Assignment 2 (ML for TS) - MVA 2023/2024

Matteo Marengo matteo.marengo@ens-paris-saclay.fr
Hugo Robert hugo.robert@ens-paris-saclay.fr

December 4, 2023

Part I

Introduction

Objective. The goal is to better understand the properties of AR and MA processes, and do signal denoising with sparse coding.

Warning and advice.

- Use code from the tutorials as well as from other sources. Do not code yourself well-known procedures (e.g. cross validation or k-means), use an existing implementation.
- The associated notebook contains some hints and several helper functions.
- Be concise. Answers are not expected to be longer than a few sentences (omitting calculations).

Instructions.

- Fill in your names and emails at the top of the document.
- Hand in your report (one per pair of students) by Tuesday 5$^{\text{th}}$ December 11:59 PM.
- Rename your report and notebook as follows:
  `FirstnameLastname1_FirstnameLastname1.pdf` and
  `FirstnameLastname2_FirstnameLastname2.ipynb`.
  For instance, `LaurentOudre_CharlesTruong.pdf`.
- Upload your report (PDF file) and notebook (IPYNB file) using this link: docs.google.com/forms/d/e/1FAIpQLSfCqMXSDU9jZJbYUMmeLCXbVeckZYNiDpPl4hRUwcJ2cBHQMw/viewform?usp=sf_link.

Part II

General questions

A time series $\{y_t\}_t$ is a single realisation of a random process $\{Y_t\}_t$ defined on the probability space $(\Omega, \mathcal{F}, P)$, i.e. $y_t = Y_t(w)$ for a given $w \in \Omega$.
In classical statistics, several independent realisations are often needed to obtain a “good” estimate (meaning consistent) of the parameters of the process.
However, thanks to a stationarity hypothesis and a "short-memory" hypothesis, it is still possible to make “good” estimates. The following question illustrates this fact.

Question 1

An estimator $\hat{\theta}_n$ is consistent if it converges in probability when the number n of samples grows to ∞ to the true value $\theta \in \mathbb{R}$ of a parameter, i.e. $\hat{\theta}_n \xrightarrow{\mathcal{D}} \theta$.

- Recall the rate of convergence of the sample mean for i.i.d. random variables with finite variance.
- Let $\{Y_t\}_{t\geq 1}$ a wide-sense stationary process such that $\sum_k |\gamma(k)| < +\infty$.
  Show that the sample mean $\bar{Y}_n = (Y_1 + \cdots + Y_n)/n$ is consistent and enjoys the same rate of convergence as the i.i.d. case.
  (Hint: bound $\mathbb{E}[(\bar{Y}_n - \mu)^2]$ with the $\gamma(k)$ and recall that convergence in L_2 implies convergence in probability.)

Answer 1

1.1: Rate of Convergence for IID Variables

For i.i.d. random variables $X_1, X_2, \ldots, X_n$ with finite variance σ^2 and mean μ, the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an estimator for μ.
Its rate of convergence can be found thanks to the **Central Limit Theorem**. Indeed, this theorem states that as n approaches infinity, the distribution of $\sqrt{n}(\bar{X}_n - \mu)$ converges in distribution to a normal distribution with mean 0 and variance σ^2.

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \tag{1}$$

$$\frac{(\bar{X}_n - \mu)}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1) \tag{2}$$

As a result, the rate of convergence of $\bar{X}_n$ to μ is $O(\frac{1}{\sqrt{n}})$. Indeed, as we want to measure the distance between the distribution $\bar{X}_n$ and μ, this distance will decrease as $1/\sqrt{n}$

1.2: Wide-sense stationary process

The Law of Large Numbers is a law that states that as the number of samples n increases, the sample mean $\bar{X}_n$ of i.i.d. random variables converges in probability to the expected value μ. It is written as $\bar{X}_n \xrightarrow{P} \mu$ when $n \to \infty$.

