Semiparametric estimation for time series: a frequency domain approach based on optimal transportation theory

Manon Felix
Advisor: Prof. Davide La Vecchia

May 2021

Abstract

In this master thesis, we make use of some results available in the optimal transportation literature to develop a novel methodology for parameters estimation in time series models based. The key idea is to use the Wasserstein distance and Sinkhorn divergence to derive minimum distance (or divergence) estimators for short- and long-memory time series models. Working on the Fourier transform of a time series we conduct inference in the frequency domain: we compute the distance/divergence between the empirical distribution of the standardized periodogram ordinates and their theoretical distribution, as implied by standard asymptotic theory. To study the properties of these new estimators, we perform several Monte-Carlo simulations. Our numerical results suggest that the estimators belonging to our novel class are root-n consistent. Their performance, in terms of Mean Squared-Error, is similar to the one yielded by the state-ofthe-art estimation method (Whittle's estimator) in the case of short- and long-memory Gaussian process. Furthermore, when the underlying innovation density of a long-memory process is skewed, our estimators overperform the Whittle's estimator.

Contents

1 Introduction

1.1 Motivation

The aim of this master thesis is to combine some results from optimal transport theory with the statistical analysis time series analysis. Working with the frequency domain approach, we aim at developing a novel methodology for estimating the parameters of a univariate stochastic process.

We do not rely on the standard information divergence-based methods, among which the standard maximum likelihood estimator approach, and consider instead the mathematical theory of optimal measure transportation.

The optimal transport theory has been applied in many research areas, like e.g. mathematics, differential geometry, partial differential equations and applied mathematics (see e.g. Santambrogio (2015) and Villani (2009)). The Wasserstein distance is also useful for contrasting complex objects and can be applied to signal processes and engineering (see e.g. Kolouri et al. (2017)). We refer to Bernton et al. (2019) for an survey of the Wasserstein distance. Many data analysis techniques in computer vision, imaging (e.g. for color/texture processing or histograms comparisons), and more general machine learning problems about regression, classification, and generative modeling are often based on optimal transportation theory (see Peyré, Cuturi, and others (2019)). For an overview on statistical use of the optimal transportation see Panaretos and Zemel (2020) and for a book-length discussion see Ambrosio and Gigli (2013). Recently, the Wasserstein distance has become popular specially for inference in generative adversarial networks (see e.g., Arjovsky, Chintala, and Bottou (2017)).

However, to the best of our knowledge, only a limited number of papers has been investigating the use of the optimal transport theory in the statistical analysis of time series analysis (see Ni et al. (2020)). Our purpose is to fill this gap in the literature and study the applicability of the Wasserstein distance (or, more generally, of some results from optimal transportation theory) for the statistical analysis of time series, via frequency domain techniques. The key argument for moving from the time domain to the frequency domain is that we are dealing with data that are independent and identically distributed (i.i.d). The assumption of i.i.d. data facilitates, as it is often the case in statistics, the estimation of parameters in a model.

We propose a novel class of minimum distance estimator (see Basu, Shioya, and Park (2019)) by minimizing the distance between the theoretical and empirical distribution of the standardized periodogram ordinates (SPOs). This program is supported by the fact that the method to replace the maximum likelihood estimator with minimum Wasserstein distance has already been applied, for instance in astronomy and climate science (see Bernton et al. (2019)). Additionally, consistency properties of estimators based on the Wasserstein distance has already been studied by Bassetti, Bodini, and Regazzini (2006) and Bernton et al. (2019). In our case, we study the properties (bias, variance, consistency, etc.) of our new estimators by means of Monte-Carlo experiments and compare them to the state-of-the-art estimator, the Whittle's estimator (see Whittle (1953)). We analyze our results for several types of distribution (standard, heavy-tailed and skewed) and focus mainly on large sample sizes to satisfy the i.i.d conditions. Our results are promising and open the possibility of further research.

1.2 Organization

In the first chapter, we review the main concepts of the optimal transport theory and provide the definitions and key mathematical tools necessary to understand our methodological development. In the second chapter, we briefly recall the Whittle's estimation theory and the key related results. Then, we introduce our novel estimators and compare them to the state-of-art estimator (Whittle). We conclude mentioning some possible research's directions which can make use of this thesis as a stepston. The R-codes needed to reproduce the exercises/plots/tables included in this thesis are available on Github at the link https://github.com/ManonFelix/Semidiscrete_estimation_ts.

