Imperial College London

Department of Mathematics

Modelling non-stationary time series with application to pairs trading

Tristan de Certaines

CID: 02108498

Supervised by Bennet Ströh and Axel Gandy

Summer 2022

Submitted in partial fulfilment of the requirements for the MSc in Statistics of Imperial College London

The work contained in this thesis is my own work unless otherwise stated.

Signed: Tristan de Certaines Date: September 2, 2022

Abstract

In this work, we introduce a general theory based on stationary approximations in the L^q sense for discrete-time locally stationary processes. Using these approximations, we derive and study time-varying estimates for an autoregressive model, focussing on the Yule-Walker and maximum-likelihood methods. A localized law of large numbers is then established to prove the consistency of such estimates. Numerical experiments are also conducted to verify their asymptotic behaviours. Finally, we use a time-varying autoregressive process to develop a trading strategy benefitting from pricing anomalies between Bitcoin and Ethereum. This strategy includes a forecasting methodology as well as evaluation metrics.

Acknowledgements

First and foremost, I would like to thank my supervisor Bennet Ströh for his constant assistance and dedicated involvement throughout this thesis. The generous time taken every week helped me greatly to stay in the right direction and was always an inspiring and pleasant time. I am very grateful for his helpful advice as well as his time spent rigorously reviewing my work while giving me complete freedom.

Additionally, I am grateful for my family, whose constant love and support keeps me motivated, as well as for my friends Simon and Alexandre without whom this work would have been far less joyful.

Contents

Notations

- For $q > 0$ and Z a random vector in $\mathbb{R}^k$, denote: $||Z||_q = [\mathbb{E}(Z^q)]^{1/q}$.
- For $q > 0$ and $y \in \mathbb{R}^k$, $|y|_q = \left(\sum_{i=1}^k |y_i|^q\right)^{1/q}$.
- If y is a vector or a matrix, then $y^\top$ denotes its transpose.
- Given a set A, $\mathbb{1}_A(x)$ equals 1 if $x \in A$ and 0 otherwise.
- $\delta_{i,j}$ denotes the Kronecker function defined as: $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise.

1. Introduction

Time series analysis is a major field of statistics which helps in understanding the underlying structure in observed data. Throughout several decades, a large theory has been developed, providing many models and methods to study a time series (Brockwell and Davis, 2009). Consequently, many of these models have been successfully applied in diverse areas over time (e.g. finance, physics, environmental data), whether the goal was to obtain insight into the data or forecast future values (Chan, 2004; Hipel and McLeod, 1994). In finance, Pairs trading, where two similar assets (e.g. stocks, bonds) are simultaneously traded to benefit from mispricing, has proven to achieve great results in the past by modelling and forecasting the spread between the two assets (Vidyamurthy, 2004). However, most of the methods from the existing literature rely on one assumption: the time series of interest must be stationary. Stationarity is a property ensuring that the statistical structure of the process is time invariant and is rarely satisfied in practice. For that reason, many time series are first transformed to meet this criterion before then being analysed. Although this method is effective, it comes at the price of losing some interpretability since analysis is performed on transformed data. In Section 2, we start by introducing important concepts and methods of time series analysis.

Recently, a new theory mainly developed by Rainer Dahlhaus emerged, aiming to go beyond the investigation of time series under the assumption of stationarity and allowing for more general models (Dahlhaus, 2012; Dahlhaus et al., 2019). The key idea of this theory is to assume that a process can be locally approximated in the L^q sense by a stationary process. This allows us to derive new methods and new time-varying models that can be fitted directly on the original data. In Section 3, we investigate the simplest example of a locally stationary process, a time-varying autoregressive process, as a way to give the intuition behind this theory and show how a model is fitted to the data. We study different estimation methods where we notice similarities with the stationary case and discuss the properties of the estimators. We also discuss the choice of window size where the process is approximately stationary to achieve the best estimation. In order to make asymptotic considerations on the estimators, we need to look at the data in a rescaled time domain (Dahlhaus, 1996). This is because a sample from a non-stationary process holds no information about the probabilistic structure of the process in the future, so asymptotics need to be different. In this framework, we derive a localized law of large numbers which is then used to study the convergence and consistency of the estimators. Finally, we run a Monte-Carlo simulation to illustrate in an example the asymptotic behaviour of the estimators, in accordance with the theory.

Over the last five years, global interest in the cryptocurrencies market and intensive

trading has led to highly speculative price movements, making traditional time series analysis hardly applicable. The theory based on the locally stationary approximation being more general, we are interested to see if it can successfully be applied in a pairs trading application on the cryptocurrencies market. In Section 4, we make use of locally stationary processes to model the spread between two cryptocurrencies (Bitcoin and Ethereum), with the aim to develop simple trading strategies. These strategies are then backtested on historical data, during a high-fluctuating period, to evaluate their performances.

2. Background

In this section, we aim to give the reader some understanding of the general results used in time series analysis. Indeed, these concepts represent the starting point of the theory of locally stationary processes which will be used to model non-stationary time series. In several cases, results about locally stationary processes will be adaptations of known results on stationary time series.

2.1. Real-valued discrete-time stationary processes

2.1.1. Definitions

First of all, a time series is a series of data points indexed in time order, where each point is represented by the realisation of a random variable from a stochastic process. To clarify this definition, we define what is meant by a stochastic process.

Definition 2.1.1 (Stochastic Process)**.** A stochastic process is a collection of random variables $\{X_t,\, t \in T\}$. The index t represents time and T is the index set of the process. The most common index sets are $T = \mathbb{N}$ (discrete-time) and $T = \mathbb{R}_+$ (continuous time).

Since we only consider real-valued discrete-time stochastic processes, we will always have $T = \mathbb{N}$ and X_t defined on the probability space $(\Omega, \mathcal{F}, P)$.

Times series analysis can then be described as a branch of applied stochastic processes. Most of the theory that was developed in that branch relies on the assumption that the process is *stationary*.

2.1.2. Stationarity

Intuitively, stationarity means that the statistical properties of the process generating a time series do not change over time. This property plays a crucial role in time series analysis as it is used to analyse, fit and predict such series.

When stating that a process $\{X_t\}_{t\in T}$ is stationary, we usually refer to second-order stationarity, which is a weaker definition than strict stationarity.

Definition 2.1.2 (strict stationarity)**.** The process $\{X_t\}_{t\in T}$ is said to be strictly stationary if, for all $n \geq 0$, for any $t_1, \ldots, t_n \in T$, and for any s such that $t_1 + s, \ldots, t_n + s \in T$, the joint CDF of $\{X_{t_1}, \ldots, X_{t_n}\}$ is the same as that of $\{X_{t_1+s}, \ldots, X_{t_n+s}\}$. This means that the probabilistic structure of a strictly stationary process is invariant under a shift in time.

This definition is however very restrictive and does not apply to most time series that we can encounter. Hence, we usually rely on a weaker form of stationarity.

Definition 2.1.3 (weak / second-order stationarity)**.** The process $\{X_t\}_{t\in T}$ is said to be weakly stationary if, for all $t \in T$ and τ such that $t+\tau \in T$, $\mathbb{E}(X_t)$ is finite and constant, and $\mathrm{Cov}(X_t, X_{t+\tau})$ is finite and depends only on τ. Under this assumption, only the joint first and second moments are invariant under a shift in time.

Remark 2.1.4. If $\{X_t\}_{t\in T}$ is strictly stationary, then it follows that it is also weakly stationary. The converse is not always true however. Indeed, the process:

$$X_t \sim \begin{cases} \mathrm{Exp}(1) & \text{if } t \text{ is odd} \\ \mathcal{N}(1,1) & \text{if } t \text{ is even} \end{cases}$$

is stationary with mean 1 and $\mathrm{Cov}(X_t, X_{t+\tau})$ equal to 1 if $\tau = 0$ and 0 otherwise. Yet, X_1 and X_2 have different distributions so the process cannot be strictly stationary.

2.1.3. Examples of real-valued processes

Example 2.1.5 (White noise)**.** The process $\{\epsilon_t\}_{t\in T}$ is a white noise process if, for all $t \in T$, there exist $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_+$ such that $\mathbb{E}(\epsilon_t) = \mu$, $\mathrm{Var}(\epsilon_t) = \sigma^2$ and $\mathrm{Cov}(\epsilon_t, \epsilon_{t+\tau}) = \sigma^2\delta_{\tau,0}$, where $\delta_{\tau,0}$ denotes the Kronecker function that is equal to 1 if $\tau = 0$ and 0 otherwise.

Example 2.1.6 (q-th order moving average process (MA(q)))**.** The process $\{X_t\}_{t\in T}$ is a q-th order moving average process if it can be expressed as:

$$X_t = \mu - \sum_{j=1}^{q} \phi_j \epsilon_{t-j} + \epsilon_t, \quad t \in T. \tag{2.1}$$

where $\mu, \phi_1, \dots \phi_q$ are constants ($\phi_q \neq 0$) and $\{\epsilon_t\}_{t\in T}$ is a zero-mean white noise process.

Example 2.1.7 (p-th order autoregressive process (AR(p)))**.** The process $\{X_t\}_{t\in T}$ is a p-th order autoregressive process if it can be expressed as:

$$X_t = -\sum_{j=1}^{p} \alpha_j X_{t-j} + \epsilon_t, \quad t \in T. \tag{2.2}$$

where $\alpha_1, \dots \alpha_p$ are constants ($\alpha_p \neq 0$) and $\{\epsilon_t\}_{t\in T}$ is a zero-mean white noise process.

While an MA(q) process is always stationary, an AR(p) process is only stationary when the roots of the polynomial $P(X) = 1 + \alpha_1 X + \cdots + \alpha_p X^p$ are outside the unit circle (Brockwell and Davis (2009)). If this is the case, an AR(p) process can be seen as an 'MA(∞)' process.

2.1.4. Estimation

In practice, when considering a time series, we only have N observations from one realisation of the process. Hence, we cannot directly know the values of some statistics (mean, variance, ...) and need to estimate them. Being able to estimate some properties is essential to understand and fit a model. An important statistic in the context of stationary time series is the autocovariance sequence $\{c_\tau\}_{\tau\in\mathbb{Z}} = \{\mathrm{Cov}(X_t, X_{t+\tau})\}_{\tau\in\mathbb{Z}}$. Methods for estimating quantities are based on ergodicity (Krengel, 2011), which is the strategy of replacing ensemble averages by their corresponding time averages.

Definition 2.1.8. The usual autocovariance sequence estimator is given by:

$$\hat{c}_\tau = \frac{1}{N}\sum_{t=1}^{N-|\tau|}(X_t - \bar{X})(X_{t+\tau} - \bar{X}), \quad \tau \in \mathbb{Z} \tag{2.3}$$

with $\bar{X} = \frac{1}{N}\sum_{t=1}^{N} X_t$ the sample mean.

Remark 2.1.9. Although this estimator is biased, it is to be preferred to its unbiased version where we replace $\frac{1}{N}$ with $\frac{1}{N-|\tau|}$. Indeed, only this estimator (2.3) verifies the positive semidefinite property of the autocovariance sequence.

In the case where the mean of the process is known, we can also replace $\bar{X}$ with $\mathbb{E}(X_t)$.

2.1.5. Model fitting

Being able to estimate the autocovariance sequence of a time series gives us insight into the structure of the process and is useful to fit a model. Here, we only consider an AR(p) process with mean zero. We present two different methods to fit such a model: the Yule-Walker method and the maximum likelihood method.

Yule-Walker method

Let $\{X_t\}_{t\in T}$ denote an AR(p) process with mean 0 and let $k \in \mathbb{Z}$. We start by multiplying Equation 2.2 by X_{t-k} and take the expectation:

$$\begin{aligned}
X_t X_{t-k} &= -\sum_{j=1}^{p}\alpha_j X_{t-j}X_{t-k} + \epsilon_t X_{t-k} \\
\mathbb{E}(X_t X_{t-k}) &= \mathbb{E}\Big(-\sum_{j=1}^{p}\alpha_j X_{t-j}X_{t-k} + \epsilon_t X_{t-k}\Big) \\
\mathbb{E}(X_t X_{t-k}) &= -\sum_{j=1}^{p}\alpha_j \mathbb{E}(X_{t-j}X_{t-k}) + \mathbb{E}(\epsilon_t X_{t-k}) \\
c_{-k} &= -\sum_{j=1}^{p}\alpha_j c_{k-j} + \sigma^2\delta_{k,0}
\end{aligned} \tag{2.4}$$

where the last line was obtained using $\mathbb{E}(X_t) = 0$ and the fact that X_t is independent of ϵ_s for $s \neq t$.

We introduce the notations $\gamma_p = (c_1, \ldots, c_p)^\top$, $\alpha = (\alpha_1, \ldots, \alpha_p)^\top$ and

$$\Gamma_p = \begin{pmatrix} c_0 & c_1 & \ldots & c_{p-1} \\ c_1 & c_0 & \ldots & c_{p-2} \\ \vdots & \vdots & & \vdots \\ c_{p-1} & c_{p-2} & \ldots & c_0 \end{pmatrix}$$

which is a symmetric Toeplitz matrix. Then, by taking $k = 0, \ldots, p$ and using $c_\tau = c_{-\tau}$ in Equation 2.4 we can write the system of equations in matrix form:

$$\begin{aligned} \gamma_p &= -\Gamma_p \alpha \\ \sigma^2 &= c_0 + \sum_{j=1}^{p} \alpha_j c_j \end{aligned}$$

If we now replace each autocovariance by its sample estimate given in Definition 2.1.8, we obtain the set of equations:

$$\begin{aligned} \hat{\gamma}_p &= -\hat{\Gamma}_p \hat{\alpha} \\ \hat{\sigma}^2 &= \hat{c}_0 + \sum_{j=1}^{p} \hat{\alpha}_j \hat{c}_j \end{aligned}$$

The matrix $\hat{\Gamma}_p$ is positive definite if $\hat{c}_0 > 0$, in which case it is non-singular (Brockwell and Davis (2009)). With this assumption, the Yule-Walker estimates are given by:

$$\hat{\alpha} = -\hat{\Gamma}_p^{-1} \hat{\gamma}_p. \tag{2.5}$$

$$\hat{\sigma}^2 = \hat{c}_0 + \sum_{j=1}^{p} \hat{\alpha}_j \hat{c}_j. \tag{2.6}$$

As a statistical guarantee, Section 8.10 of Brockwell and Davis (2009) shows that the Yule-Walker estimates are unbiased and weakly consistent.