For a wide-sense stationary process Y_t, the mean $\mu = \mathbb{E}[Y_t]$ is constant and the autocovariance function $\gamma(k) = \mathbb{E}[(Y_t - \mu)(Y_{t+k} - \mu)]$ depends only on the lag k and not on time t.

$$\begin{aligned}
\mathbb{E}[(\bar{Y}_n - \mu)^2] &= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu)\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\sum_{i=1}^{n}(Y_i - \mu)^2 + 2\sum_{i<j}(Y_i - \mu)(Y_j - \mu)\right] \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n}\mathbb{E}[(Y_i - \mu)^2] + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathbb{E}[(Y_i - \mu)(Y_j - \mu)]\right] \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{n}\sigma^2 + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\gamma(j-i)\right] \\
&= \frac{1}{n^2}\left[n\sigma^2 + 2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\gamma(j-i)\right] \\
&= \frac{1}{n^2}\left[n\sigma^2 + 2\sum_{\tau=1}^{n-1}(n-\tau)\gamma(\tau)\right] \\
&= \frac{\sigma^2}{n} + \frac{2}{n^2}\sum_{\tau=1}^{n-1}(n-\tau)\gamma(\tau)
\end{aligned}$$

Since $\sum_k |\gamma(k)| < +\infty$, the second term in this expression diminishes faster than $\frac{1}{n}$ as $n \to \infty$.

$$\lim_{n\to\infty} \mathbb{E}[(\bar{Y}_n - \mu)^2] = 0$$

This implies convergence in L2. and hence, by Chebyshev's inequality, convergence in probability:

$$\mathbb{P}\left(|\bar{Y}_n - \mu| > \epsilon\right) \leq \frac{\mathbb{E}[(\bar{Y}_n - \mu)^2]}{\epsilon^2} \to 0 \text{ as } n \to \infty.$$

We can conlude that $\bar{Y}$ is consistent. Lastly, the rate of convergence of $\bar{Y}_n$ to μ is $\frac{1}{n}$, so it is at least as fast as $O(\frac{1}{\sqrt{n}})$, the same as in the i.i.d. case.

Part III

AR and MA processes

Question 2 *Infinite order moving average MA(∞)*

Let $\{Y_t\}_{t\geq 0}$ be a random process defined by

$$Y_t = \varepsilon_t + \psi_1\varepsilon_{t-1} + \psi_2\varepsilon_{t-2} + \cdots = \sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k} \tag{3}$$

where $(\psi_k)_{k\geq 0} \subset \mathbb{R}$ ($\psi = 1$) are square summable, i.e. $\sum_k \psi_k^2 < \infty$ and $\{\varepsilon_t\}_t$ is a zero mean white noise of variance σ_ε^2. (Here, the infinite sum of random variables is the limit in L_2 of the partial sums.)

- Derive $\mathbb{E}(Y_t)$ and $\mathbb{E}(Y_t Y_{t-k})$. Is this process weakly stationary?
- Show that the power spectrum of $\{Y_t\}_t$ is $S(f) = \sigma_\varepsilon^2 |\phi(e^{-2\pi i f})|^2$ where $\phi(z) = \sum_j \psi_j z^j$. (Assume a sampling frequency of 1 Hz.)

The process $\{Y_t\}_t$ is a moving average of infinite order. Wold's theorem states that any weakly stationary process can be written as the sum of the deterministic process and a stochastic process which has the form (3).

Answer 2

2.1: Derivation of $\mathbb{E}(Y_t)$ and $\mathbb{E}(Y_t Y_{t_k})$

In order to analyze whether the process $\{Y_t\}_{t\geq 0}$ is weakly stationary or not we have to verify two points; if its expected value $\mathbb{E}(Y_t)$ and its autocovariance function $\mathbb{E}(Y_t Y_{t-k})$ are time-invariant.

