Missing, random and irregular data in time series: an overview.

Natalia Bahamonde

Pontificia Universidad Católica de Valparaíso, Chile.
natalia.bahamonde@pucv.cl

September 30, 2024

Why are Missing Values Special in Time Series Data?

- Missing value treatment is a key part of data preparation.
- Most often time series are accompanied by forecasting tasks and most algorithms won't allow missing data.
- Removing missing data points altogether might reduce information contained in other features.

Missing data can induce and compromise the accuracy of the analysis and forecasting models: dealing with them is critical in time series analysis in order to achieve valid results.

Key questions:

- Is there any identifiable reason for missing data?
- Are they missing at random?

Types of Missing Data

- Missing Completely at Random (MCAR): probability of missing values in a variable is the same for all samples.
- Missing at Random (MAR): missing values are randomly distributed, however, they are related to some other variables' values.
- Missing Not at Random (MNAR): the missingness of data is related to events or factors which are not measured.

Approaches to Handling Missing Values

There are several approaches to dealing with missing values; the most popular are imputation strategies, which attempt to generate a complete time series by filling in all missing points.

- Deletion
- Constant Imputation
- Last Observation Carried Forward (LOCF) and Next Observation Carried Backward (NOCB)
- Mean/Median/Mode Imputation
- Moving Average Statistics Imputation
- Linear Interpolation
- Spline Interpolation
- K-Nearest Neighbors (KNN) Imputation
- Decomposition for Time Series
- ...

Pros/Cons of imputation methods

- Pros
  - easy to implement,
  - can not require any model fitting,
  - not computationally expensive.
- Cons
  - Can lead to biased results if the missing data is not MCAR,
  - can distort statistical analysis and modeling results,
  - it can doesn't not consider the possible variability around the missing value,
  - may vary depending on the proportion of missing data,
  - the choice of window size can significantly impact the results,
  - many method may not be suitable for handling large gaps of missing data.

Effect on model-building process of imputation methods

- can reduces the size of the dataset,
- can also introduce bias in the models,
- can lead to overestimation or underestimation of the model parameters,
- can lead to an underestimate of the variance.

Challange: Not taking missing data structure into account in analysis may seriously bias inferences

Missing data

Gaps or intervals in a time series where no observations are available can occur for a variety of causes:

- data collection issues,
- technical concerns,
- machinery disorder,
- clerical error or the inability to observe data in a bad weather or holidays,
- or just a lack of records for a specific period.

Missing, random, unevenly or irregular data: incomplete data

There is a distinction between missing data and unevenly spaced data in the context of time series.

- The absence of observations at specific time points is commonly referred to as missing data.
  - this missing mechanism corresponds to the missing mechanism completely at random (MCAR),
  - there is a vast literature devoted to the case when the missing mechanism is independent of an underlying time series of interest X_t.
- Whereas irregularly spaced or unevenly data concerns observations that are not equally or consistently distributed in time.
  - this case the missing is called missing not at random (MNAR),
  - there are considerably fewer references.

Missing observations as an amplitude-modulated process

Following the pioneering publications by Parzen (1963), a time serie with missing observations can be regarded as an amplitude-modulated version of the original time series, i.e.,

$$Y_t = a_t X_t,$$

where X_t is assumed to be defined for all time, a_t is given by

$$a_t = \begin{cases} 1 & \text{if } X_t \text{ is observed,} \\ 0 & \text{if } X_t \text{ is missing,} \end{cases}$$

and Y_t represents the actually observed value of X_t, with 0 inserted in the series whenever the value of X_t is missing.

- In the time domain, asymptotic properties of non-parametric estimators of autocovariance and autocorrelation of the amplitude-modulated time series are established by Dunsmuir and Robinson (1981) and Yajima and Nishino (1999).
- These results can be used to build Yule-Walker type estimators for an AR process with missing observations.
- It is worth noticing that the asymptotic distributions derived by these authors apply to the case of linear stationary processes (X_t) with conditionally homoscedastic innovations (ε_t), i.e., (X_t) satisfies

$$X_t = \sum_{j=0}^{\infty} \beta_j \varepsilon_{t-j}, \qquad \sum_{j=0}^{\infty} \beta_j^2 < \infty,$$

where $\mathrm{E}(\varepsilon_t \mid \mathscr{F}_{t-1}^{\varepsilon}) = 0$ and $\mathrm{E}(\varepsilon_t^2 \mid \mathscr{F}_{t-1}^{\varepsilon}) = \sigma^2$, $\mathscr{F}_t^{\varepsilon}$ being the σ-algebra generated by $(\varepsilon_j)_{j \leq t}$.

Estimation of the ACF of a stationary time series with missing observations

Is one of the important aspects throughout time series model building procedure, particularly in the identification stage.
Following Parzen (1963), if we assume that $\{X_t\}$ and $\{a_t\}$ are independent hereafter.
Next, we put

$$\widetilde{\gamma}_{X,N}(\ell) = \frac{\widehat{\gamma}_{Y,N}(\ell)}{\widehat{\gamma}_{a,N}(\ell)}$$

when $\widehat{\gamma}_{a,N}(\ell) \neq 0$.