Maximum likelihood

For a sample $X_1, \ldots X_N$ from an AR(p) process, the maximum likelihood estimates are given by:

$$\hat{\alpha}, \hat{\sigma}^2 = \arg\max_{\alpha, \sigma^2} \ell(\alpha, \sigma^2)$$

where $\ell(\alpha, \sigma^2)$ is the log-likelihood, which can be written as:

$$\begin{aligned}
\ell(\alpha, \sigma^2) &= \log(f(X_1, \dots, X_N,\, \alpha, \sigma^2)) \\
&= \log\left(f(X_1, \dots, X_{N-1} \,|\, \alpha, \sigma^2) f(X_N \,|\, X_1, \dots, X_{N-1}, \alpha, \sigma^2)\right) \\
&= \log\left(f(X_1, \dots, X_{N-1} \,|\, \alpha, \sigma^2) f(X_N \,|\, X_1, \dots, X_{N-p},\, \alpha, \sigma^2)\right) \\
\ell(\alpha, \sigma^2) &= \log\left(f(X_1, \dots, X_p \,|\, \alpha, \sigma^2) \prod_{t=p+1}^{N} f(X_t \,|\, X_{t-1}, \dots, X_{t-p}, \alpha, \sigma^2)\right)
\end{aligned}$$

where the last line was obtained by iterating. In order to obtain a closed form approximate solution, we can consider the initial values $X_1, \dots, X_p$ to be deterministic and the innovations to be Gaussian. That way, we only need to work with the conditional log-likelihood, which is known since $X_t \,|\, X_{t-1}, \dots, X_{t-p} \sim \mathcal{N}(-\sum_{j=1}^{p} \alpha_j X_{t-j},\, \sigma^2)$:

$$\ell^c(\alpha, \sigma^2) = \sum_{t=p+1}^{N} \log\left(f(X_t \,|\, X_{t-1}, \dots, X_{t-p},\, \alpha, \sigma^2)\right) \tag{2.7}$$

$$\ell^c(\alpha, \sigma^2) = -\frac{N-p}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=p+1}^{N} \left(X_t - \sum_{j=1}^{p} (-\alpha_j) X_{t-j}\right)^2 \tag{2.8}$$

And we equivalently have:

$$\hat{\alpha}, \hat{\sigma}^2 = \arg\max_{\alpha, \sigma^2} \ell^c(\alpha, \sigma^2). \tag{2.9}$$

Moreover, we notice $\hat{\alpha}$ that maximizes this expression is the least square estimator of $X = -F\alpha + \epsilon$, with:

$$X = \begin{pmatrix} X_{p+1} \\ \vdots \\ X_N \end{pmatrix}, \quad F = \begin{pmatrix} X_p & X_{p-1} & \dots & X_1 \\ X_{p+1} & X_p & \dots & X_2 \\ \vdots & & & \vdots \\ X_{N-1} & X_{N-2} & \dots & X_{N-p} \end{pmatrix}, \quad \alpha = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix}, \quad \epsilon = \begin{pmatrix} \epsilon_{p+1} \\ \vdots \\ \epsilon_N \end{pmatrix}.$$

Hence, its expression and distribution are known (for $N - p > p$, i.e. $N > 2p$):

$$\hat{\alpha} = -(F^\top F)^{-1} F^\top X \;\sim\; \mathcal{N}(\alpha,\, \sigma^2 (F^\top F)^{-1}) \tag{2.10}$$

To find the estimator of σ^2, we differentiate Expression 2.8, where each α_j is replaced by $\hat{\alpha}_j$, with respect to σ^2 and set to zero to obtain the biased estimator:

$$\hat{\sigma}^2 = \frac{(X + F\hat{\alpha})^\top (X + F\hat{\alpha})}{N - p} \tag{2.11}$$

Example 2.1.10 (AR(1) process)**.** Let $X_t + \alpha X_{t-1} = \epsilon_t$ with $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ and $t \in \mathbb{N}$.

Then for $t = 2, \dots, N$, we know that $X_t \,|\, X_{t-1} \sim \mathcal{N}(-\alpha X_{t-1}, \sigma^2)$. Using what is above,

$$X = \begin{pmatrix} X_2 \\ \vdots \\ X_N \end{pmatrix}, \qquad F = \begin{pmatrix} X_1 \\ \vdots \\ X_{N-1} \end{pmatrix}, \qquad \epsilon = \begin{pmatrix} \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix}$$

and

$$\hat{\alpha} = -\frac{\sum_{t=1}^{N-1} X_t X_{t+1}}{\sum_{t=1}^{N-1} X_t^2}$$
$$\hat{\sigma}^2 = \frac{1}{N-1}\left(\sum_{t=2}^{N} X_t^2 + \hat{\alpha} \sum_{t=1}^{N-1} X_t X_{t+1} \right)$$

Remark 2.1.11. These estimates can be expressed as functions of the autocovariance estimators given in Equation 2.3. We notice that they are similar to the Yule-Walker estimates.

The above results illustrate some of the most important models and powerful methodologies used in time series analysis. However, despite their success in various areas as finance (Chan (2004)) or environmental data (Hipel and McLeod (1994), they heavily rely on the assumption of stationarity, which is violated for many time-series we observe in practice. Classical approaches to overcome this issue modify the time series at hand, e.g. by looking at differences to remove a trend or seasonality. Recently, a theory for so-called locally stationary processes has been developed that allows for non-stationarity in the model and avoids modifying the available dataset. In the next section we will investigate such processes.

3. Locally stationary processes

In this section, we introduce the concept of locally stationary processes, namely that a non-stationary process can be locally seen as stationary. Through new models and methods developed by Dahlhaus and others in (Dahlhaus, 2012; Vogt, 2012), we explore how the locally stationary approximation provides a new framework to study such time series.

3.1. The locally stationary approximation

Let $\{X_t\}_{t=1,\dots,T}$ be an observed time series with $T \in \mathbb{N}$ the sample size. Usually in time series analysis, one tends to fit a model on the observations in order to make inference. To carry out that statistical inference, it is sometimes necessary to be able to derive the distributions of various statistics used for the estimation of parameters from the data. However, when dealing with a non-stationary time series, the situation is slightly different. Indeed, although the steps remain the same, making asymptotic considerations has a different meaning. Since present observations hold no information on the probabilistic structure of the process in the future, it is contradictory to make asymptotic consideration. Hence, as $T \to \infty$, any estimator of the time series would become increasingly wrong. To overcome this problem and study estimators for larger sample sizes, we need to work in the framework of infill asymptotics, which originates from non parametric statistics (Chen et al., 2000). This framework can be seen as a rescaling in time: the process is always sampled on the same interval, but as $T \to \infty$, the mesh becomes increasingly finer and the sample density increases. Thus, instead of working with X_t, we work with the triangular array of observations $\{X_{t,T},\ t = 1, \dots, T,\ T \in \mathbb{N}\}$. Therefore, it is clear that infill asymptotics have a very different character than the increasing domain asymptotics familiar from traditional time series analysis.

Figure 3.1 illustrates how the process is sampled on the same interval for two different sample sizes. As we can see, the lower figure has more observations on the same interval than the top figure.

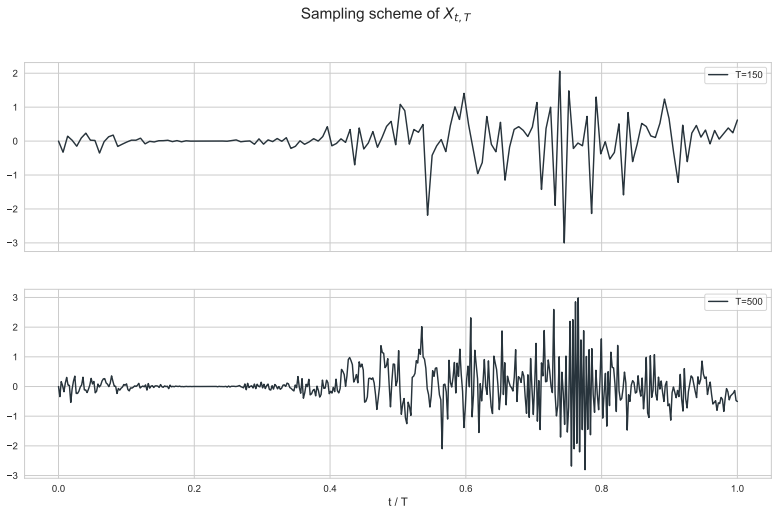

Figure 3.1.: Illustration of the sampling scheme on the interval $[0, 1]$.

Now, in order to fit a model to the observations, one intuitive idea is to assume that the process is locally stationary, meaning that its statistical properties vary slowly enough over time to be locally approximated by some stationary processes. Therefore, in some neighborhood of a fixed point $u_0 = t_0/T \in [0, 1]$, we assume that the process $X_{t,T}$ can be approximated by a stationary process $\tilde{X}_t(u_0)$. Under this assumption, we can estimate some parameters of interest (e.g. the autocovariance sequence) using classic estimation techniques. By studying the properties of these estimates, we can try to evalute the optimal neighborhood in which the process is approximately stationary. We can then iteratively compute these local estimates for different points u_0 covering the interval of interest $[0, 1]$ in order to obtain a sequence of parameters describing the process $X_{t,T}$. More formally, the locally stationary approximation can be defined as follow:

Definition 3.1.1 (Locally Stationary Approximation)**.** Let $q > 0$. Let $\{X_{t,T}\}_{t=1,\dots,T}$, be a triangular array of stochastic processes. For each $u \in [0, 1]$, let $\tilde{X}_t(u)$ be a stationary and ergodic process such that the following holds:

- $\sup_{u\in[0,1]} ||\tilde{X}_t(u)||_q < \infty$
- There exists $C_B > 0$ such that uniformly in $t = 1, \dots, T$ and $u, v \in [0, 1]$,

$$||\tilde{X}_t(u) - \tilde{X}_t(v)||_q \leq C_B\,|u - v|, \qquad ||X_{t,T} - \tilde{X}_t\big(\frac{t}{T}\big)||_q \leq C_B T^{-1}. \tag{3.1}$$

Remark 3.1.2. As we can see, this definition requires both a Lispchitz-like condition on $\tilde{X}_t$ to ensure the stationary approximation is smooth in u, and a mixing condition between $X_{t,T}$ and $\tilde{X}_t$. These conditions allow us to replace $X_{t,T}$ by the stationary approximation $\tilde{X}_t$ with rate $|t/T - u|^\alpha + T^{-\alpha}$.

Furthermore, a powerful result proved in (Dahlhaus et al., 2019) shows that these assumptions hold for any process $X_{t,T}$ defined by the recursion:

$$X_{t,T} = G_{\epsilon_t}\big(X_{t-1,T}, \dots, X_{t-p,T}, \max(\frac{t}{T}, 0)\big), \qquad t \leq T. \tag{3.2}$$

where $(\epsilon_t)_{t\in\mathbb{Z}}$ are i.i.d. random variables and $G : \mathbb{R} \times \mathbb{R}^p \times [0,1] \to \mathbb{R}$. For $u \in [0,1]$ and $t \in \mathbb{Z}$, the stationary approximation $\tilde{X}_t(u)$ is then given by:

$$\tilde{X}_t(u) = G_{\epsilon_t}\big(\tilde{X}_{t-1}(u), \dots, \tilde{X}_{t-p}(u), u\big), \qquad t \in \mathbb{Z}. \tag{3.3}$$

This result will let us apply the locally stationary approximation to some well known models such as AR(p) processes defined in 2.2.

Invariance property of the assumptions with respect to transformations

For $k \in \mathbb{N}$, define $Z_{t,T} = (X_{t,T}, \dots, X_{t-k+1,T})^\top$ and $\tilde{Z}_t(u) = (\tilde{X}_t(u), \dots, \tilde{X}_{t-k+1}(u))^\top$. An interesting feature is that, provided the assumptions from Definition 3.1.1 hold for the process $X_{t,T}$, then they also hold for the process $g(Z_{t,T})$ where g belongs to the following class of functions:

Definition 3.1.3 (class $\mathcal{L}_k(M,C)$)**.** A function $g : \mathbb{R}^k \to \mathbb{R}$ is in the class $\mathcal{L}_k(M,C)$ if there exist $M, C \geq 0$ such that:

$$\sup_{y\neq y'} \frac{|g(y) - g(y')|}{|y - y'|_1(1 + |y|_1^M + |y'|_1^M)} \leq C.$$

Proposition 3.1.4. Let $g \in \mathcal{L}_k(M,C)$. If Assumption 3.1.1 is fulfilled for $X_{t,T}$ with $\tilde{q} = q(M+1) > 0$, then the same assumptions are fulfilled for $g(Z_{t,T})$ with q.

Proof. First of all, let us show that $\sup_{u\in[0,1]} ||g(\tilde{Z}_t(u))||_q < \infty$. Let $u \in [0,1]$ and take $y' = (0, \dots, 0) \in \mathbb{R}^k$ in Definition 3.1.1:

$$\begin{aligned} ||g(\tilde{Z}_t(u)) - g(y')||_q &\leq |C| \left|\left| |\tilde{Z}_t(u)|_1 (1 + |\tilde{Z}_t(u)|_1^M) \right|\right|_q \\ &\leq |C| \left|\left| \sum_{i=0}^{k-1} |\tilde{X}_{t-i}(u)| \left(1 + \big(\sum_{i=0}^{k-1} |\tilde{X}_{t-i}(u)|^M\big)^{1/M}\right) \right|\right|_q \end{aligned}$$

By using Minkowski's inequality and $\sup_{u\in[0,1]} ||\tilde{X}_t(u)||_q < \infty$, we find that the right term is finite and bounded by a constant C'. Hence,

$$\begin{aligned} &||g(\tilde{Z}_t(u))||_q \leq C' + ||g(y')||_q \\ &||g(\tilde{Z}_t(u))||_q < C'' \qquad\qquad \text{where } C'' \text{ is a constant} \end{aligned}$$

As this is true for all $u \in [0,1]$, $\sup_{u\in[0,1]} ||g(\tilde{Z}_t(u))||_q < \infty$.

Then to prove the second condition, we use Definition 3.1.3 and Hölder's inequality with $\frac{1}{q} = \frac{1}{q(M+1)} + \frac{M}{q(M+1)}$:

$$\begin{aligned}
\big|\big|g(Z_{t,T}) - g(\tilde{Z}_t(\frac{t}{T}))\big|\big|_q &\leq |C| \left|\left| |Z_{t,T} - \tilde{Z}_t(\frac{t}{T})|_1 \left(1 + |Z_{t,T}|_1^M + |\tilde{Z}_t(\frac{t}{T})|_1^M\right)\right|\right|_q \\
&\leq |C| \left|\left| |Z_{t,T} - \tilde{Z}_t(\frac{t}{T})|_1 \right|\right|_{q(M+1)} \left|\left| 1 + |Z_{t,T}|_1^M + |\tilde{Z}_t(\frac{t}{T})|_1^M \right|\right|_{q\frac{M+1}{M}}
\end{aligned}$$

For the first term, we use Minkowski's inequality and Definition 3.1.1:

$$\begin{aligned}
\left|\left| |Z_{t,T} - \tilde{Z}_t(\frac{t}{T})|_1 \right|\right|_{q(M+1)} &= \left|\left| \sum_{i=0}^{k-1} |X_{t-i,T} - \tilde{X}_{t-i}(\frac{t}{T})| \right|\right|_{q(M+1)} \\
&\leq \sum_{i=0}^{k-1} \big|\big|X_{t-i,T} - \tilde{X}_{t-i}(\frac{t}{T})\big|\big|_{q(M+1)} \\
&\leq k C_B T^{-1}.
\end{aligned}$$

For the second term, we use Minkowski's inequality and

$$\left|\left| \left(\sum_{i=1}^{n} Y_i \right)^M \right|\right|_{\frac{q}{M}} \leq \sum_{i=1}^{n} ||Y_i||_q^M,$$

which can be proven also using Minkowski's inequality, to bound it as follow:

$$\begin{aligned}
\left|\left| 1 + |Z_{t,T}|_1^M + |\tilde{Z}_t(\frac{t}{T})|_1^M \right|\right|_{q\frac{M+1}{M}} &= \left|\left| 1 + \big(\sum_{i=0}^{k-1} |X_{t-i,T}|\big)^M + \big(\sum_{i=0}^{k-1} |\tilde{X}_{t-i}(\frac{t}{T})|\big)^M \right|\right|_{q\frac{M+1}{M}} \\
&\leq 1 + \left|\left| \big(\sum_{i=0}^{k-1} |X_{t-i,T}|\big)^M \right|\right|_{q\frac{M+1}{M}} + \left|\left| \big(\sum_{i=0}^{k-1} |\tilde{X}_{t-i}(\frac{t}{T})|\big)^M \right|\right|_{q\frac{M+1}{M}} \\
&\leq 1 + \sum_{i=0}^{k-1} \big|\big|X_{t-i,T}\big|\big|_{q(M+1)}^M + \sum_{i=0}^{k-1} \big|\big|\tilde{X}_{t-i}(\frac{t}{T})\big|\big|_{q(M+1)}^M \\
&< C
\end{aligned}$$

where C is a constant since $\sup_{u\in[0,1]} ||\tilde{X}_t(u)||_{q(M+1)}$ is finite by assumption and $||X_{t,T}||_{q(M+1)} \leq ||X_{t,T} - \tilde{X}_t(\frac{t}{T})||_{q(M+1)} + ||\tilde{X}_t(\frac{t}{T})||_{q(M+1)} < \infty$.