2 Measure transportation

This chapter aims to explain the main principles behind the theory of optimal transport. The original formulation of the optimal transport problem was given by Monge (1781). He proposed a way to calculate the most effective strategy for moving a large amount of sand from one place to another with the least amount of effort required. In mathematical terms, given a source measure μ, target measure ν supported on sets X and Y respectively and a transportation cost function $c(x, y)$ the goal is to find a transport map $T : X \to Y$ such as

$$\min_{\nu = T_{\#}\mu} \int_X c(x, T(x)) \mathrm{d}\nu(x)$$

where the constraint $\mu(T^{-1}(A) = \nu(A) \implies \nu = T_{\#}\mu$,$\forall A$ ensures that all the mass from μ is transported to ν by the map T.

The notation $\nu = T_{\#}\mu$ means that the map T pushes forward μ to a new measure ν and therefore $T_{\#}$ is called the pushforward operator.

Monges problem (defined using $c(x,y) = |x-y|$) remained open until the 1940 s, when it was revisited by Kantorovich (1942). In his reformulation, he seeks for a transport plan and allows mass splitting. Therefore, he proposed to compute how much mass gets moved from x to y and define a joint measure (a coupling) $\pi \in \mathcal{P}\left(\mathbb{R}^d, \mathbb{R}^d\right)$ which satisfies for all $A, B \in \mathcal{B}\left(\mathbb{R}^d\right)$: $\pi\left(A \times \mathbb{R}^d\right) = \mu(A),\ \pi\left(\mathbb{R}^d \times B\right) = \nu(B)$. We denote by $\Pi(\mu,\nu)$ the set of transport plans between μ and ν (i.e. couplings). Then, p-Wasserstein distance is defined as

$$W_p(\mu,\nu) = \left(\min_{\pi \in \Pi(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^p d\pi(x,y)\right)^{\frac{1}{p}}. \tag{1}$$

Obtaining a closed-form expression for W_p is typically impossible. However, the case of one dimensional probability densities, say $f_S(x)$ and $f_T(x)$ with cumulative distribution functions $F_S(x)$ and $F_T(x)$, is specifically interesting as the Wasserstein distance (a.k.a the Earth Mover's Distance (EMD)) has the following expression

$$\mathcal{W}_1(\mu,\nu) = \int_{\mathbb{R}} |F_S(x) - F_T(x)|\, dx \tag{2}$$

The possibility of using this closed-formed solution to conduct inference on time series motivates the thesis.

Cuturi (2013) introduced a modified version of the Wasserstein distance:

$$W_\lambda(\mu,\nu) = \int_{x,y} C(x,y)\mathrm{d}\pi(x,y) + \lambda \int \log\left(\frac{\pi(x,y)}{\mathrm{d}\mu(x)\mathrm{d}\nu(y)}\right)\mathrm{d}\pi(x,y). \tag{3}$$

Minimizing Eq. 3 leads to the so called Sinkhorn divergence. This divergence is obtained adding to the original optimal transportation problem an entropic regularization term (right part). When λ is small, the Sinkhorn divergence approximates the Wasserstein distance. In contrast to the Wasserstein distance, the regularized Sinkhorn divergence is differentiable and smooth, so it yields some computational advantages in terms of optimization problems.

In this thesis, we use all the concepts presented in this section in order to establish new minimum distance estimators in time series analysis.

3 Inference in the frequency domain

Before introducing our estimators, let us first provide an overview of the theory that is commonly applied to conduct inference in the frequency domain.

Consider a stationary process $\{Y_t\}$ of n observations $y_{1:n} = y_1, \ldots, y_n$. We focus on linear time series $\{Y_t\}$ satisfying

$$\phi(L)(1-L)^d Y_t = \varphi(L)\epsilon_t$$

where $LX_t = X_{t-1}$ (back shift operator).

$\phi(z)$ and $\varphi(z)$ are the auto-regressive and moving average polynomial of order p and q respectively. The time series $\{Y_t\}$ may or may not have long memory depending on the value of d. When $0 < d < 0.5$ the process is called a long-memory process and are extensively applied in finance (see e.g Tsay (2005)) . In the literature, we often rewrite d as $H = d + 1/2$, which is called Hurst exponent.