Estimators of the autocorrelation function for a stationary process with missing observations.

$$\widehat{\rho}_{PDR}(\ell) = \frac{\widehat{\gamma}_{Y,N}(\ell)/\widehat{\gamma}_{a,N}(\ell)}{\widehat{\gamma}_{Y,N}(0)/\widehat{\gamma}_{a,N}(0)} = \frac{\sum_{t=1}^{N-\ell} Y_t Y_{t-\ell}/\sum_{t=1}^{N-\ell} a_t a_{t-\ell}}{\sum_{t=1}^{N} Y_t^2/\sum_{t=1}^{N} a_t^2}. \quad (1)$$

$$\widehat{\rho}_{SST}(\ell) = \frac{\sum_{t=1}^{N-\ell} Y_t Y_{t-\ell}}{\sum_{t=1}^{N-\ell} a_{t+\ell} Y_t^2}. \quad (2)$$

$$\widehat{\rho}_{T}(\ell) = \frac{\sum_{t=1}^{N-\ell} Y_t Y_{t-\ell}}{\sqrt{\sum_{t=1}^{N-\ell} a_{t+\ell} Y_t^2}\sqrt{\sum_{t=1}^{N-\ell} a_t Y_{t+\ell}^2}}. \quad (3)$$

PDR :Proposed by Parzen (1963); Dunsmuir & Robinson (1981).

SST :Proposed by Shin & Sarkar (1995); Takeuchi (1995).

T :Proposed by Takeuchi (1995).

These results are only valable in the homocedastic case.

Missing data in economy and finance

- **Market Closures:** Data may be missing for specific time periods due to holidays, unexpected market shutdowns, or trading halts (e.g., during extreme volatility or technical issues), although economic activity continues as a product of political and social phenomena that will have a direct influence on the next value of the index under study.
- **Low Liquidity:** For less actively traded stocks or securities, there may be missing data for certain time periods when no trades occurred, especially for thinly traded markets.
- **Nowcasting:** often deals with the challenge of missing or incomplete data, because official figures are not yet fully available or there are gaps in real-time data streams. (is a real-time forecasting method used to predict the current state of an economic indicator or financial variable before the official data is available).

A first contribution: missing in ARCH(p) process

- A first idea is to contribute to the treatment of missing observations in financial time series.
- Let (X_t) be the ARCH(p) time series defined by the equation

$$X_t = \sigma_t(\alpha)\varepsilon_t, \tag{4}$$

where (ε_t) is a sequence of iid random variables with $\mathrm{E}\varepsilon_0 = 0$ and $\mathrm{E}\varepsilon_0^2 = 1$, and $(\sigma_t(\alpha))$ is a non negative process satisfying the difference equation

$$\sigma_t^2(\alpha) = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2. \tag{5}$$

Here $\alpha = (\alpha_0, \alpha_1, \ldots, \alpha_p)'$ where the parameters α_i are nonnegative and α_p is positive.

We will follow Bose & Mukherjee (2003) ideas to construct estimators of the parameters of ARCH(p) models from their quadratic representation.

We express observed data $(Y_t)_{1\leq t\leq n}$ by

$$Y_t = a_t X_t,$$

where (a_t) represents the state of observation (1: observed or 0: missing).

Throughout we make the two following assumptions :

(A1) Process (a_t) is strictly stationary and weakly mixing.

(A2) Processes (a_t) and (X_t) are independent.

Furthermore, we considered the squared representation of the process.

$$X_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + (X_t^2 - \sigma_t^2(\alpha)) = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{t-i}^2 + \sigma_t^2(\alpha)(\varepsilon_t^2 - 1) \tag{6}$$

Let $A_t = \prod_{i=0}^{p} a_{t-i}$, and

$$Y_t = A_t X_t = A_t Z_t' \alpha + A_t \sigma_t^2(\alpha)\eta_t = Z_t^{\star\prime}\alpha + \sigma_t^{\star 2}(\alpha)\eta_t, \qquad p+1 \leq t \leq n. \tag{7}$$

A preliminary LSE of α is

$$\widehat{\alpha}_{\text{pr}} = \left[\sum_{t=p+1}^{n} Z_t^{\star} Z_t^{\star\prime} \right]^{-1} \sum_{t=p+1}^{n} Z_t^{\star} Y_t^{\star}. \tag{8}$$

The consistency and asymptotic normality of $\widehat{\alpha}_{\text{pr}}$ is proved.
Like Bose and Mukherjee (2003), we use $\widehat{\alpha}_{\text{pr}}$ to construct an improved estimator $\widehat{\alpha}$ of α as follows. Dividing (7) by $\sigma_t^2(\alpha)$, we get

$$\frac{Y_t^{\star}}{\sigma_t^2(\alpha)} = \frac{Z_t^{\star\prime}}{\sigma_t^2(\alpha)} \alpha + A_t \eta_t. \tag{9}$$

In this expression, the errors $A_t \eta_t$ are homoscedastic and it is possible to establish the properties of consistency and asymptotic normality of the estimator obtained from the above expression (9).