Hence, we can write $||g(Z_{t,T}) - g(\tilde{Z}_t(\frac{t}{T}))||_q \leq C_{B,1} T^{-1}$, with $C_{B,1} > 0$ a constant. The other condition $||g(\tilde{Z}_t(u)) - g(\tilde{Z}_t(v))||_q \leq C_{B,2}|u - v|$ is shown the same way.

Taking $C'_B = \max(C_{B,1}, C_{B,2}) > 0$ concludes the proof. □

This invariance property is crucial, as it implies that most asymptotic results on $X_{t,T}$ also hold for $g(Z_{t,T})$ with $g \in \mathcal{L}_{k+1}(M, C)$, under appropriate moment conditions. This is particularily important as we sometimes need to estimate parameters of $X_{t,T}$ that can be written under the form $g(Z_{t,T})$ using the method of moments.

Example 3.1.5. *covariance estimator*: $g : (x_0, \dots, x_k) \mapsto x_0 x_k \in \mathcal{L}_{k+1}(1, 1),\ k \in \mathbb{N}$. This implies that, for $k \in \mathbb{N}$, the process $X_{0,T} X_{k,T}$ also has a locally stationary approximation which is important to recover the autocovariance structure from a sample. This estimator will be used in Section 3.3.

3.2. Introductory example: time-varying autoregressive process tvAR(1)

In this subsection, we consider the simple case of a time-varying autoregressive process of order 1 $\big(\text{tvAR}(1)\big)$ to introduce some of the concepts of locally stationary processes and give some intuition. Such a process satisfies the equation:

$$X_{t,T} + \alpha\left(\frac{t}{T}\right) X_{t-1,T} = \sigma\left(\frac{t}{T}\right)\epsilon_t, \qquad t \in \mathbb{Z} \tag{3.4}$$

with $\alpha : [0,1] \to (-1,1)$, $\sigma : [0,1] \to \mathbb{R}_+$ both Lipschitz continuous and ϵ_t i.i.d. $\mathcal{N}(0,1)$. Let $u_0 = t_0/T \in [0,1]$. In some neighborhood of u_0, $X_{t,T}$ can be approximated by the stationary process $\tilde{X}_t(u_0)$ defined by:

$$\tilde{X}_t(u_0) + \alpha(u_0)\tilde{X}_{t-1}(u_0) = \sigma(u_0)\epsilon_t, \quad t \in \mathbb{Z} \tag{3.5}$$

$\tilde{X}_t(u_0)$ can be written under its general linear process (GLP) form:

$$\tilde{X}_t(u_0) = \sum_{k=0}^{\infty} (-1)^k \alpha(u_0)^k \sigma(u_0)\epsilon_{t-k} \tag{3.6}$$

$$\tilde{X}_t(u_0) = \sum_{k=-\infty}^{\infty} a(u_0, k)\epsilon_{t-k} \tag{3.7}$$

$$\text{with } a(u,k) = \begin{cases} 0 & k < 0 \\ (-1)^k \alpha(u_0)^k \sigma(u_0) & k \geq 0 \end{cases} \tag{3.8}$$

Proposition 3.2.1. The process $\tilde{X}_t(u_0)$ given in Equation 3.5 satisfies Definition 3.1.1 for $q = 2$.

Proof. Let $u \in [0,1]$ and $t \in \mathbb{Z}$. First we show that $\sup_{u\in[0,1]} ||\tilde{X}_t(u)||_2 < \infty$.

$$\begin{aligned}||\tilde{X}_t(u)||_2 &= ||-\alpha(u)\tilde{X}_{t-1}(u) + \sigma(u)\epsilon_t||_2 \\ &\leq |\alpha(u)|\,||\tilde{X}_{t-1}(u)||_2 + |\sigma(u)|\,||\epsilon_t||_2 \qquad \text{Minkowski inequality}\end{aligned}$$

Because α is continuous on the segment $[0,1]$, it is bounded and attains its maximum. Hence, there exists $u_* \in [0,1]$ such that $\sup_{u\in[0,1]} \alpha(u) = \alpha(u_*) \in (-1,1)$. In other words, $\sup_{u\in[0,1]} |\alpha(u)| < 1$. Similarily, $\sup_{u\in[0,1]} |\sigma(u)| < \infty$. Thus we have:

$$\begin{aligned}||\tilde{X}_t(u)||_2 &\leq \sup_{u\in[0,1]} |\alpha(u)|\,||\tilde{X}_{t-1}(u)||_2 + \sup_{u\in[0,1]} |\sigma(u)|\,||\epsilon_t||_2 \\ &\leq \sup_{u\in[0,1]} |\alpha(u)|\,||\tilde{X}_t(u)||_2 + \sup_{u\in[0,1]} |\sigma(u)|.\end{aligned}$$

Since $\tilde{X}_{t-1}(u) = \tilde{X}_t(u)$ in distribution (by stationarity) and $||\epsilon_t||_2 = \mathrm{Var}(\epsilon_t)^{1/2} = 1$. Finally,

$$\sup_{u\in[0,1]} ||\tilde{X}_t(u)||_2 \leq \frac{\sup_{u\in[0,1]} |\sigma(u)|}{1 - \sup_{u\in[0,1]} |\alpha(u)|} < \infty.$$

Let $u, v \in [0,1]$ and $t \in \mathbb{Z}$. We now show that $||\tilde{X}_t(u) - \tilde{X}_t(v)||_2 < C_{B,1}|u-v|$.

$$\begin{aligned}&||\tilde{X}_t(u) - \tilde{X}_t(v)||_2 \\ &\quad= ||-\alpha(u)\tilde{X}_{t-1}(u) + \sigma(u)\epsilon_t + \alpha(v)\tilde{X}_{t-1}(v) - \sigma(v)\epsilon_t||_2 \\ &\quad= ||\big(\alpha(v)-\alpha(u)\big)\tilde{X}_{t-1}(u) + \big(\sigma(u)-\sigma(v)\big)\epsilon_t + \alpha(v)\big(\tilde{X}_{t-1}(v) - \tilde{X}_{t-1}(u)\big)||_2 \\ &\quad\leq ||(\alpha(v)-\alpha(u))\tilde{X}_{t-1}(u)||_2 + ||(\sigma(u)-\sigma(v))\epsilon_t||_2 + ||\alpha(v)(\tilde{X}_{t-1}(v) - \tilde{X}_{t-1}(u))||_2\end{aligned}$$

According to the Lispchitz assumption on α and σ, there exist $C_\alpha, C_\sigma > 0$ such that

$$|\alpha(v)-\alpha(u)| \leq C_\alpha|u-v| \qquad \text{and} \qquad |\sigma(v)-\sigma(u)| \leq C_\sigma|u-v|.$$

Moreover, $||\tilde{X}_{t-1}(v) - \tilde{X}_{t-1}(u)||_2 = ||\tilde{X}_t(v) - \tilde{X}_t(u)||_2$ and $||\epsilon_t||_2 = 1$. Therefore,

$$\begin{aligned}&||\tilde{X}_t(u) - \tilde{X}_t(v)||_2 \qquad (3.9)\\ &\quad\leq C_\alpha|u-v| \sup_{u\in[0,1]} ||\tilde{X}_t(u)||_2 + C_\sigma|u-v| + \sup_{u\in[0,1]} |\alpha(u)|\,||\tilde{X}_t(u) - \tilde{X}_t(v)||_2\end{aligned}$$

By isolating $||\tilde{X}_t(u) - \tilde{X}_t(v)||_2$ on the left side of the inequality, we find:

$$\begin{aligned}||\tilde{X}_t(u) - \tilde{X}_t(v)||_2 &\leq \frac{C_\alpha \sup_{u\in[0,1]} ||\tilde{X}_t(u)||_2 + C_\sigma}{1 - \sup_{u\in[0,1]} |\alpha(u)|}|u-v| \\ &\leq C_{B,1}|u-v|. \qquad (3.10)\end{aligned}$$

Finally, we show that for $t \in \{1, \dots T\}$, $||X_{t,T} - \tilde{X}_t(t/T)||_2 \leq C_{B,2}T^{-1}$

$$\begin{aligned}
||X_{t,T} - \tilde{X}_t(\frac{t}{T})||_2 &= || - \alpha(\frac{t}{T})X_{t-1,T} + \alpha(\frac{t}{T})\tilde{X}_{t-1}(\frac{t}{T})||_2 \\
&= |\alpha(\frac{t}{T})|\,||X_{t-1,T} - \tilde{X}_{t-1}(\frac{t}{T})||_2 \\
&\leq \sup_{u\in[0,1]} |\alpha(u)|\,||X_{t-1,T} - \tilde{X}_{t-1}(\frac{t}{T})||_2
\end{aligned}$$

We develop this inequality further:

$$\begin{aligned}
||X_{t,T} - \tilde{X}_t(\frac{t}{T})||_2 &\leq \alpha(u_*)\left(||X_{t-1,T} - \tilde{X}_{t-1}(\frac{t-1}{T})||_2 + ||\tilde{X}_{t-1}(\frac{t-1}{T}) - \tilde{X}_{t-1}(\frac{t}{T})||_2\right) \\
&\leq \alpha(u_*)\left(\left|\left|X_{t-1,T} - \tilde{X}_{t-1}(\frac{t-1}{T})\right|\right|_2 + \frac{C_{B,1}}{T}\right) \quad \text{using Equation 3.10}
\end{aligned}$$

Let us call $z_t = ||X_{t,T} - \tilde{X}_t(\frac{t}{T})||_2$. Since the definition of $X_{t,T}$ and $\tilde{X}_t(0)$ coincide for $t \leq 0$, we have $z_t = 0$ for $t \leq 0$. Hence,

$$\begin{aligned}
z_t &\leq \alpha(u_*)(z_{t-1} + \frac{C_{B,1}}{T}) \\
&\leq \alpha(u_*)^2 z_{t-2} + \frac{C_{B,1}}{T}\alpha(u_*)(1 + \alpha(u_*)) \\
&\leq \alpha(u_*)^t z_0 + \frac{C_{B,1}}{T}\alpha(u_*)\sum_{i=0}^{t-1}\alpha(u_*)^i && \text{by iterating} \\
&\leq 0 + \frac{C_{B,1}}{T}\alpha(u_*)\sum_{i=0}^{\infty}\alpha(u_*)^i && \text{since } z_0 = 0 \text{ and } \alpha(u_*) \geq 0 \\
z_t &\leq \frac{\alpha(u_*)}{1-\alpha(u_*)}C_{B,1}T^{-1} = C_{B,2}T^{-1} && \text{since } |\alpha(u_*)| < 1
\end{aligned}$$

Taking $C_B = \max(C_{B,1}, C_{B,2}) > 0$ concludes the proof. □

We now know the form of the stationary approximation in the case of a tvAR(1) process. Let us explore how to use it to make inference on the time series.

3.2.1. Local estimation by stationary methods on segments

Let us consider we have a realization of the process $X_{t,T}$. We are interested in inferring $t \mapsto \alpha(t)$ and $t \mapsto \sigma(t)$. The intuitive approach is to replace $X_{t,T}$ by its stationary approximation $\tilde{X}_t(u_0)$ on a small segment around a fixed u_0, say $\{t;\, |t/T - u_0| \leq \delta\}$, where the process is almost stationary. Using this approximation, we can estimate $\alpha(u_0)$ and $\sigma(u_0)$ with classical stationary approaches described in the previous section. Then, we shift the segment by choosing another point u_0 and repeat the estimation, leading to a sequence of estimates for the time varying parameters. Figure 3.2 illustrates this

iterative approach.

Remark 3.2.2. Because this approach relies on $\tilde{X}_t(u_0)$ and not the process $X_{t,T}$ itself, it will necessarily induce a bias. This bias can be evaluated and minimized for a particular segment length.

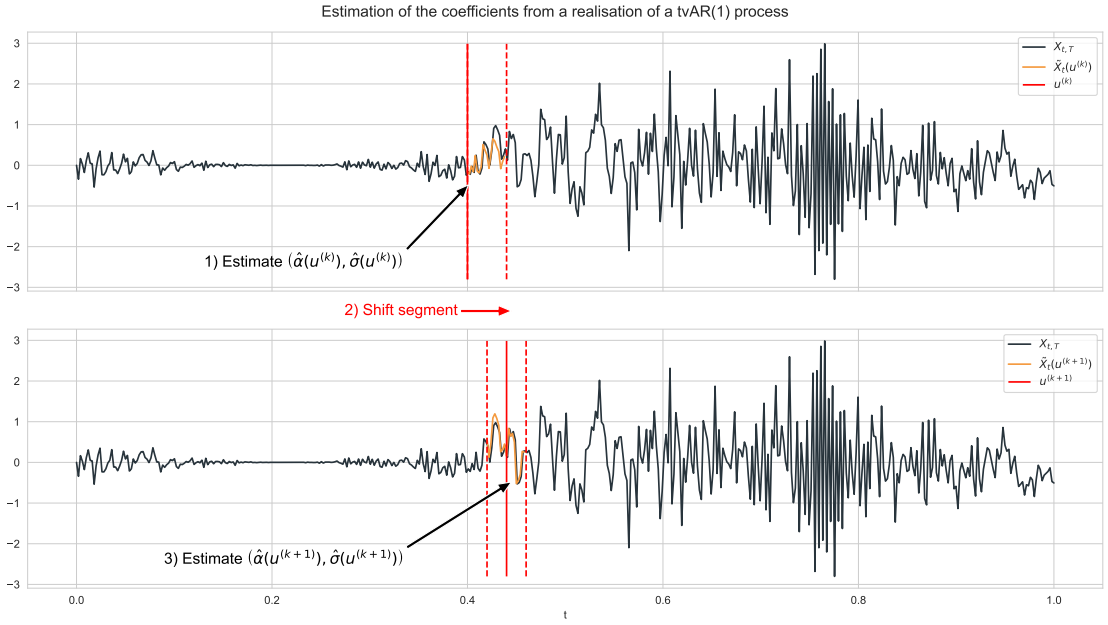

Figure 3.2.: Iterative approach: 1) select a point $u_0 \in [0,1]$ and compute the Yule-Walker estimates on a window around that point. 2) choose the next point u_0 and shift the segment around it. 3) Repeat the fist two steps until the interval of interest $[0,1]$ has been covered and a sequence of estimates has been obtained to approximate the parameter curves.

Yule-Walker estimators

Let $u_0 \in [0,1]$ and $k \in \mathbb{Z}$. For $t \in \mathbb{Z}$, let $c(u_0, k) = \mathrm{Cov}(\tilde{X}_t(u_0), \tilde{X}_{t-k}(u_0))$.
At u_0, $\tilde{X}_t(u_0)$ is an AR(1) process. Thus, as described in Section 2.1.4, the Yule-Walker estimates are given by:

$$\hat{\alpha}(u_0) = -\frac{\hat{c}(u_0,1)}{\hat{c}(u_0,0)} \qquad \text{and} \qquad \hat{\sigma}^2(u_0) = \hat{c}(u_0,0) + \hat{\alpha}(u_0)\hat{c}(u_0,1) \tag{3.11}$$

where, for $k \in \mathbb{Z}$, $\hat{c}(u_0,k)$ is an estimator of the autocovariance sequence of $\tilde{X}_t(u_0)$ at lag k. Because $\tilde{X}_t(u_0)$ is unknown in practice, $\hat{c}(u_0,k)$ will in fact be a local estimate of $\mathrm{Cov}(X_{t,T}, X_{t+k,T})$ around u_0. This notion of a localized estimator requires the introduction of localizing kernels which are special weighting functions.