Our setting is of *semiparametric nature*: we have an Euclidean parameter θ characterizing the auto-regressive and moving average polynomials and the long memory, but we do not assume any distribution for the innovation term ϵ_t (so the innovation density is an infinite dimensional nuisance parameter). For the sake on numerical illustration, we present our results for the case when $\epsilon_t \sim N\left(0, \sigma_\epsilon^2 = 1\right)$ but also underlying innovation densities with fatter tails (like e.g. Skew t (see Azzalini and Capitanio (2003)) and Student t).

To conduct inference on the model parameters $\theta = \left(\sigma_\epsilon^2, d, \phi_1, \ldots, \phi_p, \varphi_1, \ldots, \varphi_q\right)$ of long-memory processes, we could assume that ϵ_t is normally distributed. Thanks to this assumption, we can write the likelihood of the process and optimize it to find $\hat{\theta}$. Nevertheless, this approach is extremely time-consuming and can even be unfeasible due to the strong dependence and long-memory properties of the process. Instead, we tackle the problem in the frequency domain and work on Fourier frequencies rather than time data. The frequency domain approach represents a time series into combination of sinusoids.

The main tool utilized in the frequency domain is the spectral density. The spectral density $f(\lambda_j, \theta)$ of Y_t

$$f(\lambda_j, \theta) = \left|1 - e^{i\lambda}\right|^{-2d} \frac{\sigma_\epsilon^2}{2\pi} \frac{|\varphi(\exp\{-i\omega\})|^2}{|\phi(\exp\{-i\omega\})|^2}$$

where $\varphi(x) = 1 - \sum_{k=1}^{p} \varphi_k x^k$ and $\phi(x) = 1 + \sum_{k=1}^{q} \phi_k x^k$. λ_j are the fundamental Fourier frequencies where $\lambda_j = 2\pi(j/n), j \in \mathcal{J} = \{1, 2, \ldots, [(n-1)/2]\}$.

The spectrum of a time series can be estimated by the method of moment. Its sample analogue is called the periodogram $I(\lambda_j) = \frac{1}{2\pi n} \left|\sum_{t=1}^{n} \left(Y_t - \bar{Y}_n\right) e^{it\lambda_j}\right|^2$. The periodogram is asymptotically unbiased. Nevertheless, it is an inconsistent estimator; see e.g. Priestley (1981).

An key result showed by Priestley (1981) and Brillinger (2001) is that the periodogram ordinates are asymptotically independent and exponentially distributed with rate equal to the spectral density. In other words, the standardized periodogram ordinates are asymptotically independent and have an exponential distribution with rate one. This idea was introduced and exploited by Whittle (Whittle (1953)), who proposed to minimize the Whittle approximated likelihood:

$$L_W(\theta) = \frac{1}{2\pi} \left[\int_{-\pi}^{\pi} \ln f(\lambda, \theta) d\lambda + \int_{-\pi}^{\pi} \frac{I(\lambda)}{f(\lambda, \theta)} d\lambda\right] \tag{4}$$

which is derived from the fact that the SPOs are asymptotically identically distributed according to an exponential distribution. Eq. 4 can be rewritten by separating the variance component from the rest of the parameters vector as

$$L_W(\theta^*) = \sum_{j \in \mathcal{J}} \frac{I(\lambda_{j:n})}{f(\lambda_{j:n}, \theta^*)} \tag{5}$$

where $f(\lambda_{j:n}, \theta^*) = 2\pi\sigma_\epsilon^2 f(\lambda_{j:n}, \theta^*)$ and $\theta^* = (1, \eta = (d, \phi_1, \ldots, \phi_p, \varphi_1, \ldots, \varphi_q))$.

The following minimization problem can be set up to find Whittle estimator; see e.g. Beran (1994) for a book-length discussion. First, minimize $\arg\min_\eta L_W(\theta^*) = \arg\min_\eta L_W(\eta)$ which yields to $\hat{\eta}$. Then, set $\hat{\sigma}_\epsilon^2 = 2\pi L_W(\hat{\eta})$. One can prove the consistency of the parameter $\hat{\theta}^*$. Additionally, the parameter is $\sqrt{n}$-consistent and converges to a normal distribution. In the case of underlying Gaussian innovation terms, $\hat{\theta}$ achieves the Cramer-Rao lower bound. The Whittle's estimator can also be applied for ARIMA(p, q) process and remains $\sqrt{n}$-consistent and asymptotically normal. We therefore use this parameter as our reference to compare our results as it is still the state-of-the-art methodology.