- has a closed form expression,
- is easy to obtain,
- better performance than QMLE for small sample size.

Extensions

- A critique that has been directed towards the log-GARCH model is that its log-volatility specification does not exist in the presence of zero returns. Sucarrat and Escribano (2016) use the above results to provide a framework for observed zero as missing observations and replace them with estimates of their conditional expectation.
- Following the approach introduced by Dunsmuir and Robinson (1981), to take into account missing values in time series, Raissi (2024) study serial correlations, allowing for unconditional heteroscedasticity and time-varying probabilities of zero financial returns.
- A variant of the autoregressive conditional heteroscedastic (ARCH) model is proposed by Plaza et al. (2023), in which an additional covariate is included (sea surface temperature) and missing values are considered.

Types of datasets are encountered in astronomy and astrophysics (X-ray anda gamma-ray)

- Gaussian, regularly spaced bins.
- Gaussian, irregularly spaced observations.
- Poissonian, regularly spaced bins.
- Poissonian, irregularly spaced individual events.

Some types of variable high energy objects

- magnetic flares from normal stars,
- thermonuclear explosions on white dwarfs and neutron stars,
- supernovae from all types of exploding stars,
- quasars (quasi-stellar object) where the X-ray emission arises from a hot energetic active galactic nucleus (AGN),
- blazars AGN produced by accreting supermassive black holes with a relativistic jet aimed at Earth.

How to model astronomical time series?

- There are some techniques that fit unequally spaced time series, such as the continuous-time autoregressive moving average (CARMA) processes.
- These models are defined as the solution of a stochastic differential equation.
- It is not uncommon in astronomical time series, that the time gaps between observations are large.
- Therefore, an alternative suitable approach to modeling astronomical time series with large gaps between observations should be based on the solution of a difference equation of a discrete process.

An irregularly spaced IAR(1): Elorrieta et al. (2019)

- Let y_{t_j} an observation measured at time t_j,
- Considerer a increasing sequence of observational times $\{t_j\}$ for $j = 1, \ldots, n$.

Elorrieta et al. (2019) propose an autoregressive model for irregular discrete-time series based on the discrete time representation of the continuous autoregressive model of order 1.

$$y_{t_j} = \phi^{t_j - t_{j-1}} y_{t_{j-1}} + \sigma \sqrt{1 - \phi^{2(t_j - t_{j-1})}}\, \varepsilon_{t_j},$$

where ε_{t_j} are independent random variables with zero mean and unit variance.

An irregularly spaced IAR(1): Notes

- It is possible to establish a connection between Gaussian IAR(1) model and CAR(1).
- Observe that $E(y_{t_j}) = 0$ and $Var(y_{t_j}) = \sigma^2$ for all y_{t_j}.
- for $s < t$, $\gamma(t-s) = E(y_t y_s) = \sigma^2 \phi^{t-s}$
- If $t_j - t_{j-1} = 1$ for all j, then the model IAR becomes

$$y_{t_j} = \phi y_{t_{j-1}} + \sigma\sqrt{1-\phi^2}\varepsilon_{t_j}, \tag{10}$$

which corresponds to the autoregressive model of order 1 [AR(1)] for regularly space data.

An irregularly spaced isARMA(1,1) (Bahamonde et al., 2024)

We consider a particular case of time varying ARMA(1, 1) process $(Y_\ell)_{\ell\in\mathbb{Z}}$ such as $\text{Var}(Y_\ell) = \text{Var}(Y_0)$ for any $\ell \in \mathbb{Z}$. This so-called isARMA(1,1) process (irregularly spaced ARMA(1, 1)) is defined by

$$Y_\ell = \alpha_\ell\, Y_{\ell-1} + \zeta_\ell + \beta_\ell\, \zeta_{\ell-1}, \qquad \text{for } \ell \in \mathbb{Z}, \tag{11}$$

where

- $(\zeta_\ell)_{\ell\in\mathbb{Z}}$ is an i.i.d. centered sequence of random variables with variance $\nu_*^2 > 0$;
- There exists $\alpha_* \in (-1, 1)$ and a sequence of integer-valued random variables $(k_\ell)_{\ell\in\mathbb{Z}}$ independent of $(\zeta_\ell)_{\ell\in\mathbb{Z}}$ such that $\alpha_\ell = \alpha_*^{k_\ell}$ for any $\ell \in \mathbb{Z}$;
- $(\beta_\ell)_{\ell\in\mathbb{Z}}$ is a sequence of random variables such as

$$\beta_\ell = \text{sgn}(\alpha_\ell)\left(\sqrt{(\sigma^{*2}/\nu^{*2} - 1)(1 - \alpha_\ell^2)} - |\alpha_\ell|\right), \qquad \text{for } \ell \in \mathbb{Z}, \tag{12}$$

with $\sigma_*^2 > 0$ is such as $\nu_*^2 < \sigma_*^2 < 2\nu_*^2$.