Definition 3.2.3 (Localizing kernel)**.** The function $K : \mathbb{R} \to \mathbb{R}$ is a localizing kernel if:

- K is symmetric: for all $x \in \mathbb{R}$, $K(x) = K(-x)$.
- K has a compact support: there exists $c > 0$ such that for all $|x| > c$, $K(x) = 0$.
- $K : \mathbb{R} \to \mathbb{R}$ is piecewise differentiable with $\int_{\mathbb{R}} K(x)dx = 1$ and $\sup_{x\in\mathbb{R}} |K(x)| < \infty$.

Example 3.2.4. There are several well-known kernels such as:

1. Epanechnikov kernel: $K(x) = \frac{3}{4}(1 - x^2)\mathbb{1}_{[-1,1]}(x)$.
2. Uniform kernel: $K(x) = \frac{1}{2}\mathbb{1}_{[-1,1]}(x)$.
3. Triangular kernel: $K(x) = (1 - |x|)\mathbb{1}_{[-1,1]}(x)$.

They are represented Figure 3.3.

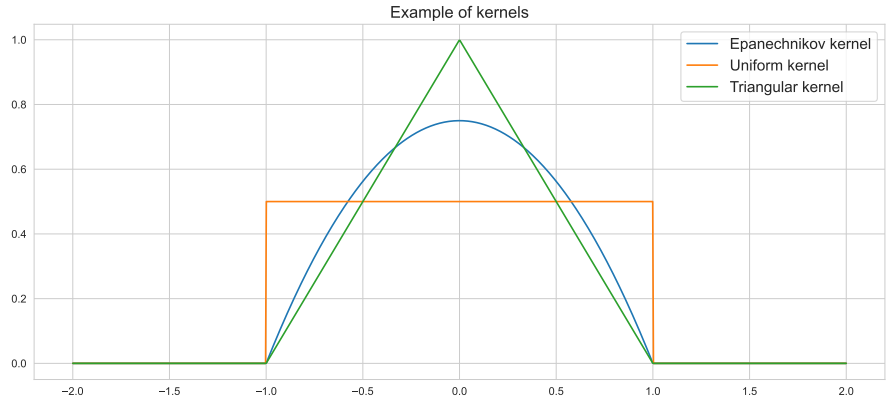

Figure 3.3.: Plot of different kernels.

We can now define the estimator of the autocovariance sequence of $\tilde{X}_t(u_0)$.

Definition 3.2.5 (Kernel estimators of the covariance)**.** Let K be a localizing kernel as defined in Definition 3.2.3 and $b_T > 0$ a bandwidth. Then we can estimate $c(u_0, k)$ by:

$$\tilde{c}_T(u_0, k) = \frac{1}{b_T T} \sum_{t=1}^{T-k} K\left(\frac{u_0 - (t + k/2)/T}{b_T}\right) X_{t,T} X_{t+k,T}. \tag{3.12}$$

Or equivalently, with $|i - j| = k$:

$$\tilde{\tilde{c}}_T(u_0, i, j) = \frac{1}{b_T T} \sum_{t=\max(i,j)}^{T+\min(i,j)} K\left(\frac{u_0 - t/T}{b_T}\right) X_{t-i,T} X_{t-j,T}. \tag{3.13}$$

These two estimators are completely equivalent in the sense that they share the same asymptotic properties. We therefore adopt the general notation $\hat{c}(u_0, k)$ to refer to any of the two.

Remark 3.2.6. The bandwidth b_T is a parameter that determines the smoothness of the estimator and that needs to be tuned. A small bandwidth leads to undersmoothing resulting in an irregular estimator while a large value leads to an oversmoothed and more biased estimator. The bandwidth selection is then a problem on its own and consists of finding a tradeoff between the two. It is the subject of active research.

For a localizing kernel with support $[-1, 1]$, the kernel will smooth observations in the window $[u_0 - b_T,\ u_0 + b_T]$. The segment length is thus $N_T := 2b_T$.

In practice, we often choose b_T such that $b_T \to 0$ and $b_T T \to \infty$ as $T \to \infty$ so that the window becomes increasingly smaller but contains increasingly more observations.

Coming back to the Yule-Walker estimates in the non-stationary setting defined in Definition 3.11, we obtain the following important first result.

Proposition 3.2.7. The Yule-Walker estimates defined in Equation 3.11 are consistent:

$$\hat{\alpha}(u_0) \xrightarrow[T\to\infty]{p} \alpha(u_0) \tag{3.14}$$

$$\hat{\sigma}^2(u_0) \xrightarrow[T\to\infty]{p} \sigma^2(u_0) \tag{3.15}$$

Proof. The proof requires to study the asymptotics of locally stationary processes and will be derived in Section 3.3. □

Maximum likelihood estimator

An important class of estimators in time series analysis are maximum likelihood estimators. These estimates are computed by maximizing the localized version of the conditional log-likelihood. Again, this is a modification of the stationary method recalled in Subsection 2.1.5, using localizing kernels.

Let K denote a localizing kernel and $b_T > 0$ its associated bandwidth. Let us assume that $b_T \to 0$ and $b_T T \to \infty$ as $T \to \infty$. We also define the local conditional log-likelihood of a tvAR(p) process:

$$\mathcal{L}_T^C(u_0, \alpha, \sigma^2) = \frac{1}{b_T T} \sum_{t=p+1}^{T} K\left(\frac{u_0 - t/T}{b_T}\right) \ell_{t,T}(\alpha, \sigma^2), \qquad u_0 \in [0, 1] \tag{3.16}$$

with $\ell_{t,T}(\alpha, \sigma^2) = \log f(X_{t,T} \,|\, X_{t-1,T}, \ldots, X_{1,T},\ \alpha, \sigma^2)$.

In the context of stationary approximations, we will approximate $\mathcal{L}_T^C(u_0, \alpha, \sigma^2)$ by $\tilde{\mathcal{L}}_T^C(u_0, \alpha, \sigma^2)$, which is the same function with $\ell_{t,T}(\alpha, \sigma^2)$ replaced by:

$$\tilde{\ell}_{t,T}(\alpha, \sigma^2) = \log f\big(\tilde{X}_t(u_0) \,|\, \tilde{X}_{t-1}(u_0), \ldots, \tilde{X}_1(u_0),\ \alpha, \sigma^2\big).$$

In the case of the tvAR(1) model given in Equation 3.4 with gaussian innovations, we

have for $t = 2, \dots, T$ that $X_{t,T} \,|\, X_{t-1,T} \sim \mathcal{N}\big(-\alpha(t/T)X_{t-1,T},\, \sigma(t/T)^2\big)$. Therefore:

$$\ell_{t,T}(\alpha, \sigma^2) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}(X_{t,T} + \alpha X_{t-1,T})^2.$$

Hence for $u_0 \in [0,1]$, we find $\hat{\alpha}$ and $\hat{\sigma}^2$ by differenciation:

$$\begin{aligned}
&\frac{\partial \mathcal{L}_T^C}{\partial \alpha}(u_0, \hat{\alpha}, \sigma^2) = 0 \\
&\qquad \Leftrightarrow 0 = \frac{1}{bT}\sum_{t=2}^{T} K\left(\frac{u_0 - t/T}{b}\right)\frac{\partial \ell}{\partial \alpha}(\hat{\alpha}, \sigma^2) && \text{by linearity} \\
&\qquad \Leftrightarrow 0 = \frac{1}{bT}\sum_{t=2}^{T} K\left(\frac{u_0 - t/T}{b}\right)\big[\frac{X_{t-1,T}}{\sigma^2}(X_{t,T} + \hat{\alpha} X_{t-1,T})\big] \\
&\qquad \Leftrightarrow 0 = \frac{1}{bT\sigma^2}\sum_{t=2}^{T} K\left(\frac{u_0 - t/T}{b}\right) X_{t,T}X_{t-1,T} + \frac{\hat{\alpha}}{bT\sigma^2}\sum_{t=2}^{T} K\left(\frac{u_0 - t/T}{b}\right) X_{t-1,T}^2 \\
&\qquad \Leftrightarrow \hat{\alpha} = -\frac{\frac{1}{bT}\sum_{t=2}^{T} K\left(\frac{u_0 - t/T}{b}\right) X_{t,T}X_{t-1,T}}{\frac{1}{bT}\sum_{t=2}^{T} K\left(\frac{u_0 - t/T}{b}\right) X_{t-1,T}^2} \\
&\qquad \Leftrightarrow \hat{\alpha} = -\frac{\tilde{\tilde{c}}_T(u_0, 0, 1)}{\tilde{\tilde{c}}_T(u_0, 1, 1)} && \text{with the translation } t' = t+1 \text{ in both sums}
\end{aligned}$$

and

$$\begin{aligned}
&\frac{\partial \mathcal{L}_T^C}{\partial \sigma^2}(u_0, \alpha, \hat{\sigma}^2) = 0 \\
&\qquad \Leftrightarrow 0 = \frac{1}{bT}\sum_{t=2}^{T} K\left(\frac{u_0 - t/T}{b}\right)\frac{\partial \ell}{\partial \sigma^2}(\alpha, \hat{\sigma}^2) \\
&\qquad \Leftrightarrow 0 = \frac{1}{bT}\sum_{t=2}^{T} K\left(\frac{u_0 - t/T}{b}\right)\big[\frac{1}{2\hat{\sigma}^2} - \frac{1}{2\hat{\sigma}^4}(X_{t,T} + \alpha X_{t-1,T})^2\big] \\
&\qquad \Leftrightarrow 0 = \frac{1}{bT}\sum_{t=2}^{T} K\left(\frac{u_0 - t/T}{b}\right)\big[\hat{\sigma}^2 - (X_{t,T}^2 + 2\alpha X_{t,T}X_{t-1,T} + \alpha^2 X_{t-1,T}^2)\big] \\
&\qquad \Leftrightarrow 0 = \hat{\sigma}^2\frac{1}{bT}\sum_{t=2}^{T} K\left(\frac{u_0 - t/T}{b}\right) - \tilde{\tilde{c}}_T(u_0, 0, 0) - 2\alpha\tilde{\tilde{c}}_T(u_0, 0, 1) - \alpha^2\tilde{\tilde{c}}_T(u_0, 1, 1).
\end{aligned}$$

Then, by replacing α with its estimator $\hat{\alpha}$, this is equivalent to:

$$0 = \hat{\sigma}^2 \frac{1}{bT} \sum_{t=2}^{T} K\left(\frac{u_0 - t/T}{b}\right) - \big(\tilde{\tilde{c}}(u_0, 0, 0) + \hat{\alpha}\tilde{\tilde{c}}(u_0, 0, 1)\big).$$

Since when T is large enough $\frac{1}{bT}\sum_{t=2}^{T} K\left(\frac{u_0 - t/T}{b}\right) \simeq \int K(x)dx = 1$, we then obtain:

$$\sigma^2 \simeq \tilde{\tilde{c}}(u_0, 0, 0) + \hat{\alpha}\tilde{\tilde{c}}(u_0, 0, 1)$$

As we can see, these estimates are equivalent to those obtained with the Yule-Walker equations and thus share the same properties.

3.2.2. Properties of the estimates and optimal segment length

As we can see from this tvAR(1) example, for a fixed $u_0 \in [0, 1]$, both the Yule-Walker and the Maximum Likelihood estimators can be expressed as a function of $\hat{c}(u_0, k)$, where $k \in \{0, 1\}$. It is therefore important to know the properties of $\hat{c}(u_0, k)$ to derive information on the estimates $\hat{\alpha}(u_0)$ and $\hat{\sigma}^2(u_0)$. Interestingly, a theoretical optimal bandwidth can be found by minimizing the mean squared error of that estimate, leading to the optimal window size.

Proposition 3.2.8 (Properties of $\hat{c}(u_0, k)$)**.**
Let $\mu(u_0) = \frac{\partial^2}{\partial^2 u_0} c(u_0, k)$, $\tau(u_0) = \sum_{l=-\infty}^{\infty} c(u_0, l)[c(u_0, l) + c(u_0, l+2k)]$, $d_K = \int x^2 K(x)dx$ and $v_K = \int K(x)^2 dx$. We have:

(i) $\mathbb{E}(\hat{c}(u_0, k)) = c(u_0, k) + \frac{b_T^2}{2} d_K \mu(u_0) + o(b_T^2) + O(\frac{1}{b_T T})$.

(ii) $\mathrm{Var}(\hat{c}(u_0, k)) = \frac{1}{b_T T} v_K \tau(u_0) + o(\frac{1}{b_T T})$.

(iii) $\mathrm{MSE}(\hat{c}(u_0, k)) = \frac{b_T^4}{4} d_K^2 \mu(u_0)^2 + \frac{1}{b_T T} v_K \tau(u_0) + o(b_T^4 + \frac{1}{b_T T})$.

Proof. The expressions of the expectation and variance of $\hat{c}(u_0, k)$ were provided in (Dahlhaus, 1996). They rely on the spectral representation of $X_{t,T}$, which are beyond the scope of the present work. For that reason, we only prove the result for the MSE and refer to (Dahlhaus, 1996) for the precise derivation of (i) and (ii).

(iii) For clarity we write $c = c(u_0, k)$ and $\hat{c} = \hat{c}(u_0, k)$.

$$\begin{aligned}
\mathrm{MSE}\big(\hat{c}(u_0,k)\big) &= \mathbb{E}(|\hat{c}(u_0,k) - c(u_0,k)|^2) \\
&= \mathbb{E}(|\hat{c} - c|^2) \\
&= \mathbb{E}(\hat{c}^2) - 2c\mathbb{E}(\hat{c}) + c^2 \\
&= \mathrm{Var}(\hat{c}) + \mathbb{E}(\hat{c})^2 - 2c\mathbb{E}(\hat{c}) + c^2 \\
&= \frac{1}{b_T T} v_K \tau(u_0) + o(\frac{1}{b_T T}) + \big(c + \frac{b_T}{2} d_K \mu(u_0) + o(b_T^2) + O(\frac{1}{b_T T})\big)^2 \\
&\quad - 2c\big(c + \frac{b_T^2}{2} d_K \mu(u_0)\big) + c^2 \\
&= \frac{b_T^4}{4} d_K^2 \mu(u_0)^2 + \frac{1}{b_T T} v_K \tau(u_0) + o(b_T^4 + \frac{1}{b_T T}).
\end{aligned}$$

□

Minimizing the MSE leads to the expression of the optimal window.

Theorem 3.2.9. The optimal window in the sense that it minimizes the mean squared error of $\hat{c}(u_0, k)$ is given by:

$$N_{\mathrm{opt}} = T b_{\mathrm{opt}} = T^{4/5} C(K_{\mathrm{opt}})^{1/5} \left[\frac{\tau(u_0)}{\mu(u_0)^2}\right]^{1/5} \tag{3.17}$$

$$K_{\mathrm{opt}}(x) = 6x(1-x)\mathbb{1}_{[0,1]}(x) \tag{3.18}$$

where $C(K) = v_K / d_K^2$

Proof. The proof of this theorem, which can be found in (Dahlhaus and Giraitis, 1998), is quite long and uses concepts we do not cover in this work. The idea is to minimize the function $f(b, K) \coloneqq \frac{b^4}{4} d_K^2 \mu(u_0)^2 + \frac{1}{bT} v_K \tau(u_0)$. □

Remark 3.2.10. As we can see, the optimal segment length N_{opt} increases when $\mu(u_0)$ decreases. This relationship was expected since $\mu(u_0)$ captures the degree of non stationarity of the process. Indeed, it only takes small values if $c(u_0, k)$ is close to constant or linear in time and takes large values for quick variations.

