequence of time intervals. There is a need to explain when we say ”corresponding sequence of time intervals”. The left bound of the first time interval is the time when the left bound of the first extended interval of Res has been recorded, and so on.

By applying this algorithm, we do not have a regular period, which is needed for a time series analysis. The period can be taken here as the average of the periods of extended intervals obtained.

We have implemented Algorithm 1 in *R* and test it on the CAC 40 stock index recorded minute by minute during five days: from June 22, to June 26, 2020. After treating the 2169 data, we obtained 787 extended intervals as shown in Figure 8. The initial data was recorded everyday from 9:00 am to 6:05 pm, except the last day which ends at 10:52 am. So, the total time of recording was 38 hours and 12 minutes. As we obtained 787 extended intervals, we can take as period for time series analysis: 3 minutes. We then assume that every extended interval is recorded during a lap time of 3 minutes.

Observe that the minimum value per day as well as the maximum value of the CAC 40 during the five days we considered is the same. So, those data could not be analyzed with min-max standard intervals. In Figure 9 are have plotted the extended intervals that we obtained.

Algorithm 1 Transform point-values to extended intervals

Require: data, time, ε, frequency=1440
  $Res \leftarrow \{\}$, $ResTime \leftarrow \{\}$
  $N \leftarrow length(data)$, $i \leftarrow 1$
  $\underline{A} \leftarrow data[i]$, $\underline{T} \leftarrow time[i]$
  while $i < N$ **do**
    if $data[i+1] \leq data[i]$ **then**
      $i \leftarrow i+1$, $j \leftarrow 1$
      while the index is decreasing or is not significantly (use ε) increasing **do**
        $i \leftarrow i+1$, $j \leftarrow j+1$
        if $j > frequency$ **then**
          break
        end if
      end while
      $\overline{A} \leftarrow data[i]$, $\overline{T} \leftarrow time[i]$
      add $A = \lfloor \underline{A}, \overline{A} \rfloor$ in Res and $T = \lfloor \underline{T}, \overline{T} \rfloor$ in $Restime$
      $i \leftarrow i+1$
    end if
    if $data[i+1] > data[i]$ **then**
      $i \leftarrow i+1$, $j \leftarrow 1$
      while the index is increasing or is not significantly (use ε) decreasing **do**
        $i \leftarrow i+1$, $j \leftarrow j+1$
        if $j > frequency$ **then**
          break
        end if
      end while
      $\overline{A} \leftarrow data[i]$, $\overline{T} \leftarrow time[i]$
      add $A = \lfloor \underline{A}, \overline{A} \rfloor$ in Res and $T = \lfloor \underline{T}, \overline{T} \rfloor$ in $ResTime$
      $i \leftarrow i+1$
    end if
  end while
  return Res and $ResTime$

```
> X <- read.csv("Documents/ExtendedInterval/data/CAC40_2020.csv")
> Z <- DataToExtint(X)
2169 observations treated
From 2020-06-22 09:00:00 to 2020-06-26 10:52:00
787 extended intervals obtained:
   Contains: 285 standard intervals, 2 degenerate intervals, 500 decreasing intervals
>
```

Figure 8. Example of data pretreated in R using Algorithm 1.

Figure 9. CAC 40 Stock Index from June 22, to June 26, 2020

7. Conclusions

In this work, we have redefined extended intervals in a more natural manner and written an algorithm to efficiently transform point-valued data to extended interval-valued data. An extended interval is a standard interval endowed with a direction α, which is an element of the Abelian group $\mathbb{Z}_2 = \{0, 1\}$. The direction 0 means you move on the real line from the left to the right, and the direction 1 means you move from the right to the left. This process can be generalized on $\mathbb{R}^n$. For example, one could define extended rectangles on $\mathbb{R}^2$ with 4 directions represented by the Abelian group $\mathbb{Z}_4$.

We have seen that by using extended intervals to record the values of a given index, every extended interval gives the value of the index and the direction of variation at the time of recording. We have proposed a language that we hope will be use in the future in the trading markets. Precisely, talking about the French CAC40 index, if we say that we got 4922^- today, this would mean that we got a value of 4922 and the index was decreasing when we got this value. This is an example of how this new structure of extended intervals can be very useful in the context of trading markets, and more. A suitable distance has been defined on extended intervals and used to define variance and covariance on random extended intervals, in a natural way. We have studied ARMA processes with extended intervals both theoretically and numerically. In the numerical part, we forecasted on CAC 40 stock index from January 2nd to July 26, 2019.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

Alefeld G, Herzberger J (2012) *Introduction to interval computation*. Academic press.