Hence, a first consequence of this choice of (β_ℓ) required in (12) is

$$\mathbb{E}[Y_\ell] = 0 \text{ and } \mathrm{Var}(Y_\ell) = \mathrm{Var}(Y_0) = \sigma_*^2 \qquad \text{for any } \ell \in \mathbb{Z}.$$

Note also that the independence of the sequence (ζ_ℓ) implies that $\mathbb{E}[Y_\ell \zeta_\ell] = \mathbb{E}[\zeta_\ell^2] = \nu_*^2$ for $\ell \in \mathbb{Z}$, and thus $\zeta_t \sim (0, \nu_*^2 = \textit{constant})$.

Remark

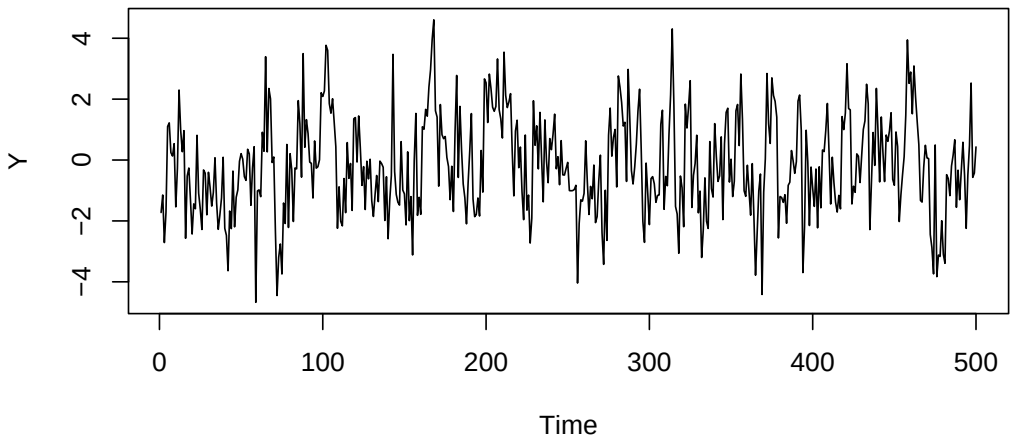
Model (11) *is inspired by classical ARMA(1,1) model*

$$X_t = \alpha X_{t-1} + \xi_t + \beta \xi_{t-1}, \quad t \in \mathbb{Z} \tag{13}$$

for i.i.d. and centered random variables (ξ_t) *with variance* ν^2*, and with* $|\alpha| < 1$ *and* $\beta \in \mathbb{R}$.
The typical situation of irregularly spaced observations will be $(X_{t_\ell})_{\ell \in \mathbb{N}}$ *for some observation times* $\ldots, t_{-2}, t_{-1}, t_0, t_1, t_2, \ldots$ *and with* $k_\ell = t_\ell - t_{\ell-1}$*. Nevertheless it provides a different model than the isARMA(1,1) process. Indeed the recursion between* X_{t_ℓ} *and* $X_{t_\ell - 1}$ *now includes a recursion with a memory* k_ℓ *: this was at the origin of choosing to develop this category of model.*

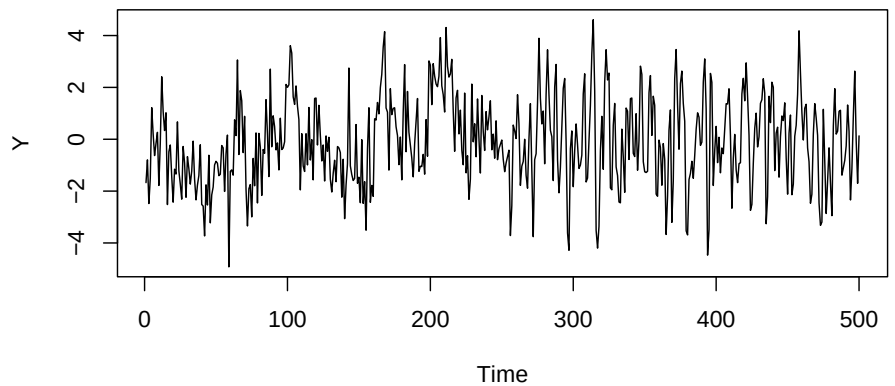
Example of trajectories of isARMA(1,1) processes

- $\alpha_* = 0.9$ using the same the same white noise (ζ_ℓ)
- Case 1 (stationary case): $(k_\ell)_\ell$ of independent random variables following a geometric distribution with parameter 0.4: this is a case where the process has random coefficients varying with time, but remains stationary.