Therefore, if the non stationarity is significant, i.e. $\mu(u_0)$ is large, the locally stationary approximation is only working well on a small neighborhood of u_0.

This formula also raises a challenging issue: the optimal segment length requires information about the process that we do not have. Thus, we can't adaptively determine the optimal window from the observed process and this remains an active topic of research. Recently, (Richter and Dahlhaus, 2019) proposed an adaptive bandwidth selector that relies on cross-validation.

3.3. Asymptotics for locally stationary processes

In this subsection, we study the asymptotics of locally stationary processes. To do so, we look at the convergence of local estimates, such as the local estimators of the autocovariance sequence, as $T \to \infty$. Indeed, they are particularily important since we saw that the parameters of a tvAR(1) process depended only on the autocovariance sequence, but this also holds for tvAR(p) processes, with $p > 1$.

3.3.1. Localized law of large numbers and convergence results

Lemma 3.3.1. Assume that Y_t is a stationary and ergodic process with $\mathbb{E}(Y_1) < \infty$. Let $u \in (0,1)$ and $b_T > 0$ such that $b_T \to 0$ and $b_T T \to \infty$ as $T \to \infty$. Then, the following convergence holds in L^1:

$$\frac{1}{b_T T} \sum_{t=1}^{T} K\left(\frac{t/T - u}{b_T}\right) Y_t \xrightarrow[T\to\infty]{L^1} \mathbb{E}(Y_0)$$

The proof of this lemma can be found in (Dahlhaus and Rao, 2006).

Theorem 3.3.2 (Localized law of large numbers)**.** Suppose that Definition 3.1.1 holds and that K is a localizing kernel with bandwidth b_T. Then, for each $u \in (0,1)$:

$$\frac{1}{b_T T} \sum_{t=1}^{T} K\left(\frac{t/T - u}{b_T}\right) X_{t,T} \xrightarrow[T\to\infty]{L^1} \mathbb{E}(\tilde{X}_0(u)) \tag{3.19}$$

Proof. (Dahlhaus et al., 2019)

$$\begin{aligned}
&\left|\left|\frac{1}{b_T T} \sum_{t=1}^{T} K\left(\frac{t/T - u}{b_T}\right) X_{t,T} - \mathbb{E}(\tilde{X}_0(u))\right|\right|_1 \\
&\leq ||\frac{1}{b_T T} \sum_{t=1}^{T} K\left(\frac{t/T - u}{b_T}\right) (X_{t,T} - \tilde{X}_t(u))||_1 + ||\frac{1}{b_T T} \sum_{t=1}^{T} K\left(\frac{t/T - u}{b_T}\right) \tilde{X}_t(u) - \mathbb{E}(\tilde{X}_0(u))||_1.
\end{aligned}$$

The second term tends to 0 as $T \to \infty$ by Lemma 3.3.1. For the first term:

$$\begin{aligned}
&\left|\left|\frac{1}{b_T T} \sum_{t=1}^{T} K\left(\frac{t/T - u}{b_T}\right) (X_{t,T} - \tilde{X}_t(u))\right|\right|_1 \\
&\qquad \leq \sup_{x\in\mathbb{R}} |K(x)| \left(\sup_{t=1,\dots,T} ||X_{t,T} - \tilde{X}_t(\frac{t}{T})||_1 + \sup_{|u-v|\leq b_T/2} ||\tilde{X}_t(u) - \tilde{X}_t(v)||_1 \right) \\
&\qquad \xrightarrow[T\to\infty]{} 0
\end{aligned}$$

Using the definition of the locally stationary approximation and that $b_T \xrightarrow[T\to\infty]{} 0$. □

Using this theorem, we can now show the convergence of the local estimate of the auto-

covariance sequence.

Corollary 3.3.3. Let $u \in (0,1)$ and $k \in \{1, \dots, T\}$. Then, if $\mathbb{E}(\tilde{X}_t(u)) = 0$:

$$\hat{c}(u,k) = \frac{1}{b_T T} \sum_{t=1}^{T-k} K\left(\frac{t/T - u}{b_T}\right) X_{t,T} X_{t+k,T} \xrightarrow[T\to\infty]{L^1} c(u,k) = \mathrm{Cov}\big(\tilde{X}_t(u),\, \tilde{X}_{t-k,T}(u)\big).$$

Proof. Let us consider the covariance estimator at lag k: $g : (x_0, \dots, x_k) \mapsto x_0 x_k$. First, we show that $g \in \mathcal{L}_{k+1}(1,1)$. Let $y \neq y' \in \mathbb{R}^{k+1}$,

$$\begin{aligned}
|g(y) - g(y')| &= |y_0 y_k - y'_0 y'_k| \\
&\leq |y_0(y_k - y'_k)| + |y'_k(y_0 - y'_0)| \\
&\leq |y_0|\,|y_k - y'_k| + |y'_k|\,|y_0 - y'_0| \\
&\leq |y|_1\,|y - y'|_1 + |y'|_1\,|y - y'|_1 \\
&\leq |y - y'|_1(|y|_1 + |y'|_1) \\
&\leq |y - y'|_1(1 + |y|_1 + |y'|_1)
\end{aligned}$$

This holds for any $y \neq y'$ so $g \in \mathcal{L}_{k+1}(1,1)$. Hence, we can apply Theorem 3.3.2 on the process $g(X_{t,T}, \dots, X_{t-k+1,T})$ which proves the result. □

Since the Yule-Walker estimates of a tvAR(p) model are identifiable (Tsay and Tiao, 1984), this result directly shows their consistency.

3.3.2. Simulations of the Yule-Walker estimates for a tvAR(1) process

In this section, we perform a Monte Carlo simulation to study the behaviour and convergence of the Yule-Walker estimates introduced in Subsection 3.2.1, in the case of a tvAR(1) process.

We start by generating 500 realisations of the tvAR(1) model defined in Equation 3.4 for different sample sizes, starting at 0 and with true coefficient functions:

$$\alpha(u) = -0.8\cos\big(1.5 - \cos(4\pi u)\big), \qquad \sigma(u) = \cos(u\frac{\pi}{2} + e^u)^2, \qquad u \in [0,1].$$

Then for each sample size T and each realisation, we estimate the coefficient functions $\hat{\alpha}$ and $\hat{\sigma}$ at equidistant points $u_1, \dots, u_{100} \in [0,1]$ using the Epanechnikov kernel from Example 3.2.4.
The bandwidth parameter has to be manually chosen since the optimal value cannot be used directly. In this experiment, we chose $b_T = 0.1\,T^{-1/5}$, such that $b_T \xrightarrow[T\to\infty]{} 0$, $b_T T \xrightarrow[T\to\infty]{} \infty$ and that the localizing kernel smooths a reasonable number of observations at each estimation point u_i over the window $[u_i - b_T,\, u_i + b_T]$.

Finally, we look at the average estimations for each sample size to see if they converge to the true values. We also compute the mean integrated squared error (MISE) between the estimates and the true parameters to measure the quality of the estimation. We

replace the integral over $[0,1]$ in the MISE by a Riemann sum over the equidistant partition $u_1, \ldots, u_{100}$. Here, we generate 500 realisations of the process and investigate $T = 100, 1000, 10000$. The same experiment is then reproduced with different kernels in Appendix A.1.2. Because the results are similar, we focus on those obtained with the Epanechnikov kernel. They are represented in Figure 3.4 and Table 3.1.

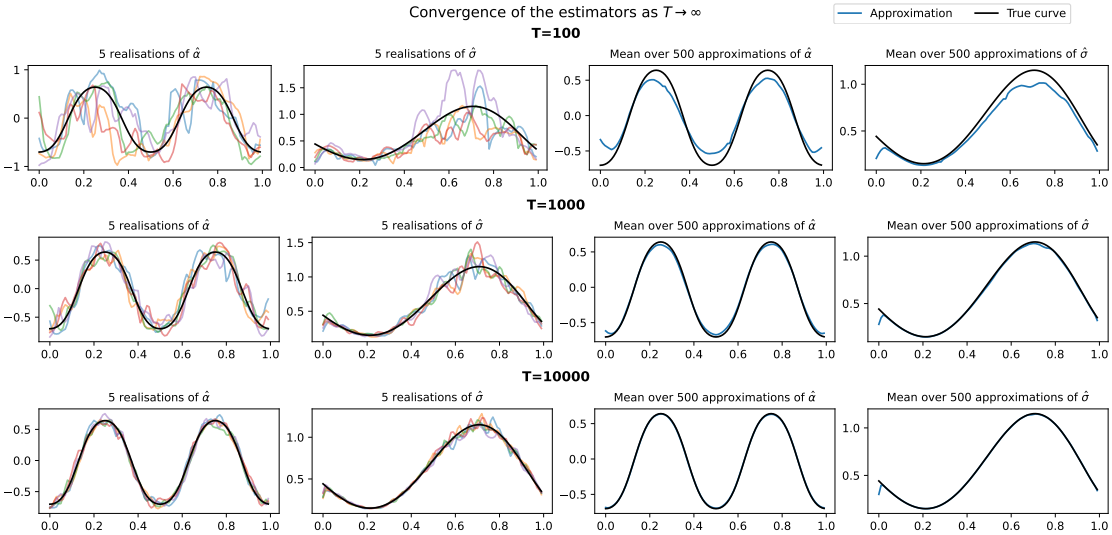

Figure 3.4.: Monte Carlo study for the Yule-Walker estimates with the Epanechnikov kernel and $b_T = 0.1 \times T^{-1/5}$.

Table 3.1.: MISE of $\hat{\alpha}$ and $\hat{\sigma}$ for $T = 100, 1000, 10000$ using the Epanechnikov kernel.

	MISE	
T	$\hat{\alpha}$	$\hat{\sigma}$
100	0.1259	0.0432
1000	0.0197	0.0063
10000	0.0029	0.0011

First of all, we graphically see that the mean estimates become increasively accurate as T increases, reflecting the consistency of the Yule-Walker estimates. This is confirmed by the decreasing MISE. Nonetheless, we observe a high bias near the extreme points of the interval. This comes from the fact that the localizing kernel smoothes observations over the window $[u_i - b_T, u_i + b_T]$ for each estimation point u_i. If u_i is close to either bounds of the interval, the observation window will become unbalanced resulting in a the smoothed average becoming biased. This bias decreases as b_T increases however, which is justified by the observation window shrinking. One way to overcome this edge effect is to estimate the coefficient curves on a subset of $[0,1]$ such as $[b_T, 1 - b_T]$. In the case where we need to estimate the coefficients on the entire interval, some methods

help reducing this bias. For example, (Hall and Wehrly, 1991) describes a geometrical method which involves reflecting the dataset on both edges of the interval, producing a new dataset with three times the extent of the original. This new dataset can then be used to get the kernel estimates on the original design interval without worrying on edge effects. We also implemented this method and notice graphically in Figure A.1 that the bias near the bounds of the interval are improved. Table A.1 also shows that the overall MISE of the estimators is better using this method.

In the above section, we described an intuitive approach to study non-stationary time series using the local stationary approximation. Through a study of a simple tvAR(1) example, we highlighted the importance of the autocovariance sequence in the expression of the parameters estimates and the importance of the segment length on which we assume the process to be stationary. Then, we found that the estimators were consistent by deriving an asymptotic theory on locally stationary processes and investigating their behaviour in a Monte-Carlo simulation study. Now that we have those asymptotic guarantees, we utilise our results in a real world application and model a financial time series using the local stationary approximation in order to develop a simple trading strategy.

4. Application to pairs trading

In this section, we use theory from locally stationary processes to implement a trading strategy on the pair Bitcoin (BTC) / Ethereum (ETH) and backtest it on past data. The trading strategy we are interested in is called 'Pairs trading' (Vidyamurthy, 2004) and involves benefitting from anomalies between two correlated assets (BTC and ETH in this case). The first subsection defines the idea behind this strategy.

4.1. Pairs trading

4.1.1. History

Pairs trading is an investment strategy which was first introduced in the mid 1980s by a group of technical analyst researchers working at Morgan Stanley. At that time, their mission was to develop quantitative arbitrage strategies using state-of-the-art statistical techniques.

The strategy behind pairs trading requires to first identify a pair of assets whose prices tend to move together. Then, if an anomaly in the relationship is detected at some point, a trade is entered with the hope that the anomaly corrects itself. For instance, if an anomaly is detected, the investor would simultaneously sell the overevaluated asset and buy the underevaluated one and later do the opposite to close the position. This is based on the *mean reversion* assumption which is a theory used in finance suggesting that the phenomenon of interest will eventually revert to its long-term average levels.

The advantage of pairs trading is that, although it is hard to determine whether an individual asset is overevaluated or not, it is easier to notice a mispricing by looking at relative pricing, hence the necessity of having two correlated assets. The mutual mispricing is captured by the notion of spread between these assets, referring to the difference between the prices or the rates. The further the spread from its equilibrium, the larger the anomaly and benefit potential.

This strategy was first used on the market in 1987 with great success. It then gradually spread to other trading firms and is still a common strategy used by many hedge funds today, despite being more competitive.

4.1.2. Definitions

To describe more precisely pairs trading, we need to refer to other financial terms that we briefly introduce.

First of all, the application of pairs trading, just like any other strategy, is done with a *portfolio* which is a collection of financial investments (e.g. stocks, bonds, cryptocur-

rencies, cash). The result of the strategy is then reflected by the value of the portfolio, describing how much it is worth at time t. In this application, the portfolio will only hold BTC, ETH and cash.

In some cases, it is possible to develop a portfolio where, without any initial wealth (cash or assets), there is always a profit opportunity which is guaranteed to succeed. In this case, the value of a portfolio keeps increasing over time without ever decreasing. We call this type of portfolio an *arbitrage*. In other words, a portfolio is an arbitrage if it creates wealth 'out of nothing', without any risk of loss.

To apply any strategy, whether it is an arbitrage or not, it is necessary to engage in trades which implies buying or selling assets of the portfolio. We call a *long position* the purchase of an asset with the expectation that its value will go up and a *short position* the symmetric situation.

Furthermore, a desired outcome for many strategies is that its returns are uncorrelated with the market returns, meaning it performs in a steady manner whether the marker goes up or down. If this is the case, the strategy is said to be *market neutral*.

Pairs trading belongs to a specific category of market neutral trading strategies called statistical arbitrage, where only two assets form the portfolio. The strategy involves opening both a long position and short position simultaneously to take advantage of inefficient pricing in correlated assets. Unlike an arbitrage, it is not without any risk since it heavily depends on statistical analysis and the mean reversion property of market prices, meaning that the prices will eventually go back to an equilibrium. Because of that risk factor, statistical arbitrage strategies often require high-frequency trading algorithms to take advantage of tiny anomalies that often last for a short time period.

As described in (Vidyamurthy, 2004), the main way to perform the statistical analysis is by using a cointegration technique to make the spread stationary which is desirable from the forecasting perspective, but other methods exist (Do et al., 2006). In this work, we attempt to model and forecast the spread with locally stationary processes allowing us to incorporate behaviour beyond stationarity in our model.

In order to then quantitatively determine whether pairs trading is an effective strategy, it is important to backtest it, i.e. test it on historical data. One way to monitor its performances is to look at its profit and loss (P&L) referring to its net gains or losses.

Example 4.1.1. If a portfolio has the following composition:

- -1 BTC (short position)
- 10 ETH (long position)
- \$2000 in cash

and the current values for BTC (resp. ETH) are \$21000 (resp. \$1200), then the current P&L will be: $P\&L = -1 \times 21000 + 10 \times 1600 + 2000 = -\$3000$ which is a loss.