The expected value of Y_t is:

$$\mathbb{E}(Y_t) = \mathbb{E}\left(\sum_{k=0}^{\infty} \psi_k \varepsilon_{t-k}\right)$$

Since $\mathbb{E}(\varepsilon_t) = 0$ for all t as it is a zero mean white noise, and ψ_k are constants, we have:

$$\mathbb{E}(Y_t) = \sum_{k=0}^{\infty} \psi_k \mathbb{E}(\varepsilon_{t-k}) = \sum_{k=0}^{\infty} \psi_k \cdot 0 = 0$$

Now for the autocovariance function $\mathbb{E}(Y_t Y_{t-k})$:

$$\mathbb{E}(Y_t Y_{t-k}) = \mathbb{E}\left(\left(\sum_{i=0}^{\infty} \psi_i \varepsilon_{t-i}\right)\left(\sum_{j=0}^{\infty} \psi_j \varepsilon_{t-k-j}\right)\right)$$

We then obtain this result:

$$\mathbb{E}(Y_t Y_{t-k}) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \psi_i \psi_j \mathbb{E}(\varepsilon_{t-i}\varepsilon_{t-k-j})$$

As ε_t is a white noise, $\mathbb{E}(\varepsilon_{t-i}\varepsilon_{t-k-j}) = 0$ for $i \neq k+j$ (ε are i.i.d, so we obtain a product of $\mathbb{E}(\varepsilon_{t-i})\mathbb{E}(\varepsilon_{t-k-j}) = 0$) and σ_ε^2 for $i = k+j$. Therefore, the expression can be simplified to:

$$\mathbb{E}(Y_t Y_{t-k}) = \sigma_\varepsilon^2 \sum_{i=0}^{\infty} \psi_i \psi_{i+k}$$

As defined previously, for a process to be weakly stationary, its mean and autocovariance function must not depend on the time t but only depend on the lag k.

1. Constant Mean: As shown, $\mathbb{E}(Y_t) = 0$ for all t, satisfying the constant mean criterion.

2. Autocovariance Function: The autocovariance function $\mathbb{E}(Y_t Y_{t-k})$ depends only on the lag k and not on t, satisfying the second criterion.

Given these two properties, the process $\{Y_t\}_{t\geq 0}$ is weakly stationary.

2.2: Power spectrum of Y_t

Let's compute the power spectrum of Y_t.

The power spectrum of a stationary stochastic process is the Fourier transform of its autocovariance function. For a discrete-time process, the power spectrum $S(f)$ at frequency f is defined as:

$$S(f) = \sum_{k=-\infty}^{\infty} \gamma_Y(k) e^{-2\pi i f k}$$

where $\gamma_Y(k)$ is the autocovariance function of the process $\{Y_t\}_t$ at lag k.

To show that the power spectrum of the process $\{Y_t\}_t$ is $S(f) = \sigma_\varepsilon^2 |\phi(e^{-2\pi i f})|^2$, where $\phi(z) = \sum_j \psi_j z^j$, we need to carry out a detailed calculation step by step. We are given that the process $\{Y_t\}_t$ is a moving average of infinite order with coefficients ψ_k and that $\{\varepsilon_t\}_t$ is a white noise process with zero mean and variance σ_ε^2.

The autocovariance function at lag k is:

$$\gamma_Y(k) = \mathbb{E}(Y_t Y_{t-k})$$

$$\gamma_Y(k) = \sigma_\varepsilon^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k}$$

Therefore the power spectrum can be written like that:

$$S(f) = \sum_{k=-\infty}^{\infty} \gamma_Y(k) e^{-2\pi i f k}$$

$$S(f) = \sum_{k=-\infty}^{\infty} \left(\sigma_\varepsilon^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k} \right) e^{-2\pi i f k}$$

$$S(f) = \sigma_\varepsilon^2 \sum_{k=-\infty}^{\infty} \left(\sum_{j=0}^{\infty} \psi_j \psi_{j+k} \right) e^{-2\pi i f k}$$

The generating function $\phi(z)$ is defined as:

$$\phi(z) = \sum_{j=0}^{\infty} \psi_j z^j$$

To compute the squared magnitude of the Fourier series representation of a function, we consider the expression $\left| \sum_{j=0}^{N} \psi_j e^{-2\pi i f j} \right|^2$.

$$\begin{aligned} \left| \sum_{j=0}^{N} \psi_j e^{-2\pi i f j} \right|^2 &= \left(\sum_{j=0}^{N} \psi_j e^{-2\pi i f j} \right) \left(\sum_{i=0}^{N} \psi_i e^{2\pi i f i} \right) \\ &= \sum_{j=0}^{N} \sum_{i=0}^{N} \psi_j \psi_i e^{-2\pi i f (j-i)} \\ &= \sum_{\tau=-N+1}^{N-1} \sum_{n=0}^{N-\tau-1} \psi_n \psi_{n+\tau} e^{-2\pi i f \tau} \end{aligned}$$

In the above computation, we use the index change $i = n + \tau$ to transform the double sum into a single sum over τ, where τ represents the difference in indices between the terms in the original series.