4 Methodology

4.1 Problem settings

Our goal is to find the parameter $\eta = \theta^*$ of a time series model in the parameter space $\theta^* \in \Theta$ with dimension $\Theta \subset \mathbb{R}$ that minimizes the distance between the empirical and theoretical cumulative distributions of the SPOs. We denote the distance or divergence used by $\mathcal{D}$ and write our minimum distance estimator such as

$$\hat{\theta}^* = \underset{\theta^* \in \Theta}{\operatorname{argmin}}\ \mathcal{D}\left(\mu, \nu\right).$$

where μ is the theoretical exponential distribution and ν is the empirical distribution of the SPOs. In our study, several estimators are proposed and therefore $\mathcal{D}$ is redefined for each optimization problem. For instance, when $\mathcal{D}$ is the Wasserstein distance, we denote the corresponding minimum Wasserstein estimator (MWE) as $\hat{\theta}^*_{MWE}$. For the sake of simplicity, we assumed the variance of the innovation term σ_ϵ^2 to be known and equal to one. Hence, our parameter vector to be estimated is

$$\theta^* = (d, \phi_1, \ldots, \phi_p, \varphi_1, \ldots, \varphi_q).$$

In addition to that, we focus first on processes with underlying Gaussian distribution and then extend to other distributions with fat tails.

4.2 Estimation methods

4.2.1 Minimum Wasserstein Estimator

The Wasserstein distance when $p = 1$ is given in Eq. 2 when F_ν is the empirical cumulative distribution of the SPOs, we estimate it by

$$\hat{F}_\nu(x) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}_{X_j \leq x}$$

where X_j are the SPOs of a time series process asymptotically exponentially distributed.

To compute F_μ, as it is common in machine learning literature about generative models, we initially thought to generate exponential random variables and stack then in a vector (Z, say). In mathematical terms, for a sample size n, we generate $(n-1)/2 = m$ observations $z_1, ..., z_m$ of random variables $Z_1, ..., Z_m$ following an exponential distribution with rate one and store them in a vector Z . Since $p = 1$, the Wasserstein distance can be approximated by

$$\mathcal{D}(\mu, \nu) = \frac{1}{m} \sum_{j=1}^{m} |x_j - z_j| \tag{6}$$

where x_j are the SPOs of a time series process asymptotically exponentially distributed and z_j are the observations generated using $Z \sim Exp(1)$. Minimizing Eq. 6 leads to the minimum Wasserstein estimator noted

$$\hat{\theta}^*_{MWE} = argmin\ \mathcal{D}(\mu, \nu).$$

Figure 1 displays the Wasserstein loss function of two FARIMA$(0, d, 0)$ processes. The top plot shows a smooth and concave loss function with a global minimum that is the same value as the Whittle's estimator's. On the other hand, the lower shows a wiggly function around the true value of the parameter. Thus, if one uses the Wasserstein loss to estimate the parameter d, these two phenomena arise: we end up with either a smooth function containing a global minimum or a function that fluctuates and has several local minima. Through this thesis, we tackle these issues and introduce estimators that are defined by loss functions which do not suffer from these problems.

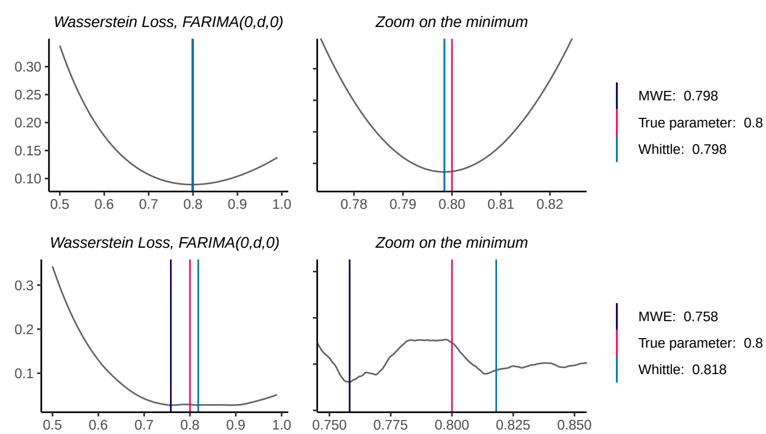


Figure 1: *Wasserstein loss functions of two FARIMA(0,d,0) processes where H = 0.8 (d = 1.2). The sample size is 3001. The left column display the entire loss functions for all possible parameter values that a long-memory process can take (0.51 < H < 0.99). The right column is a zoom on the functions.*