4.1.3. Assumptions

In this work, we make the following simplifying assumptions:

- BTC and ETH can be traded with immediate effect, meaning that we can open or close a position as soon as a signal is detected.
- The market is frictionless. This means that buying or selling shares does not involve any transaction costs or taxes, and that the current market price per share applies to both buying and selling and is the same for any traded amount (no bid-ask spread).
- It is possible to hold negative values of an asset, i.e. selling some BTC or ETH without actually holding them (short selling).

The goal of these assumptions is to focus on the direct results of the strategy, without taking into account the costs incurred by other phenomena. We will attempt to implement a pairs trading algorithm on the pair BTC / ETH, both evaluated in US dollars.

4.2. Bitcoin (BTC) / Ethereum (ETH)

In this subsection, we offer some justification on the choice of the pair BTC / ETH and derive the relevant expression of the spread. Indeed, the main factor of success in pairs trading lies in the pair of assets we want to trade since the strategy relies on a strong correlation between the two which eventually results in a mean reverting effect.

To verify this graphically, we plot the evolution of the prices of BTC and ETH between 17/08/2017 and 14/08/2022 Figure 4.1. The data was downloaded online (CryptoDataDownload, 2022) and gives hourly prices in US dollars of the two cryptocurrencies from the Binance Exchange.
We call p_t^{BTC} (resp. p_t^{ETH}) the price of Bitcoin (resp. Ethereum) at time t.

Figure 4.1.: Hourly prices of BTC and ETH in US dollars.

As we can see, the prices of both BTC and ETH tend to move together which is needed to form a pair. This is verified by the value of Pearson's correlation which is 0.93 on the sample, with a p-value smaller than 1%, meaning that there is a significant linear relationship between p_t^{BTC} and p_t^{ETH} at a 1% level.

Remark 4.2.1. Although this could be a 'spurious correlation', where the relationship between the two assets are coincidental or due to an unseen confounder, there are reasons to believe that it is not the case here, given how closely related the two cryptocurrencies are.

However, they are on two very different scales since Bitcoin's price is more than 10 times larger than Ethereum's price. For that reason, the direct spread $s_t^{\text{direct}} = p_t^{\text{BTC}} - p_t^{\text{ETH}}$ would just be similar to Bitcoin's price making it hard to predict.

To overcome this problem, we first notice that if BTC and ETH have the same movements but at different levels, they must share similar log-returns defined as follow:

$$r_t^{\text{BTC}} = \log\left(\frac{p_t^{\text{BTC}}}{p_{t-1}^{\text{BTC}}}\right) \qquad \text{and} \qquad r_t^{\text{ETH}} = \log\left(\frac{p_t^{\text{ETH}}}{p_{t-1}^{\text{ETH}}}\right).$$

Reasoning in hourly returns instead of hourly prices is a form of normalisation that still captures the movements of the time series. We indeed find that Pearson's correlation of the log-returns time series is larger than 0.8 at a 1% level, which means that the pair can reasonably be considered for a pairs trading strategy. Therefore, we use the following

expression of the spread:

$$s_t = r_t^{\text{BTC}} - r_t^{\text{ETH}}. \tag{4.1}$$

Remark 4.2.2. At time t, the current situation of the market can be analysed through the sign of the spread. A positive spread means that the log-return of BTC is larger than the log-returns of ETH, so that the value of BTC increases more (or decreases less) than the value of ETH, resulting in a potential mispricing. A negative spread, on the other hand, means the opposite situation.

As we can see from Figure 4.2, the time series of the spread is clearly non-stationary as the variance changes over time.

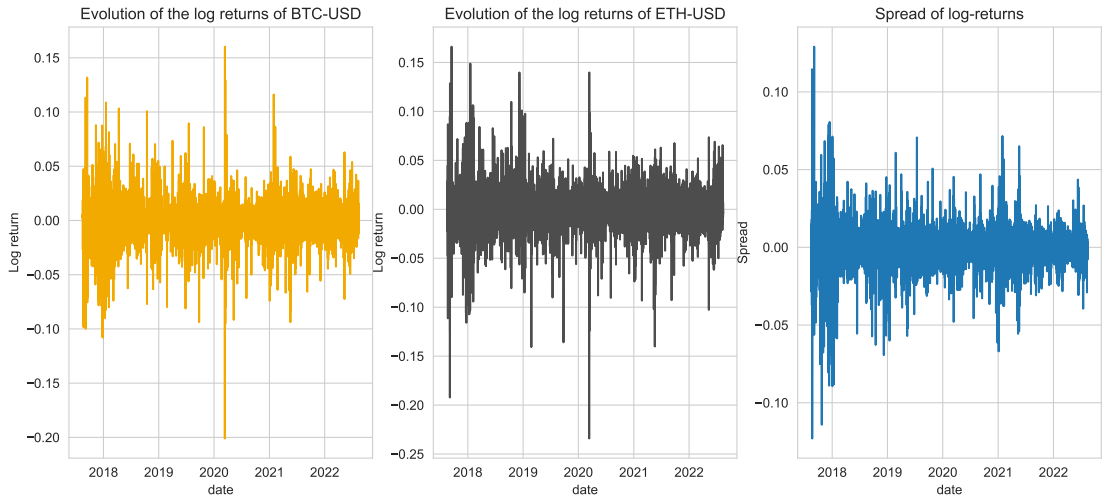

Figure 4.2.: Hourly log-returns of BTC and ETH and spread of the log-returns.

Remark 4.2.3. In finance, log-returns are often preferred over simple returns because they are unbounded and additive in time. Besides, returns are sometimes assumed to be log-normally distributed so the log-returns are conviniently normally distributed.

Now that we have defined the spread, we can start doing some analysis to determine when to enter a trade. To do so, we use concepts from locally stationary processes.

4.3. Locally stationary approximation of the spread

In this subsection, we model the spread time series as a tvAR(1) process defined in Equation 3.4, which is locally mean-reverting by nature. Using this model and the assumption that the parameter curve $u \mapsto \alpha(u)$ is smooth, we employ interpolation techniques to forecast the spread at time $t + 1$. Then, depending on the value of the forecast, we develop different strategies for when to enter a trade.

4.3.1. Modelling the spread as a tvAR(1) process

Let us assume that the spread time series is a tvAR(1) process. We recall that it satisfies the expression:

$$s_{t,T} + \alpha\left(\frac{t}{T}\right) s_{t-1,T} = \sigma\left(\frac{t}{T}\right)\epsilon_t, \qquad t \in \{1,\dots,T\}$$

with $\alpha : [0,1] \to (-1,1)$, $\sigma : [0,1] \to \mathbb{R}_+$ both Lipschitz continuous, ϵ_t i.i.d. $\mathcal{N}(0,1)$ and $T \in \mathbb{N}$. The reason this model is particularly suitable to model the spread time series is because it is locally mean-reverting. Hence, it makes sense to assume that any mispricing anomaly, caused by a deviation of the spread from its equilibrium, will eventually correct itself.

To make explanations clearer, we take for example the spread time series where the observations are from 23/06/2021 08:00:00 to 14/08/2022 00:00:00 (represented by the interval of interest $[0,1]$). In this case, $T = 10000$.

First, we start by estimating the parameter curves using the Yule-Walker method from Subsection 3.2.1, with the Epanechnikov kernel and $b_T = 0.1\,T^{-1/5} = 0.0158$. The spread as well as the estimated parameter curves $\hat{\alpha}$ and $\hat{\sigma}$ are shown Figure 4.3.

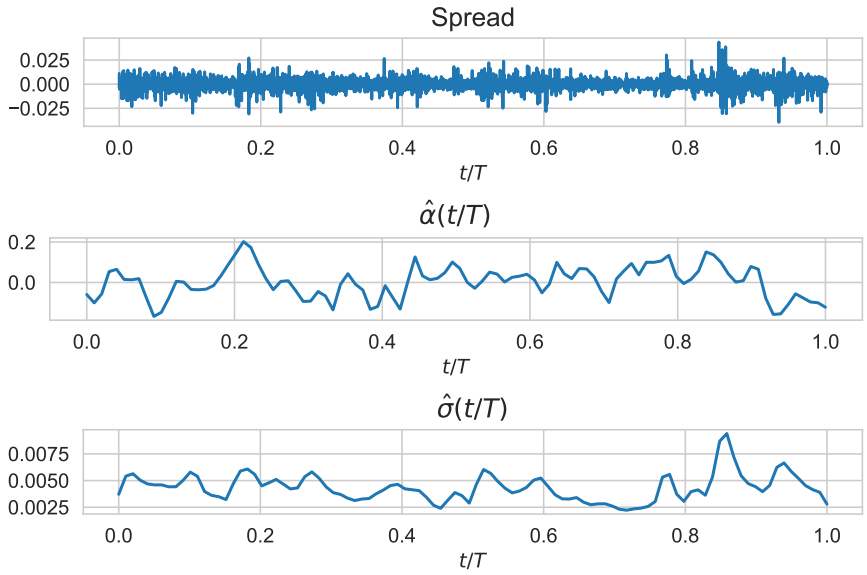

Figure 4.3.: Yule-Walker estimation of the parameter curves for the spread time series between 23/06/2021 08:00:00 and 14/08/2022 00:00:00.

We notice two things from the above figure. First, the sign of $\hat{\alpha}$ regularily changes, meaning that the one-step correlation varies from being positive to negative. We also see that $\hat{\sigma}$ has small values and captures noticeable changes in the variance of the spread. Indeed, $\hat{\sigma}(0.85)$ is four times as high as $\hat{\sigma}(0.7)$ for example.

4.3.2. Forecasting

The idea is then to use these parameter curves to predict the future value of the spread. To do so, we aim to extrapolate $u \mapsto \hat{\alpha}(u)$ and $u \mapsto \hat{\sigma}(u)$ using cubic splines, a form of smooth polynomial interpolation, to then forecast the spread as:

$$s_{T+1} = -\hat{\alpha}\left(\frac{T+1}{T}\right) s_{T,T}. \tag{4.2}$$

This method of forecasting is justified by the locally stationary approximation where the stationarity of the process varies slowly enough on a small interval and the parameter curves $u \mapsto \alpha(u)$ and $u \mapsto \sigma(u)$ also vary slowly and smoothly. To continue with our example, we briefly explain how the extrapolation part is done.

Cubic spline interpolation

Cubic spline interpolation, described in (McKinley and Levine, 1998), is the mathematical equivalent of drawing a smooth curve between a number of points. We only consider the last 10 points of the time series we want to extrapolate in order to build that curve, namely $\hat{\alpha}(\frac{T-9}{T}), \ldots, \hat{\alpha}(1)$. The general idea of cubic spline is to fit a piecewise function of the form:

$$P_{\hat{\alpha}}(u) = \begin{cases} p_1(u) & \text{if } \frac{T-9}{T} \leq u \leq \frac{T-8}{T} \\ \vdots & \vdots \\ p_9(u) & \text{if } \frac{T-1}{T} \leq u \leq 1 \end{cases}$$

where, for $i = 1, \ldots, 9$, p_i is a third degree polynomial. The function $P_{\hat{\alpha}}$ needs to satisfy four conditions:

1. $P_{\hat{\alpha}}$ interpolates every point $\hat{\alpha}(\frac{T-9}{T}), \ldots, \hat{\alpha}(1)$.
2. $u \mapsto P_{\hat{\alpha}}(u)$ is continuous on $[\frac{T-9}{T}, 1]$.
3. $u \mapsto P'_{\hat{\alpha}}(u)$ is continuous on $[\frac{T-9}{T}, 1]$.
4. $u \mapsto P''_{\hat{\alpha}}(u)$ is continuous on $[\frac{T-9}{T}, 1]$.

These four conditions provide a system of equations that can be numerically solved to find the expressions of the polynomials. The last polynomial is then used to extrapolate the next value of the time series as such:

$$\hat{\alpha}\left(\frac{T+1}{T}\right) = p_9\left(\frac{T+1}{T}\right). \tag{4.3}$$

This then allows us to forecast the spread according to Expression 4.2. Figure 4.4 illustrates the extrapolation of $\hat{\alpha}$ with the example data and Figure 4.5 shows the associated forecast of the spread.

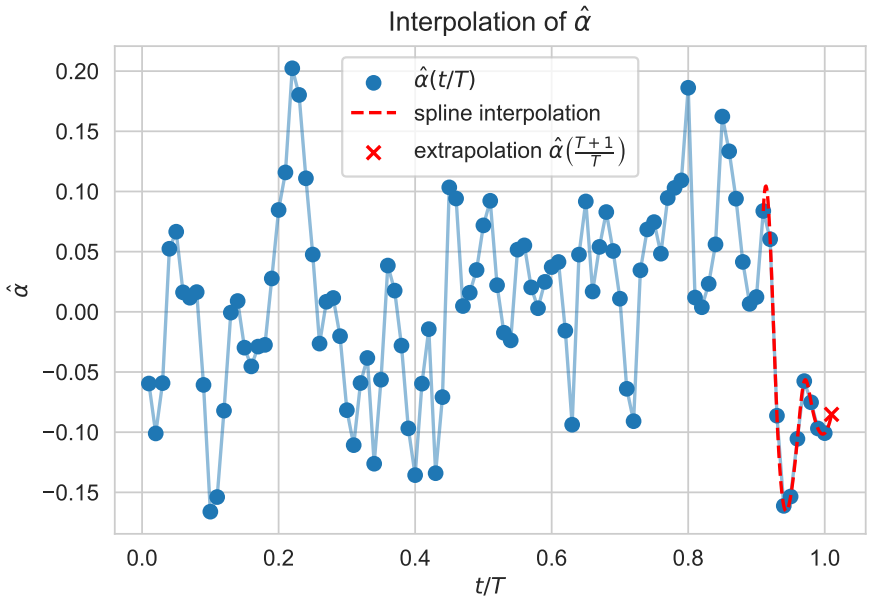

Figure 4.4.: Cubic spline interpolation and extrapolation of $u \mapsto \hat{\alpha}(u)$.

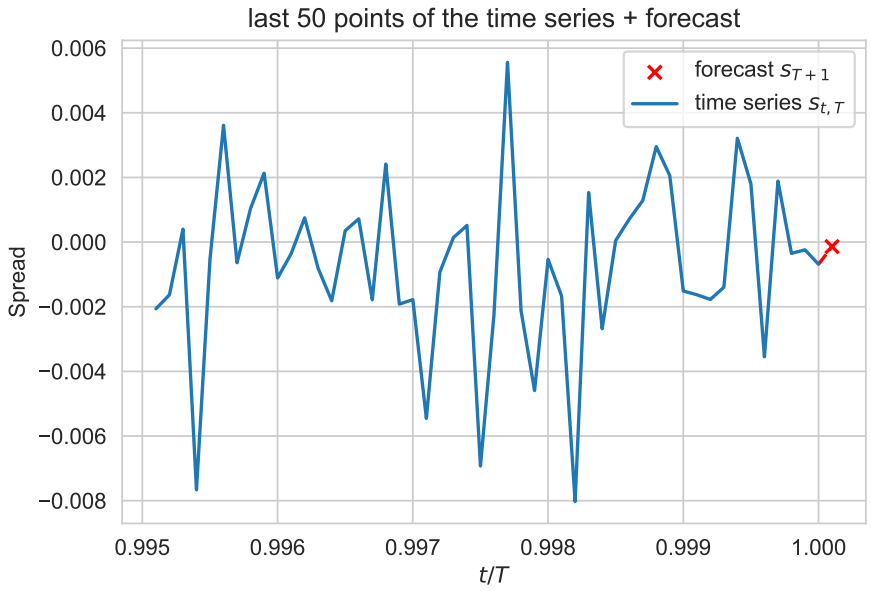

Figure 4.5.: Forecast of the spread time series using Equation 4.2.