Example of trajectories of isARMA(1,1): non stationary case

- $\alpha_* = 0.9$ using the same the same white noise (ζ_ℓ)
- Case 2 (non-stationary case): we have chosen a deterministic sequence of positive integer number $(k_\ell)_\ell$ such that $k_\ell = 1$ for $\ell \in \{1, \ldots, n/2\}$ and $k_\ell = 100$ for $\ell \in \{n/2+1, \ldots, n\}$ (then $\alpha_*^{100} \simeq 2.10^{-5}$, (Y_ℓ) is almost a MA(1) process).
- In this case, the process is no longer at all stationary, and we can see a change appearing in the plot, corresponding to a greater (at the right part of the trajectory) or lesser dependence (at the left part of the trajectory).



Study of empirical estimators of the parameters

Now, with $L \in \mathbb{N}^*$ and based on an observed trajectory $(Y_1, \ldots, Y_L)$ of an isARMA(1,1) process (Y_t) satisfying (11), assume that

- $\alpha_* \in (-1, 1)$,
- $\sigma_*^2 > 0$, and
- $\nu_*^2 = \mathrm{Var}(\zeta_0) > 0$

are unknown parameters such as $\nu_*^2 < \sigma_*^2 < 2\,\nu_*^2$.

Empirical estimators

Empirical estimators based on the model are obtained:

$$\widehat{\sigma^2} = \frac{1}{L}\sum_{\ell=1}^{L} Y_\ell^2, \qquad \widehat{\alpha}_{(j)} = \left(\frac{\widehat{\eta}_j}{\widehat{\lambda}_j}\right)^{1/j}.$$

$$\widehat{\lambda}_j = \frac{1}{\#E_{j,L}}\sum_{i\in E_{j,L}} Y_i Y_{i-1}, \qquad \widehat{\eta}_j = \frac{1}{\#E_{j,L}}\sum_{i\in E_{j,L}} Y_i Y_{i-2}.$$

The estimators $\widehat{\sigma^2}$, $\widehat{\lambda}_j$, $\widehat{\eta}_j$ and $\widehat{\alpha}_{(j)}$ satisfy the following theorems:

- L2 convergence
- Convergence in probability (with the convergence rate $\sqrt{L}$)
- Almost sure convergence following from a simple application of Borel-Cantelli lemma, and convergence in distribution.

Gaussian Quasi-Maximum Likelihood estimation of isARMA(1,1) process

- We would like to estimate $\theta_* = (\alpha_*, \sigma_*^2, \nu_*^2)$.
- We are going to use a QMLE-type estimator and for this we are going to rely on the AR(∞) representation of (Y_ℓ).
- We have already seen that this requires $|\alpha_*| < 1$ and $\nu_*^2 \leq \sigma_*^2 < 2\,\nu_*^2$, ensuring $\sup_{\ell\in\mathbb{Z}} |\alpha_\ell| < 1$ and $\sup_{\ell\in\mathbb{Z}} |\beta_\ell| < 1$.
- As a consequence we define the compact set of parameters Θ as

$$\Theta = \left\{ \theta = (\alpha, \sigma^2, \nu^2) \in \mathbb{R}^3 \Big/ \underline{\nu}^2 \leq \nu^2 \leq \sigma^2 \leq \delta\,\nu^2 \leq \overline{\sigma}^2,\ |\alpha| \leq \overline{\alpha} \right\},$$

where $0 < \underline{\nu}^2 < \overline{\sigma}^2$, $1 \leq \delta < 2$ and $0 \leq \overline{\alpha} < 1$ are given. Then, for $\theta \in \Theta$ define the objective function $\widehat{\mathscr{L}_L}(\theta)$ by:

$$\widehat{\mathscr{L}_L}(\theta) = \frac{1}{\nu^2} \sum_{\ell=1}^{L} \left(Y_\ell + \sum_{i=1}^{\ell-1} c_{i,\ell}(\theta)\, Y_{\ell-i} \right)^2 + L \log \nu^2,$$

with the convention $\sum_{i=1}^{0} = 0$.

Remark

The function $\mathscr{L}_L(\theta)$ is explicit in case one knows all the inter-arrival times $k_1, \ldots, k_L$. This means that the estimate of θ^* is given conditionally with respect to those $k_1, \ldots, k_L$. Hence observations both include $Y_1, \ldots, Y_L$ and $k_1, \ldots, k_L$.

We define the Gaussian Quasi-Maximum Likelihood Estimator (QMLE) of θ_*:

$$\widehat{\theta}_L = \underset{\theta \in \Theta}{\operatorname{argmin}} \widehat{\mathscr{L}_L}(\theta). \tag{14}$$

The estimators $\theta_* = (\alpha_*, \sigma_*^2, \nu_*^2)$ satisfy the following theorems:

- Almost sure convergence
- Convergence in distribution.

Time domain methods for X-ray and gamma-ray astronomy

Four basic challenges of temporal analysis in high energy astronomy can be identified

1. Variations occur on an enormous range of timescales from milliseconds to decades.
2. Brightness changes exhibit a wide range of behaviors that can be deterministic or stochastic, aperiodic or periodic.
3. The data sets may have some unusual features, such as irregular observations with gaps and multiple dimensions. Gaps in the data stream can occur every $\sim$90 minutes on low-Earth orbiting satellites, and are often due to suboptimal scheduling of telescope assignment.
4. High energy light curves are often treated as a multivariate process. The most common multivariate studies concern how brightness variations relate to the source spectra (hardness ratios, plasma temperature, among others).