Our first finding is that the computed distance might vary widely around the true parameter value and its value depends heavily on the sample simulated from an Exp(1), i.e. observations $z_1, ..., z_m$. As a consequence, the estimated model parameter(s) $\hat{\theta}_{MWE}$ typically depend heavily on the random exponential variables generated to conduct the optimization.

To illustrate the problem, on Figure 2 we continue with the process used to plot the second line on Figure 1 and simply change the seed with which the vector Z is generated. We remark that we are now dealing with a loss function that is smooth and has a global minimum that is precisely the true value of the parameter $\theta^* = H = 0.8$. Therefore, we see that by modifying the vector Z, we can obtain a more appropriate loss function. We find this aspect potentially problematic for numerical optimization: by definition a random vector cannot be controlled and we may obtain biased estimates simply because of the simulated variables needed for the optimization procedure.

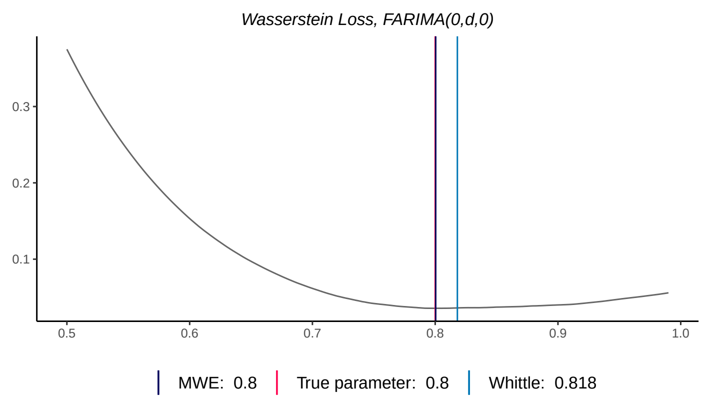


Figure 2: *Wasserstein loss function of the FARIMA(0,d,0) process (bottom one) of Figure 1 computed with another random vector.*

In order to get a better overview of the behavior of the MWE when the vector Z changes, we compute $k = 200$ times $\hat{\theta}^*_{MWE}$ for a given process of size $n = 3001$. Then, we plot the results on Figure 3. We can observe that, even for large sample size, the estimated parameter depends heavily on the random vector Z. Nevertheless, the mean remains relatively close to the true value.

To cope with this problem of dependence between the random vector Z and the parameter estimate, we are going to explore two options.

4.2.2 Mean of Minimum Wasserstein Estimators

Option A: for the simulated times series, we generate several exponential random variables and stack them in vectors. Then, we estimate the model parameter for each of the simulated vector and report the mean of the estimated parameter.

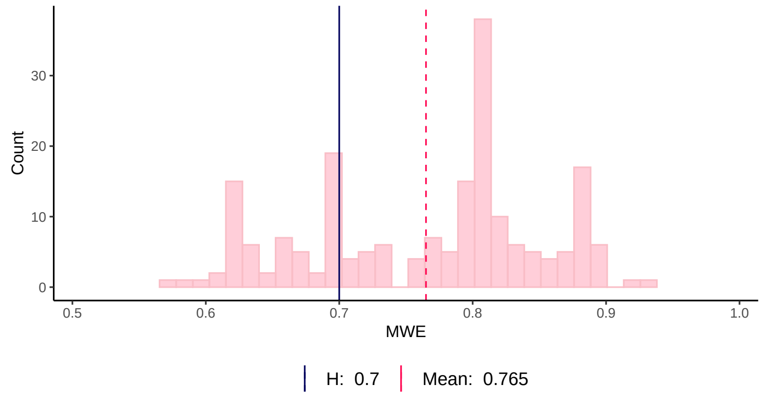


Figure 3: *We simulate 200 vectors following an exponential distribution and then compute the MWE of a FARIMA(0,d,0) process with n = 3001. The blue line is the true parameter H = 0.7 value and the red dashed line is the mean of the MWE.*