4.3.3. Strategies

Now that we are able to obtain a one-step prediction of the spread, we need to compare this value to the equilibrium point. If the prediction deviates a lot from it, then we enter a trade. Otherwise, we do not enter any trade. To account for the variance of the spread, we measure the distance between the forecast and the equilibrium point in terms of standard deviations using the z-score.

Definition 4.3.1 (Z-score)**.** Let s_{t+1}^{pred} be the prediction of the spread at time $t+1$. The associated z-score is given by:

$$z_t = \frac{s_{t+1}^{\text{pred}} - \hat{\mu}^{(\ell)}}{\hat{\sigma}^{(\ell)}} \tag{4.4}$$

where $\hat{\mu}^{(\ell)}$ and $\hat{\sigma}^{(\ell)}$ are the local estimates of respectively the mean and the standard deviation of the spread time series near $t = T$. They are computed at the end of the interval using a localizing kernel K, a bandwidth b_T and the reflected data $X_{t,T}$ from the edge effects reduction method (Hall and Wehrly, 1991), according to the equations:

$$\hat{\mu}^{(\ell)} = \sum_{t=1}^{T} K\left(\frac{2/3 - t/T}{b_T}\right) X_{t,T} \tag{4.5}$$

$$\hat{\sigma}^{(\ell)} = \sqrt{\sum_{t=1}^{T} K\left(\frac{2/3 - t/T}{b_T}\right) X_{t,T}^2}. \tag{4.6}$$

Remark 4.3.2. Taking $u_0 = 2/3$ with the reflected time series in Equations 4.5 and 4.6 is equivalent to taking $u_0 = 1$ with the initial time series. It thus represents the local mean and local standard deviation at the end of the time series.

The z-score gives us insight on whether to enter a trade or not. Let $z_* \in \mathbb{R}_+$ be a threshold value. We distinguish three situations:

1. $z_t \geq z_*$: we predict that the spread becomes significantly larger than its equilibrium point, meaning that BTC may be overpriced or ETH may be underpriced. In this situation, we short BTC and long ETH.
2. $z_t \leq -z_*$: we predict that the spread becomes significantly smaller than its equilibrium point, meaning that BTC may be underpriced or ETH may be overpriced. In this situation, we long BTC and short ETH.
3. $|z_t| < z_*$: we do not predict a movement of the spread that is significant enough to engage in a trade.

The threshold value z_* clearly impacts the strategy and needs to be tuned in order to maximize profits. It can be seen as a risk managing parameter. The larger its value, the fewer trades will be made. The trade that will actually be made correspond to those where there is a clear signal that the spread is diverging. Using this logic to enter a trade, we investigate three simple strategies.

Strategy 1

A trade is entered at time t if $|z_t| \geq z_*$. It is then immediately closed at time $t+1$. This arises from the fact that the forecast of the spread is only valid *locally*. Indeed, the spread being non-stationary, we don't have any information about its future.

Strategy 2

A trade is entered at time t if $|z_t| \geq z_*$ and all previous positions were closed (the portfolio is limited to one opened trade). To close a position, we are waiting to get the opposite signal (e.g. we are waiting for a signal to long BTC if we previously shorted BTC) or we are waiting for the end of the trading interval to be reached.

Strategy 3

A trade is entered at time t if $|z_t| \geq z_*$ and all previous positions were closed. A position is closed if the spread starts diverging in the opposite direction or if the end of the trading interval is reached. For example, if we have a long position on BTC, we will close the position when $z_t > z'_*$ where z'_* is another threshold value smaller than z_*. In all simulations, $z'_* = 0.75$ will be used.

Intuitively, we expect strategy 1 to make smaller gains or losses at each trade. Indeed, since opened positions are immediately closed, the spread is not allowed to diverge more, reducing the benefit / loss potential at each step in time.

In strategy 2, since closing a position is equivalent to detecting a signal to open a new one, it implies to always have an opened position after the first trade. Although the strategy is market neutral, it can be risky to hold onto a position for too long.

Strategy 3 seems to be a better version of the two. Indeed, first it waits for the spread to diverge in the opposite direction before closing a position. That way, the profit potential keeps increasing as long as the spread keeps diverging in the same direction. Then, since it requires a smaller threshold to close a position than to open a new one, a failing position will be closed quicker than with strategy 2. This also allows the portfolio to not always have an opened position.

4.4. Implementation and results

In this subsection, we implement, compare and evaluate the three strategies by backtesting them on the interval 17/01/2022 17:00:00 - 14/08/2022 00:00:00 ($n = 5000$ hours). The study was conducted in Python on the NextGen Maths Clusters. The code is available on Github. For each strategy, the method is the same and is described in Algorithm 1.

Algorithm 1 Pseudocode of the backtesting method for each strategy.

$t_{\text{start}} \leftarrow$ 17/01/2022 17:00:00
$t_{\text{end}} \leftarrow$ 14/08/2022 00:00:00
$T \leftarrow 10000$
for $t = t_{\text{start}}, \dots, t_{\text{end}}$ (5000 hours) **do**

1. time series $\leftarrow (s_{t-T+1}, \dots, s_t)$ # T observations.
2. fit a tvAR(1) model, i.e. compute the Yule-Walker estimates of $u \mapsto \alpha(u)$ and $u \mapsto \sigma(u)$ at 100 equidistant points, using the Epanechnikov kernel and the bandwidth $b_T = 0.1\, T^{-1/5}$.
3. extrapolate the coefficient curves, predict s_{t+1} and compute the z-score.
4. check if a trade should be entered.

end for

Before backtesting our strategies, we start by tuning the threshold value z_* using the hit-ratio metric, which looks at the trades that were entered.

Definition 4.4.1 (Hit-ratio)**.** In a trading interval, the hit-ratio h computes the ratio of successful predictions that led to enter a trade. Let n be the length of the trading interval, s_t be the spread process and s_t^{pred} represent the associated 1-step forecasts. Let $I \subset \{1, \dots, n\}$ be the set of times where a position was entered. Then, h is defined by:

$$h = \frac{1}{|I|} \sum_{i \in I} \delta_{\text{sign}(s_i),\, \text{sign}(s_i^{\text{pred}})}. \tag{4.7}$$

Remark 4.4.2. For $i \in I$, if $\text{sign}(s_i) = \text{sign}(s_i^{\text{pred}})$, then the right movement of the spread has been predicted and the trade entry is considered as successful.

This definition gives the general version of the hit ratio which takes every trade into account but we can filter on the type of trade to see if a strategy performs better when it comes to short and long positions.

This metric evaluates the entry of a trade but, depending on the strategy employed, a trade which is failing at first can end up become profitable.

To find the values of z_* that historically work the best for each strategy, we compute their hit ratio for 15 values of z_* between 1 and 2.5 from 13/01/2019 06:00:00 to 17/01/2022 16:00:00 (26410 hours). The selected values are given Table 4.1.

Table 4.1.: Values of z_* corresponding to the best hit ratios on the interval 13/01/2019 06:00:00 - 17/01/2022 16:00:00.

	Strategy 1	Strategy 2	Strategy 3
z_*	1.3214	1.4285	1.214286
hit ratio	0.6458	0.6363	0.6316

As we can see, the associated hit ratios are larger than 0.5 which is encouraging since it means that our predictions are more often correct when entering a trade than not.

Using these values, we backtest the three strategies with no initial position in the portfolio. In order to take into account the different levels of prices and remain market neutral, we spend \$20000 worth of BTC and ETH at each trade, which means we don't trade the same volume. The evolution of their P&L and their hit ratios are given Figure 4.6 and Table 4.2.

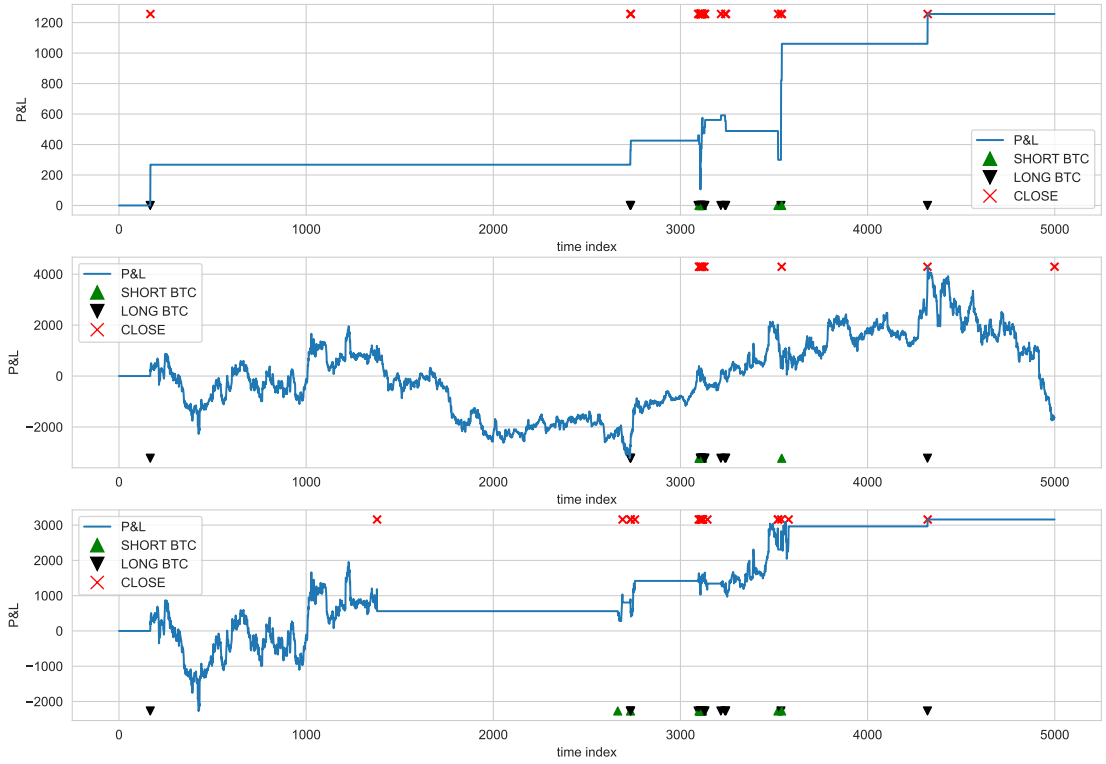

Figure 4.6.: Evolution of the P&L during backtests of the three strategies over 5000 hours. Note: as two positions are always opened simultaneously, the figure only shows the BTC trades.

Table 4.2.: Hit ratios of the three strategies per type of trade.

	Strategy 1	Strategy 2	Strategy 3
hit ratio (general)	0.3333	0.4444	0.4211
hit ratio LONG	0.1875	0.2	0.1818
hit ratio SHORT	0.625	0.75	0.75

First of all, if we just look at the hit ratios, we notice that all three strategies achieve poor results. The scores being smaller than 0.5, more trades were considered as failed than successful. We notice that these bad overall scores are the results of mostly failed long trades for BTC: for strategy 2 for example, when a long trade signal was detected, the price then immediately dropped 80% of the time. These long signals are most likely failing because the prices of the cryptocurrencies have been dropping during the trading period as can be seen in Figure 4.1. As a result, only few of the long positions have been successful while most of the short positions turned out successful. This market trend leading to poor long trades is also probably the reason why the overall hit ratios are so low compared to the tuning interval (Table 4.1) where the prices were more consistent.

However, this metric does not give a fair evaluation of the performances of the strategies on the entire interval since it only looks at the entry of the trades but not the exit. Indeed, what ultimately matters is to see if the trades made profits, i.e. it is more interesting to look at the evolution of the P&L as an evaluation metric.

In Figure 4.6, we see that despite the poor hit ratios, two out of the three strategies made profits.

Strategy 1 makes smaller gains than strategy 3 which was expected as the magnitude of the P&L is smaller since positions are immediately closed.

Strategy 2 seems to keep opened positions for too long, which can lead to significant losses (for example the first trade). If a position is first opened and the model is uncertain, then the position will be held long enough for the differences of levels between BTC and ETH to reflect on the P&L. For example, if a portfolio has a long position on BTC and a short position on ETH in a period where the market goes down, the differences of levels implies that the price of BTC will drop more in absolute value than the price of ETH. Although this effect is reduced by the fact that we are not trading an equal volume of both cryptocurrencies, it cannot be removed on the long term. This means that strategy 2 loses in market neutrality, which is a required property.

Strategy 3 seems to perform steadily despite losing at some point because of the first trade. This first trade was kept for a little too long because it was an uncertain period for our model, which is reflected by small values of the z-score. The time series of the z-score is given in appendix Figure A.6. Otherwise, most of the other trades seem profitable and quickly closed.

Therefore, out of the three strategies, we would favour the third one as the most profitable while being seemingly reliable.

All strategies's success heavily rely on the 1-step forecast which depends on the chosen model. In this application, we chose a simple tvAR(1) model which is easy to estimate but limited in complexity. Hence, other models could have been used such as tvAR(p), tvARCH (Dahlhaus and Rao, 2006) or tvGARCH (Rohan and Ramanathan, 2013). To keep simulations simple, we chose to adopt a limit to the number of opened positions but theory on optimal portfolios with locally stationary returns of assets could also have been used to develop a more efficient strategy (Shiraishi and Taniguchi, 2007).

5. Conclusion

In this work, we have introduced a recently developed theory aiming to overcome the restriction of stationarity in time series analysis. Using the locally stationary approximation, we studied methods to fit a model on any time series, focussing on the tvAR(1) model and the Yule-Walker estimates. The consistency of these estimates then gave us a statistical guarantee to use them in practice.

We applied this theory in a pairs trading application to model and predict the hourly spread between Bitcoin and Ethereum with a tvAR(1) process. From this, we then developed three simple trading strategies that showed promising results. In a time period where the market does not seem to follow a trend, the predictions were more often correct than not. On the tested trading period where the prices of cryptocurrencies were dropping however, the predictions were less accurate, although the strategies mostly performed well. In order to determine whether these strategies relying on a tvAR(1) model actually outperform the existing stationary methods, the next step would be to implement and backtest a stationary method, like cointegration, and compare the P&L on the same interval.

Despite not having a direct comparison of the application, we can still observe some general advantages and drawbacks from using locally stationary processes instead of stationary processes. Indeed the main advantage is that a model can be fitted on any time series, without having to transform it, providing insight that was previously impossible to obtain. Besides, since most time-varying models are modifications of stationary models, the scope of possible applications is very wide. However, we saw that this theory was built in the framework of infill asymptotics and thus requires a lot of data to obtain accurate estimates. Simple stationary models, on the other hand, need very few data points to work. Moreover, from a forecasting perspective, stationarity provides more guarantees and a more comprehensive theory than local stationarity.

Theory based on the locally stationary approximation recently showed a promising breakthrough in time series analysis and is an active research area as it keeps getting updated to offer new methods competing with traditional tools. For example, a general theory on stationary approximations for non-stationary continuous-time processes is currently being developed to establish inference methodologies in a more general setting (Stelzer and Ströh, 2021; Ströh, 2021).

References

Peter J. Brockwell and Richard A. Davis. *Time series: theory and methods*. Springer science & business media, 2009.

Ngai Hang Chan. *Time series: applications to finance*. John Wiley & Sons, 2004.

Huann-Sheng Chen, Douglas G. Simpson, and Zhiliang Ying. Infill asymptotics for a stochastic process model with measurement error. *Statistica Sinica*, pages 141–156, 2000.

CryptoDataDownload. Cryptodatadownload, 2022. URL `https://www.cryptodatadownload.com/data/`. [Online; data downloaded 14-August-2022].

Rainer Dahlhaus. Asymptotic statistical inference for nonstationary processes with evolutionary spectra. In *Athens conference on applied probability and time series analysis*, pages 145–159. Springer, 1996.