Bauch H, Neumaier A (1990) Interval methods for systems of equations. cambridge university press, *Zamm-Z Angew Math Me* 72: 590–590.

Bertoluzza C, Corral Blanco N, Salas A (1995) On a new class of distances between fuzzy numbers. *Mathware soft comput* 2 .

Bierens HJ (2012) The wold decomposition. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.231.2308.

Billard L, Diday E (2000) Regression analysis for interval-valued data. In *Data Analysis, Classification, and Related Methods*, 369–374. Springer.

Billard L, Diday E (2003) From the statistics of data to the statistics of knowledge: symbolic data analysis. *J Am Stat Assoc* 98: 470–487. https://doi.org/10.1198/016214503000242

Dai X, Cerqueti R, Wang Q, et al. (2023) Volatility forecasting: a new garch-type model for fuzzy sets-valued time series. *Ann Oper Res* 1–41. https://doi.org/10.1007/s10479-023-05746-z

Goldsztejn A, Daney D, RueherM, et al. (2005) Modal intervals revisited: a mean-value extension to generalized intervals. In *Proceedings of QCP-2005 (Quantification in Constraint Programming), Barcelona, Spain*.

González-Rivera G, Lin W (2013) Constrained regression for interval-valued data. *J Bus Econ Stat* 31: 473–490. https://doi.org/10.1080/07350015.2013.818004

Han A, Hong Y, Wang S (2012) Autoregressive conditional models for interval-valued time series data. *The 3rd International Conference on Singular Spectrum Analysis and Its Applications*, 27.

Han A, Hong Y, Wang S (2015) Autoregressive conditional models for interval-valued time series data. Working Paper.

Han A, Hong Y, Wang S, et al. (2016) A vector autoregressive moving average model for interval-valued time series data. *Essays in Honor of Aman Ullah*, 417–460. Emerald Group, Publishing Limited.

Hsu HL, Wu B (2008) Evaluating forecasting performance for interval data. *Comput Math Appl* 56: 2155–2163. https://doi.org/10.1016/j.camwa.2008.03.042

Jaulin L, Kieffer M, Didrit O, et al. (2001) *Interval analysis*. *Appl Interval Anal*, Springer London. https://doi.org/10.1007/978-1-4471-0249-6_2

Kamdem JS, Kamdem BRG, Ougouyandjou C (2020) S-arma model and wold decomposition for covariance stationary interval-valued time series processes. *New Math Natl Comput* 17: 191–213. https://doi.org/10.1142/S1793005721500101

Kaucher E (1973) *Über metrische und algebraische Eigenschaften einiger beim numerischen Rechnen auftretender Räume*. na.

Körner R, Näther W (2002) On the variance of random fuzzy variables. In *Statistical modeling, analysis and management of fuzzy data*, 25–42. Springer.

Lu Q, Sun Y, Hong Y, et al. (2022) Forecasting interval-valued crude oil prices using asymmetric interval models. *Quantit Financ* 22: 2047–2061.https://doi.org/10.1080/14697688.2022.2112065

Maia ALS, de Carvalho FdA, Ludermir TB (2008) Forecasting models for interval-valued time series. *Neurocomputing* 71: 3344–3352. https://doi.org/10.1016/j.neucom.2008.02.022

Moore RE (1966) *Interval analysis Prentice-Hall Englewood Cliffs*, NJ.

Ortolf HJ (1969) *Eine Verallgemeinerung der Intervallarithmetik*. Gesellschaft für Mathematik und Datenverarbeitung.

R Core Team (2021) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sun Y, Han A, Hong Y, et al. (2018) Threshold autoregressive models for interval-valued time series data. *J Econometrics* 206: 414–446. https://doi.org/10.1016/j.jeconom.2018.06.009

Wang X, Li S (2011) The interval autoregressive time series model. In *Fuzzy Systems (FUZZ), 2011 IEEE International Conference on*, 2528–2533. IEEE.

Wang X, Zhang Z, Li S (2016) Set-valued and interval-valued stationary time series. *J Multivariate Anal* 145: 208–223. https://doi.org/10.1016/j.jmva.2015.12.010

Wu D, Dai X, Zhao R, et al. (2023) Pass-through from temperature intervals to china's commodity futures' interval-valued returns: Evidence from the varying-coefficient its model. *Financ Res Lett* 58: 104289. https://doi.org/10.1016/j.frl.2023.104289

Yang X, Li S (2005) The Dp-metric space of set-valued random variables and its application to covariances. *Int J Innov Comput Inf Contr 1*: 73–82.

AIMS Press