For illustration, based on the same process than in Figure 3 with $n = 3001$, we generate $k = 10, 20, 50, 100, 200, 500, 1000$ random vectors, estimate the k parameters and then report the mean. Thus, the mean becomes our estimator and we note it $\hat{\theta}^*_{MMWE}$. The results are listed in Table 1. As k increases, the average becomes progressively closer to the true parameter.

k	1	10	20	50	100	200	500	1000
$\hat{\theta}^*_{MMWE}$	0.807	0.73	0.727	0.726	0.721	0.719	0.714	0.714

Table 1: *Mean of the minimum Wasserstein estimators for a FARIMA$(0, d, 0)$ of size $n = 3001$ by varying the value of k, i.e. the number of exponential random vectors generated. The true value is 0.7.*

4.2.3 Minimum Semidiscrete Wasserstein Estimator

Option B: instead of using the empirical cumulative distribution function (c.d.f) of exponential random variables generated from a computer, we plan to use the c.d.f of exponential variables with rate one for the SPOs, namely $F(x) = 1 - e^{-x}$. Therefore, the Wasserstein distance becomes:

$$\int_{\mathcal{X}} \left| \hat{F}(x) - (1 - e^{-x}) \right| \mathrm{d}x \tag{7}$$

where $\hat{F}(x) = \frac{1}{m} \sum_{j=1}^{m} 1_{X_j \leq x}$ and x_j are the SPOs of a process. To compute this distance, we replace $\hat{F}_\mu(z) = \frac{1}{m} \sum_{j=1}^{m} \mathbf{1}_{Z_j \leq z}$ by $F_\mu = 1 - e^{-x}$. We use a trapezoidal integration to approximate the integral. Thanks to this second option, there is no longer randomness in our process estimation. The corresponding loss functions for two FARIMA$(0, d, 0)$ processes using this method are showed in Figure 4

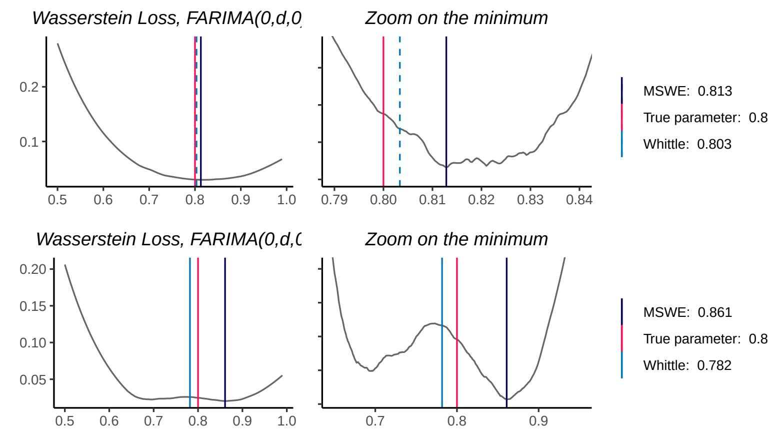


Figure 4: *Semidiscrete Wasserstein loss functions for two FARIMA(0,d,0) processes. The sample size is 3001 and the true parameter value is 0.8.*

Still, another problem persists. The Wasserstein loss, even for large sample size, is often not well-shaped (i.e smooth and concave): it may contain several local minima (see e.g. Figure 1). This concern leads to biased estimates with large variance. It should also be noted that the loss shape degenerates even more when n decreases (see Figure 5). So far, we are not able to explain why there is such diversity in the shape of the loss functions.

Figure 5: *Wasserstein loss function of a FARIMA(0,d,0) process with sample size equal to 51.*

4.2.4 Minimum Weighted Wasserstein Estimator

After searching a way to fix this problem, we found that by putting some weights in the loss function defined by the Wasserstein distance, we obtain a much more regular optimization problem. Therefore, we seek the parameter that minimizes

$$\mathcal{W}_1(\mu, \nu) = \int_{\mathbb{R}} |F_\mu(t) - F_\nu(t)| \, dt \tag{8}$$

where the empirical cumulative distribution of the SPOs is $\hat{F}_\nu(x) = \sum_{i=1}^{m} w_j 1\{X_i \leq x\}$ and the weights w_j are

$$w_j = \frac{\frac{I(\lambda_j)}{f(\lambda_j;\theta)}}{\sum_{j=1}^{m} \frac{I(\lambda_j)}{f(\lambda_j;\theta)}}. \tag{9}$$

We call the estimator minimizing this new distance our minimum weighted Wasserstein estimator (MWWE) θ^*_{MWWE}.