Rainer Dahlhaus. Locally stationary processes. In *Handbook of statistics*, volume 30, pages 351–413. Elsevier, 2012.

Rainer Dahlhaus and Liudas Giraitis. On the optimal segment length for parameter estimates for locally stationary time series. *Journal of Time Series Analysis*, 19(6): 629–655, 1998.

Rainer Dahlhaus and Suhasini Subba Rao. Statistical inference for time-varying ARCH processes. *The Annals of Statistics*, 34(3):1075–1114, 2006.

Rainer Dahlhaus, Stefan Richter, and Wei Biao Wu. Towards a general theory for nonlinear locally stationary processes. *Bernoulli*, 25(2):1013–1044, 2019.

Binh Do, Robert Faff, and Kais Hamza. A new approach to modeling and estimation for pairs trading. In *Proceedings of 2006 financial management association European conference*, volume 1, pages 87–99, 2006.

Peter Hall and Thomas E. Wehrly. A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *Journal of the American Statistical Association*, 86(415):665–672, 1991.

Keith W. Hipel and A. Ian McLeod. *Time series modelling of water resources and environmental systems*. Elsevier, 1994.

Ulrich Krengel. *Ergodic theorems*, volume 6. Walter de Gruyter, 2011.

Sky McKinley and Megan Levine. Cubic spline interpolation. *College of the Redwoods*, 45(1):1049–1060, 1998.

Stefan Richter and Rainer Dahlhaus. Cross validation for locally stationary processes. *The Annals of Statistics*, 47(4):2145–2173, 2019.

Neelabh Rohan and TV Ramanathan. Nonparametric estimation of a time-varying GARCH model. *Journal of Nonparametric Statistics*, 25(1):33–52, 2013.

Hiroshi Shiraishi and Masanobu Taniguchi. Statistical estimation of optimal portfolios for locally stationary returns of assets. *International Journal of Theoretical and Applied Finance*, 10(01):129–154, 2007.

Robert Stelzer and Bennet Ströh. Asymptotics of time-varying processes in continuous-time using locally stationary approximations. *arXiv e-prints*, pages arXiv–2105, 2021.

Bennet Ströh. Statistical inference for continuous-time locally stationary processes using stationary approximations. *arXiv preprint arXiv:2105.04390*, 2021.

Ruey S. Tsay and George C. Tiao. Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARMA models. *Journal of the American Statistical Association*, 79(385):84–96, 1984.

Ganapathy Vidyamurthy. *Pairs Trading: quantitative methods and analysis*, volume 217. John Wiley & Sons, 2004.

Michael Vogt. Nonparametric regression for locally stationary time series. *The Annals of Statistics*, 40(5):2601–2633, 2012.

A. Appendix

A.1. Monte-Carlo simulation

A.1.1. Reduction of the edge effect

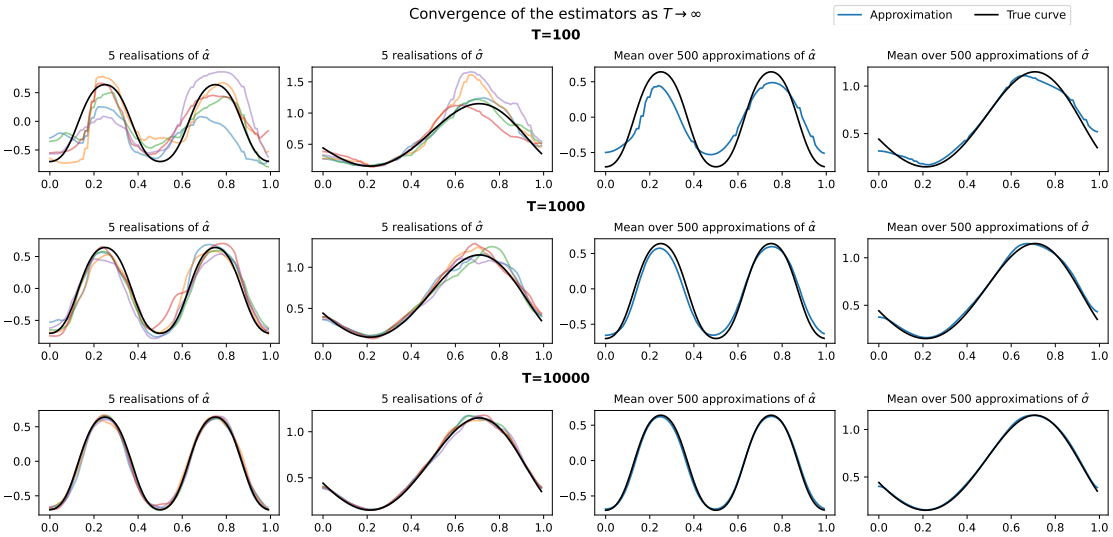

Figure A.1.: Monte Carlo study for the Yule-Walker estimates with the Epanechnikov kernel and $b_T = 0.1 \times T^{-1/5}$, with the edge effects reduction method of (Hall and Wehrly, 1991).

Table A.1.: MISE of $\hat{\alpha}$ and $\hat{\sigma}$ for $T = 100, 1000, 10000$ using the Epanechnikov kernel with the edge effects reduction method of (Hall and Wehrly, 1991).

	MISE	
T	$\hat{\alpha}$	$\hat{\sigma}$
100	0.0923	0.0176
1000	0.0137	0.0029
10000	0.0021	0.0005

A.1.2. Monte-Carlo simulation with other kernels

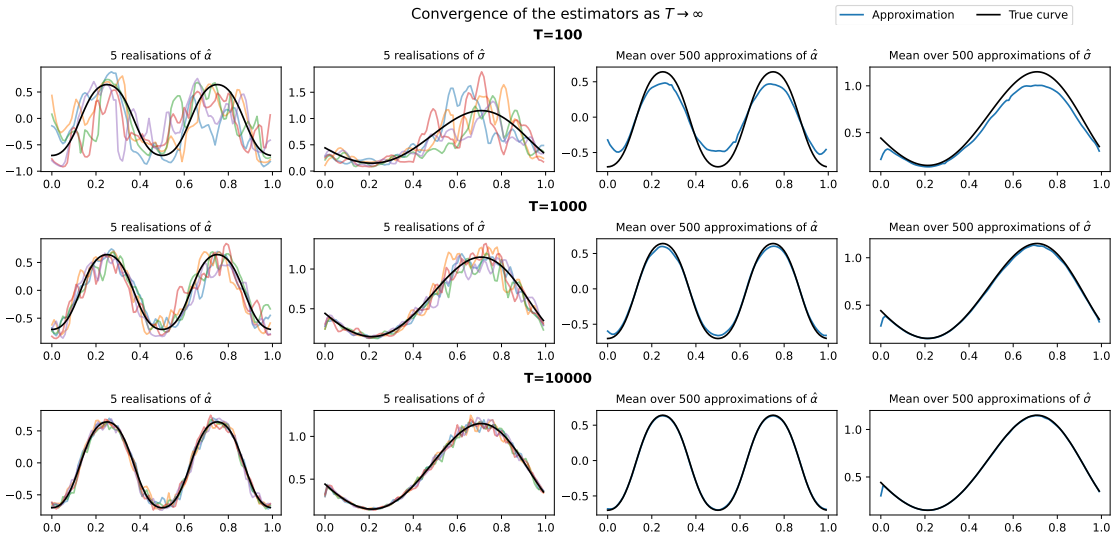

Figure A.2.: Monte Carlo study for the Yule-Walker estimates with the quadratic kernel and $b_T = 0.1 \times T^{-1/5}$. No reduction of edge effects.

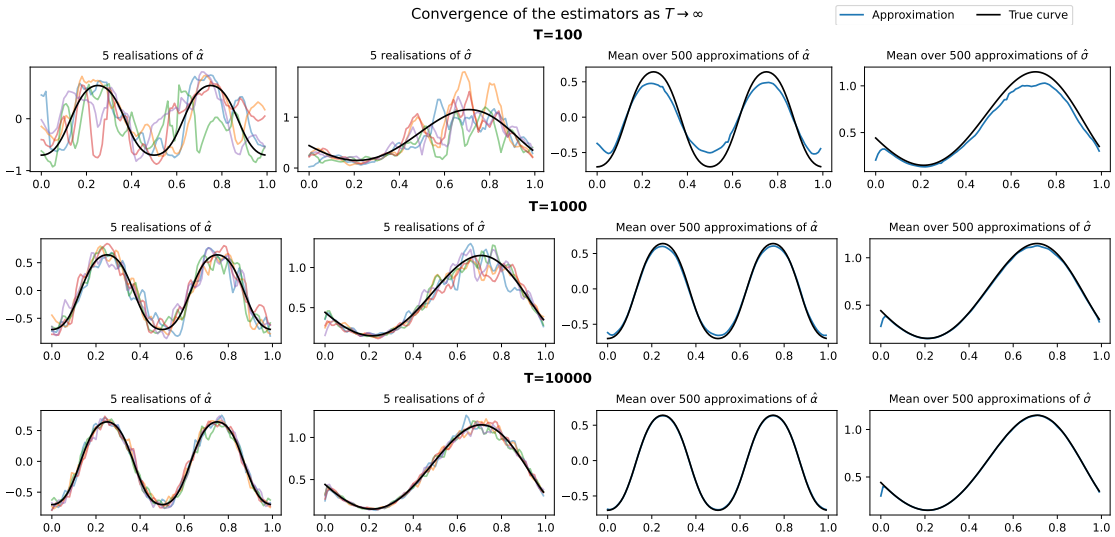

Figure A.3.: Monte Carlo study for the Yule-Walker estimates with the triangular kernel and $b_T = 0.1 \times T^{-1/5}$. No reduction of edge effects.

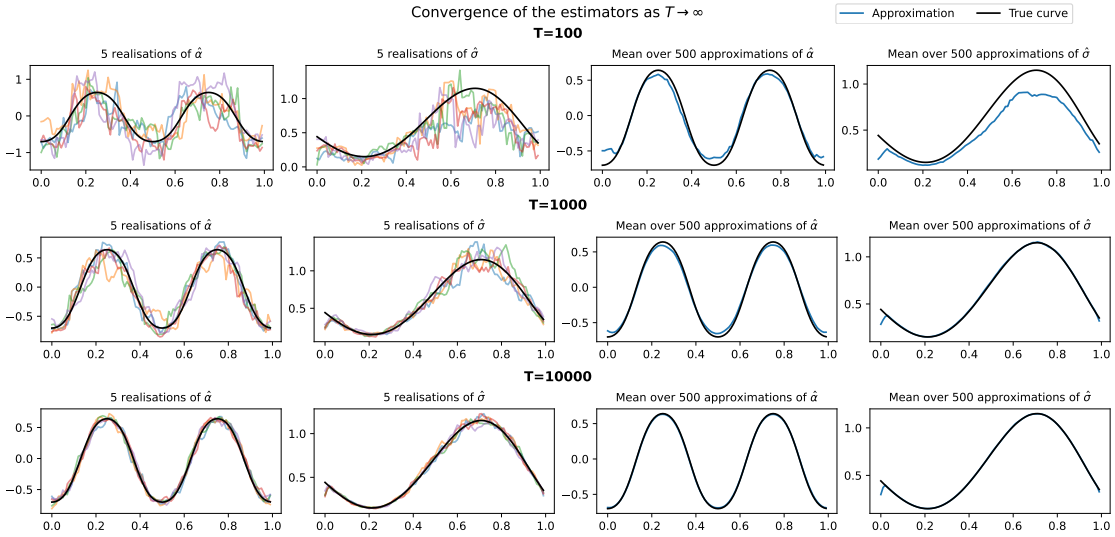

Figure A.4.: Monte Carlo study for the Yule-Walker estimates with the uniform kernel and $b_T = 0.1 \times T^{-1/5}$. No reduction of edge effects.

Table A.2.: Mean integrated squared error for every kernel.

	MISE							
Kernel	Epanechnikov		Quadratic		Triangular		Uniform	
T	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\alpha}$	$\hat{\sigma}$	$\hat{\alpha}$	$\hat{\sigma}$
100	0.1259	0.0432	0.1384	0.0454	0.1303	0.0453	0.1607	0.0575
1000	0.0197	0.0063	0.0227	0.0072	0.0210	0.0069	0.0175	0.0055
10000	0.0029	0.0011	0.0035	0.0013	0.0033	0.0013	0.0025	0.0010

As we can see from the different figures, the estimators $\hat{\alpha}$ and $\hat{\sigma}$ behave very similarily despite being computed using different kernels. Indeed, Table A.2 shows that the MISE are almost identical for the different kernels. Here, the uniform kernel seems to perform slightly better, although the difference is minor.

A.2. Pairs trading application

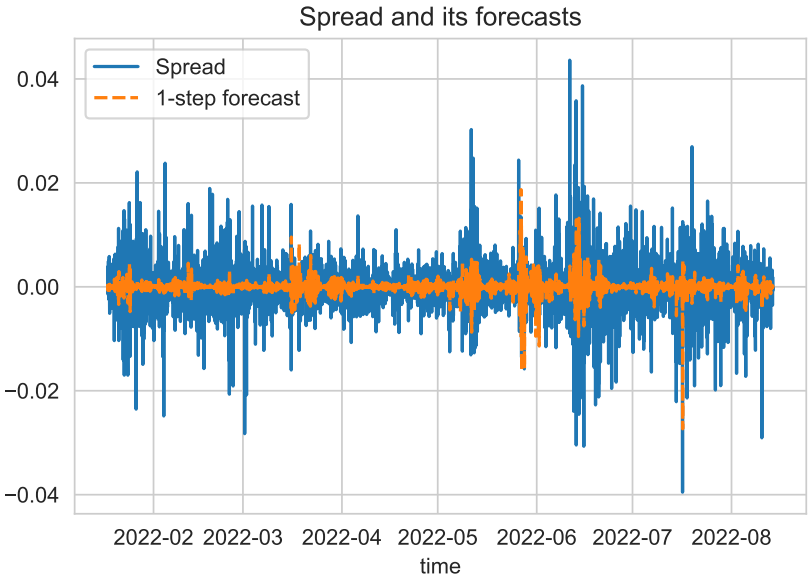

Figure A.5.: Spread over the trading interval with its 1-step associated forecasts.

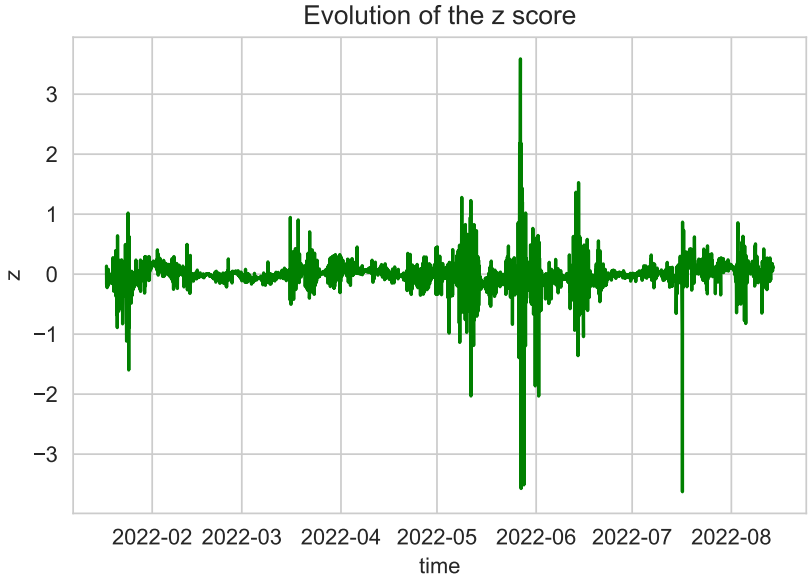

Figure A.6.: Evolution of the z score over the trading interval.