As we extend the limits of the sum to infinity, we have:

$$\phi\left(e^{-2\pi i f}\right)^2 = \lim_{N\to\infty} \sum_{\tau=-N+1}^{N-1} \sum_{n=0}^{N-\tau-1} \psi_n \psi_{n+\tau} e^{-2\pi i f \tau}$$

Therefore, we can conclude that:

$$S(f) = \sigma_\varepsilon^2 \left| \phi(e^{-2\pi i f}) \right|^2$$

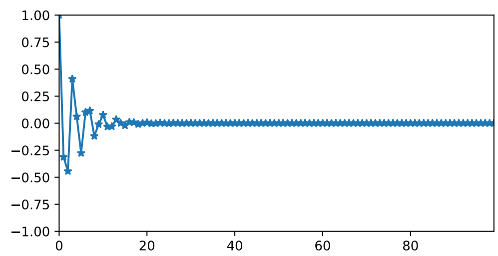
Question 3 *AR(2) process*

Let $\{Y_t\}_{t\geq 1}$ be an AR(2) process, i.e.

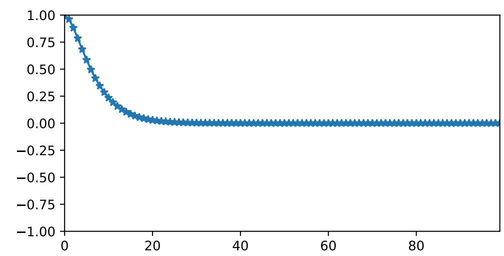
$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t \tag{4}$$

with $\phi_1, \phi_2 \in \mathbb{R}$. The associated characteristic polynomial is $\phi(z) := 1 - \phi_1 z - \phi_2 z^2$. Assume that ϕ has two distinct roots (possibly complex) r_1 and r_2 such that $|r_i| > 1$. Properties on the roots of this polynomial drive the behaviour of this process.

- Express the autocovariance coefficients $\gamma(\tau)$ using the roots r_1 and r_2.
- Figure 1 shows the correlograms of two different AR(2) processes. Can you tell which one has complex roots and which one has real roots?
- Express the power spectrum $S(f)$ (assume the sampling frequency is 1 Hz) using $\phi(\cdot)$.
- Choose ϕ_1 and ϕ_2 such that the characteristic polynomial has two complex conjugate roots of norm $r = 1.05$ and phase $\theta = 2\pi/6$. Simulate the process $\{Y_t\}_t$ (with $n = 2000$) and display the signal and the periodogram (use a smooth estimator) on Figure 2. What do you observe?



Correlogram of the first AR(2)



Correlogram of the second AR(2)

Figure 1: Two AR(2) processes

Answer 3

3.1: Autocovariance coefficients $\gamma(\tau)$

Let $\{Y_t\}_{t\geq 1}$ be an AR(2) process given by:

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t$$

where $\phi_1, \phi_2 \in \mathbb{R}$, and ε_t is a white noise process.