But stochastic effects are also present with both random and structured variations

Monte Carlo Experiences: Case of stationary inter-arrival times of sampling

We report the results of Monte Carlo experiments after 1000 independent replications. We will consider the following parameters:

- Sample length $L = 200$, $L = 500$, $L = 1000$ and $L = 2000$ for exhibiting the asymptotic behaviour.
- $\alpha_* = 0.2$ and $\alpha_* = 0.8$;
- $\nu_*^2 = 2$ and $\sigma_*^2 = 3$;
- $(k_\ell)_{\ell \in \mathbb{N}^*}$ is a sequence of i.i.d. r.v.'s following the two probability distributions:
  1. A Geometric distribution with parameter 0.4, denoted $\mathscr{G}$;
  2. A mixture of probability distribution denoted $\mathscr{M}$ such as $k_\ell = B_\ell + (1 - B_\ell)(1 + G_\ell)$ where (B_ℓ) is a sequence of i.i.d. Bernoulli random variables with parameter 0.9 and (G_ℓ) is a sequence of i.i.d. Geometric random variables with parameter 0.1.

- $(\zeta_\ell)_{\ell\in\mathbb{Z}}$ is a sequence of i.i.d. r.v.'s, following the three probability distributions:
  1. A standard Gaussian distribution $\mathscr{N}(0,1)$, denoted $\mathscr{N}$;
  2. A normalized symmetric uniform distribution $\mathscr{U}\left[-\sqrt{3},\sqrt{3}\right]$, denoted $\mathscr{U}$;
  3. A normalized Student distribution with 3 degrees of freedom, that is a $\sqrt{1/3}\,t(3)$ (with a unit variance), and denoted $t(3)$. In such a case, the condition $\mathbb{E}[\zeta_0^4] < \infty$ is not satisfied and this case is considered for investigating the possible extensions of the obtained limit theorems.

Results of Monte-Carlo experiments

- The Root Mean Squared Error defined has $RMSE_\theta = \sqrt{\frac{\sum_{j=1}^{L}(\theta_0 - \hat{\theta}_j)^2}{L}}$,
- θ_0 is the true value of the generic θ parameter and $\hat{\theta}_j$ is the estimate generated by the j-th iteration in the MC experiment.
- $\widehat{\alpha}_E$ and $\widehat{\sigma}^2_E$, denotes the empirical estimator for α_* and σ^2_* in an isARMA(1,1) model
- $\widehat{\alpha}_Q$ and $\widehat{\sigma}^2_Q$, denotes the Gaussian QMLE estimator for α and σ^2 respectively.

Table 1: RMSE for $\alpha_* = 0.8$

Sample length L		200				500				1000				2000			
Estimators		$\widehat{\alpha}_Q$	$\widehat{\alpha}_E$	$\widehat{\sigma}^2_Q$	$\widehat{\sigma}^2_E$	$\widehat{\alpha}_Q$	$\widehat{\alpha}_E$	$\widehat{\sigma}^2_Q$	$\widehat{\sigma}^2_E$	$\widehat{\alpha}_Q$	$\widehat{\alpha}_E$	$\widehat{\sigma}^2_Q$	$\widehat{\sigma}^2_E$	$\widehat{\alpha}_Q$	$\widehat{\alpha}_E$	$\widehat{\sigma}^2_Q$	$\widehat{\sigma}^2_E$
$\zeta_0 \sim \mathcal{N}$	$k_0 \sim \mathcal{G}$	0.070	1.116	0.412	0.411	0.038	0.273	0.264	0.266	0.027	0.231	0.184	0.184	0.019	0.202	0.134	0.136
	$k_0 \sim \mathcal{M}$	0.081	0.135	0.505	0.485	0.049	0.088	0.318	0.296	0.037	0.072	0.256	0.210	0.029	0.063	0.213	0.146
$\zeta_0 \sim \mathcal{U}$	$k_0 \sim \mathcal{G}$	0.081	0.630	0.370	0.367	0.039	0.270	0.224	0.227	0.028	0.227	0.154	0.156	0.020	0.203	0.111	0.113
	$k_0 \sim \mathcal{M}$	0.082	0.128	0.434	0.393	0.050	0.091	0.306	0.255	0.035	0.073	0.234	0.179	0.029	0.063	0.208	0.130
$\zeta_0 \sim t(3)$	$k_0 \sim \mathcal{G}$	0.099	2.844	4.091	3.951	0.045	0.311	5.920	6.193	0.027	0.263	2.185	2.189	0.019	0.242	1.112	1.100
	$k_0 \sim \mathcal{M}$	0.086	0.137	17.214	16.542	0.048	0.093	1.612	1.656	0.034	0.074	1.383	1.384	0.028	0.066	2.437	2.351