The autocovariance function $\gamma(\tau)$ is defined as:

$$\gamma(\tau) = E(Y_t Y_{t-\tau})$$

which, for an AR(2) process, can be expressed as:

$$\begin{aligned}
\gamma(\tau) =& E\left[(\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t) Y_{t-\tau}\right] \\
=& \phi_1 E(Y_{t-1} Y_{t-\tau}) + \phi_2 E(Y_{t-2} Y_{t-\tau}) + E(\varepsilon_t Y_{t-\tau}) \\
=& \begin{cases} \phi_1 \gamma(\tau-1) + \phi_2 \gamma(\tau-2), & \text{for } \tau \neq 0, \\ \phi_1 \gamma(\tau-1) + \phi_2 \gamma(\tau-2) + \sigma_\varepsilon^2, & \text{for } \tau = 0. \end{cases}
\end{aligned}$$

The initial conditions for $\gamma(0)$, $\gamma(1)$, and $\gamma(2)$ are obtained by solving:

$$\begin{aligned}
\gamma(0) &= \phi_1\gamma(1) + \phi_2\gamma(2) + \sigma_\varepsilon^2, \\
\gamma(1) &= \phi_1\gamma(0) + \phi_2\gamma(1), \\
\gamma(2) &= \phi_1\gamma(1) + \phi_2\gamma(0).
\end{aligned}$$

Solving these equations yields:

$$\begin{aligned}
\gamma(0) &= \frac{(1-\phi_2)\frac{\sigma_\varepsilon^2}{1+\phi_2}}{(1-\phi_2)^2 - \phi_1^2}, \\
\gamma(1) &= \frac{\phi_1\frac{\sigma_\varepsilon^2}{1+\phi_2}}{(1-\phi_2)^2 - \phi_1^2}.
\end{aligned}$$

Given $\gamma(0)$ and $\gamma(1)$, we can compute $\gamma(\tau)$ for $\tau = 2, 3, \ldots$ as follows:

$$\gamma(\tau) = \phi_1\gamma(\tau-1) + \phi_2\gamma(\tau-2)$$

Furthermore, as r_1 and r_2 are the roots of $\phi(z)$, we can express them as:

$$\begin{aligned}
r_1 + r_2 &= -\frac{\phi_1}{\phi_2} \\
r_1 r_2 &= -\frac{1}{\phi_2}
\end{aligned}$$

Therefore:

$$\begin{aligned}
\phi_2 &= -\frac{1}{r_1 r_2} \\
\phi_1 &= \frac{1}{r_1} + \frac{1}{r_2}
\end{aligned}$$

Finally:

$$\gamma(\tau) = -(\frac{1}{r_1 r_2})\gamma(\tau-1) + (\frac{1}{r_1} + \frac{1}{r_2})\gamma(\tau-2)$$

3.2: Complex vs Real roots

Real Roots: The correlogram of an AR(2) process with real roots will an exponential decay towards zero without any regular oscillations. Indeed, there are no imaginary parts to introduce oscillations

Complex Roots: The correlogram with complex roots will have a damped sinusoidal pattern. In this pattern,the autocorrelations will oscillate and gradually be lowered in amplitude as the lag increases.

The correlogram on the **left** shows a damped oscillation, suggesting that this process has **complex roots**. The correlogram on the **right** issmoothly decaying without any oscillation which might show that it is a process with **real roots**.

3.3: Power spectrum of $S(f)$

The power spectrum $S(f)$ can be expressed using the noise variance σ_ε^2 and the characteristic polynomial evaluated at $e^{-i2\pi f}$. Given the sampling frequency is 1 Hz, the power spectrum $S(f)$ can be written as:

$$S(f) = \frac{\sigma_\varepsilon^2}{|\phi(e^{-i2\pi f})|^2}$$

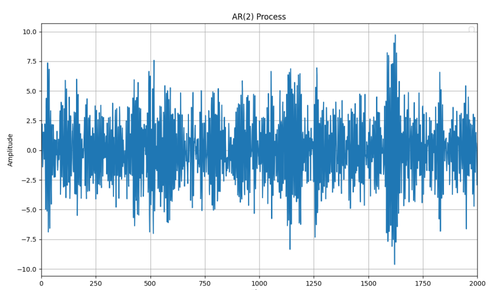
Furthermore, $\phi(e^{-i2\pi f})$ represents the characteristic polynomial evaluated at $e^{-i2\pi f}$, which includes the AR coefficients ϕ_1 and ϕ_2 of the given process:

$$\phi(e^{-i2\pi f}) = 1 - \phi_1 e^{-i2\pi f} - \phi_2 e^{-i4\pi f}$$

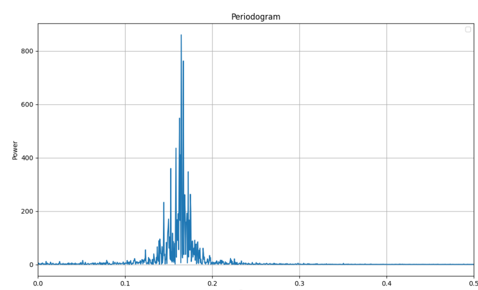
Therefore, the final expression for the power spectrum is:

$$S(f) = \frac{\sigma_\varepsilon^2}{|1 - \phi_1 e^{-i2\pi f} - \phi_2 e^{-i4\pi f}|^2}$$

3.4: Complex conjugate roots



Signal



Periodogram

Figure 2: AR(2) process

As shown in Figure 2, the signal corresponds to an AR(2) process. Concerning the periodogram, the peak is obtained at a frequency of 0.17. This is because, as the sampling frequency is $fs = 1Hz$, the frequency oscillation of the process is: $f = \frac{\theta fs}{2\pi} \approx 0.17$

Part IV

Sparse coding

The modulated discrete cosine transform (MDCT) is a signal transformation often used in sound processing applications (for instance to encode a MP3 file). A MDCT atom $\phi_{L,k}$ is defined for a length 2L and a frequency localisation k ($k = 0, \ldots, L-1$) by

$$\forall u = 0, \ldots, 2L-1, \quad \phi_{L,k}[u] = w_L[u]\sqrt{\frac{2}{L}}\cos[\frac{\pi}{L}\left(u + \frac{L+1}{2}\right)(k + \frac{1}{2})] \tag{5}$$

where w_L is a modulating window given by

$$w_L[u] = \sin\left[\frac{\pi}{2L}\left(u + \frac{1}{2}\right)\right]. \tag{6}$$

Question 4 *Sparse coding with OMP*

For the signal provided in the notebook, learn a sparse representation with MDCT atoms. The dictionary is defined as the concatenation of all shifted MDCDT atoms for scales L in $[32, 64, 128, 256, 512, 1024]$.

- For the sparse coding, implement the Orthogonal Matching Pursuit (OMP). (Use convolutions to compute the correlations coefficients.)
- Display the norm of the successive residuals and the reconstructed signal with 10 atoms.

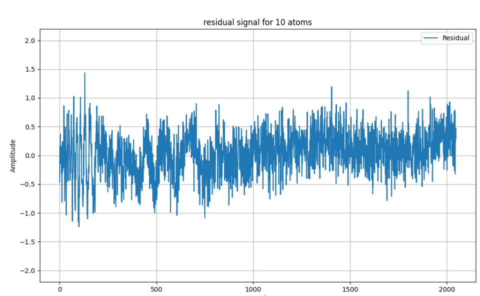
Answer 4

4.1: OMP implementation

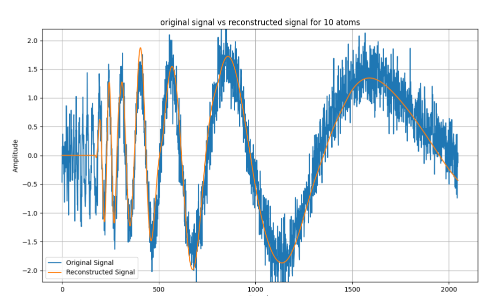
The OMP is implemented in the notebook. We compute the correlation between the residual and elements of the dictionary and add the candidate with the best correlation step by step until we get num_atoms atoms.

4.2: Norm of the successive residuals - Reconstructed signal

The results of the implemented code are in Fig 3.



Norms of the successive residuals



Reconstruction with 10 atoms

Figure 3: Results of Question 4

From this result, we can observe that the OMP method works well because the **reconstructed signal fits the original signal and the residual seems to be the data noise**. However, the reconstructed signal is zero up to sample 200, a value which decreases as num_atoms is increased. This is because the increase in num_atoms allows other higher-frequency atoms to reconstruct the start of the signal.