Table 2: RMSE for $\alpha_* = 0.2$

Sample length L		200				500				1000				2000			
Estimators		$\widehat{\alpha}_Q$	$\widehat{\alpha}_E$	$\widehat{\sigma}^2_Q$	$\widehat{\sigma}^2_E$	$\widehat{\alpha}_Q$	$\widehat{\alpha}_E$	$\widehat{\sigma}^2_Q$	$\widehat{\sigma}^2_E$	$\widehat{\alpha}_Q$	$\widehat{\alpha}_E$	$\widehat{\sigma}^2_Q$	$\widehat{\sigma}^2_E$	$\widehat{\alpha}_Q$	$\widehat{\alpha}_E$	$\widehat{\sigma}^2_Q$	$\widehat{\sigma}^2_E$
$\zeta_0 \sim \mathcal{N}$	$k_0 \sim \mathcal{G}$	0.128	0.706	0.362	0.360	0.087	0.310	0.228	0.235	0.063	0.202	0.162	0.169	0.043	0.165	0.116	0.119
	$k_0 \sim \mathcal{M}$	0.110	0.162	0.388	0.384	0.070	0.106	0.239	0.237	0.050	0.074	0.164	0.167	0.035	0.053	0.116	0.117
$\zeta_0 \sim \mathcal{U}$	$k_0 \sim \mathcal{G}$	0.133	1.597	0.272	0.279	0.083	0.271	0.165	0.174	0.062	0.207	0.120	0.125	0.044	0.168	0.083	0.089
	$k_0 \sim \mathcal{M}$	0.109	0.177	0.305	0.295	0.070	0.101	0.186	0.186	0.050	0.072	0.130	0.131	0.035	0.055	0.092	0.093
$\zeta_0 \sim t(3)$	$k_0 \sim \mathcal{G}$	0.153	20.289	4.184	4.035	0.087	0.607	5.936	5.892	0.062	0.220	2.037	2.171	0.044	0.194	1.109	1.109
	$k_0 \sim \mathcal{M}$	0.141	0.170	3.194	3.763	0.069	0.105	4.422	4.384	0.050	0.075	1.941	1.934	0.035	0.055	0.979	0.976

Conclusions of Monte-Carlo experiments

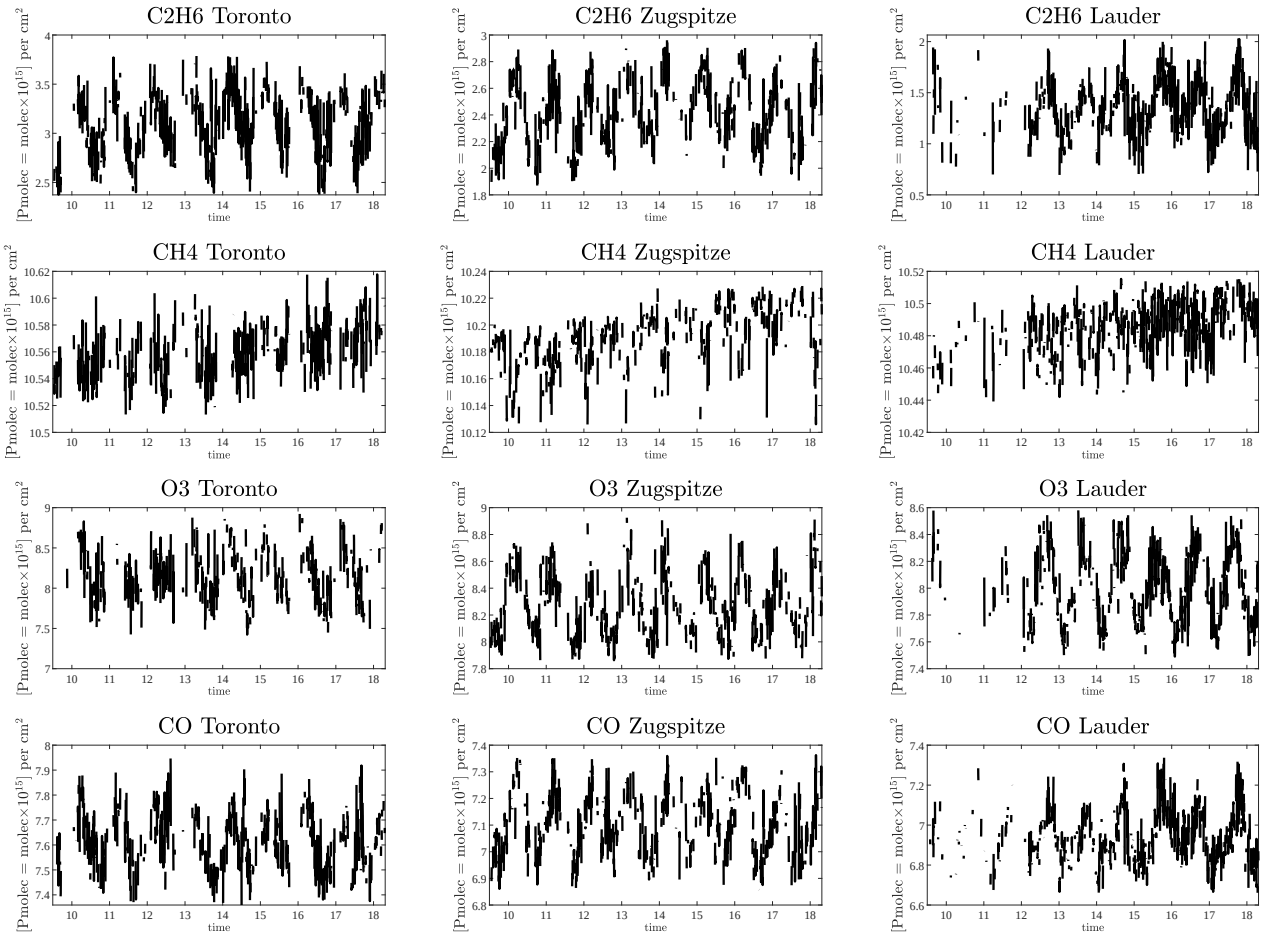
The numerical applications exhibit the following points:

- In case of stationary inter-arrival sampling times (k_ℓ), and with $\mathbb{E}[\xi_0^4] < 4$, the consistency of the QMLE and empirical estimators is checked.
- The QMLE estimators almost always provide the best consistency rates and the asymptotic normality can be exhibited. However, the estimator $\widehat{\sigma}_E^2$ seems to be almost accurate than $\widehat{\sigma}_Q^2$. Concerning $\widehat{\alpha}_E$, its convergence rate seems to depend on the distribution of (k_ℓ) and the less missing data the better its convergence rate.
- In case of non stationary inter-arrival sampling times (k_ℓ) (that has not been presented here), the QMLE estimators seem to remain as efficient as in the stationary case.

Empirical illustration

- We study an empirical illustration of the isARMA process concerning the issue of fitting the ground-based FTIR solar spectra.
- We consider four main direct and indirect greenhouse gases, namely ethane (C_2H_6), methane (CH_4), ozone (O_3) and carbon monoxide (CO),
- Data are recorded at different ground-stations located at Toronto (Canada), Zugspitze (Germany) and Lauder (New Zealand).
- The datasets are available at NOAA.

FTIR measurements of the ethane, methane, ozone and carbon monoxide abundance in the atmosphere



Analysis and results of FTIR data

- The Kalman filter (KF) procedure is here adopted as a standard benchmark to handle the missing data problem.
- We study both the isARMA(1,1) and KF methodologies in fitting the x_t residuals obtained after apply a deterministic component capturing both the annual periodicity and a quadratic trend to the 12 time series.
- We conclude that the residuals do not display significant serial correlations, suggesting that the isARMA specification leads to a good fit of the current datasets.
- The two models lead to almost identical statistics highlighting the similarity between the two approaches, even though slightly lower RMSEs are provided by the KF technique. This could suggest that the two approaches are in some way interchangeable.
- However, the isARMA methodology presents an huge advantage in the computational time, since only L operations are required to compute the log likelihood, while the KF alternative necessitates n iterations.

Other models considering incomplete data

- Spectral (Whittle) estimation in the presence of missing data (with Doukhan, 2009)
- New discrete-time nonlinear state-space formulation to parameterize GARCH models, to use these models to develop estimation techniques that are also valid in situations where observations are missing (with Ossandón, 2013).
- Linear regression model with long-memory noise and an independent set of random times given by renewal process sampling and a random sampling time called "jittered sampling" (with Araya el al., 2023).

References

1. H. Araya, N. Bahamonde, L. Fermín, T. Roa & S. Torres (2023) On the consistency of least squared estimator in models sampled at random times driven by long memory noise; the jittered case. Statistica Sinica 33.
2. N. Bahamonde, J.M.. Bardet, K. Bertin, P. Doukhan & F. Maddanu (2024) An irregularly spaced ARMA(1,1) model and an application for contamination data (Preprint).
3. P. Bondon & N. Bahamonde (2012), Least squares estimation of ARCH models with missing observations, Journal of Time Series Analysis 33
4. W. Dunsmuir & P. M. Robinson (1981) Asymptotic theory for time series containing missing and amplitude modulated observations. Sankhya Ser. A, 43.
5. F. Elorrieta, S. Eyheramendy & W. Palma (2019) Discrete-time autoregressive model for unequally spaced time-series observations. A& A, 627, A120.
6. R. J. A. Little & D. B. Rubin (2002) Statistical analysis with missing data. Wiley Series in Probability and Statistics, NJ, second edition.
7. E. Parzen (1963), On spectral analysis with missing observations and amplitude modulation. Sankhya Ser. A, 25.
8. Y. Yajima & H. Nishino (1999), Estimation of the autocorrelation function of a stationary time series with missing observations. Sankhya Ser. A, 61.

Gracias