The employment conditions of R packages to calculate the distances used in this thesis required that the weights sum to 1 and that are comprised between 0 and 1. Therefore, our first intuition is to use the weights proposed in Eq. 9. Intuitively, these weights give more leverage to extreme values while computing the Wasserstein distance.

Figure 6 shows the same process and vector Z as Figure 1 but applies the weights to calculate our weighted Wasserstein distance. We can observe that the weighted Wasserstein loss function is immediately smoother and contains a minimum which is even closer to the true parameter than the Whittle's estimator.

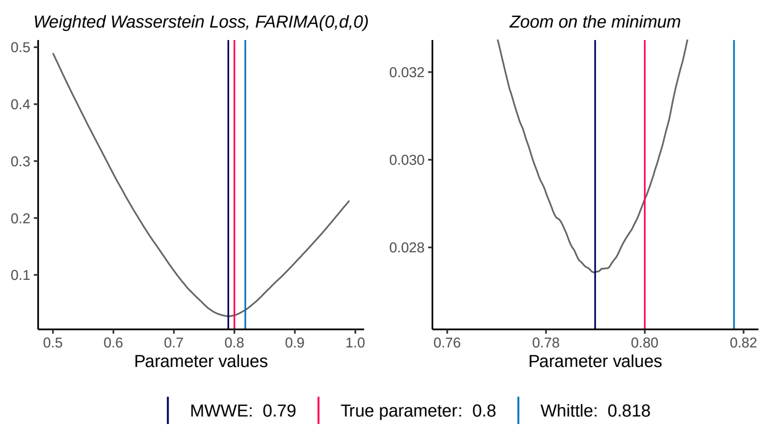


Figure 6: *Weighted Wasserstein loss function of the FARIMA(0,d,0) process on Figure 1 (bottom).*

It is important to note that the weights applied here are not optimal and, therefore, this question remains open and subject to further analysis. However, the weights presented in this section work well especially for ARMA(p, q) processes as illustrated on Figure 7. The shape of the loss function using the weighted Wasserstein estimator suggests that we could obtain an estimator with small variance. We could, for instance, implement a minimum expected Wasserstein estimator defined by Bernton et al. (2019). This option could be investigated in future research.

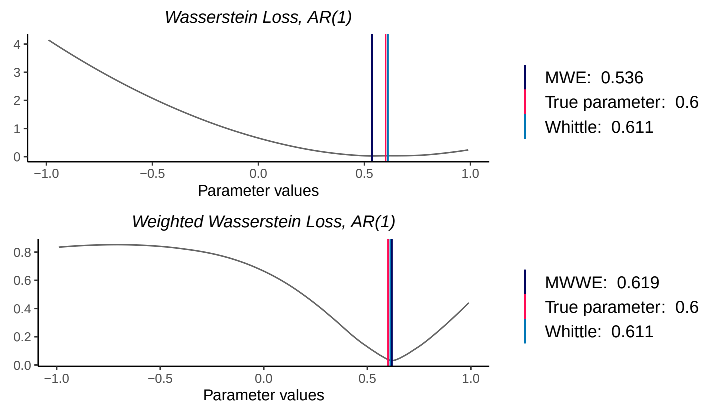


Figure 7: *Wasserstein loss function and weighted Wasserstein loss function of a Gaussian AR(1) process. The sample size is 3001 and the true parameter is 0.6.*

4.2.5 Minimum Sinkhorn Estimator

A second idea is to employ the Sinkhorn divergence (see Eq. 3) to estimate our parameter based on Cuturi (2013). The regularization term should make the loss function smoother. On Figure 8, we compare the loss function when we employ the Wasserstein distance or the Sinkhorn divergence to estimate our parameter. Indeed, we end up with a smooth and concave function. The reached minimum is very close to the true value. A good property with the Sinkhorn divergence is that it remains smooth even for a very small sample size.

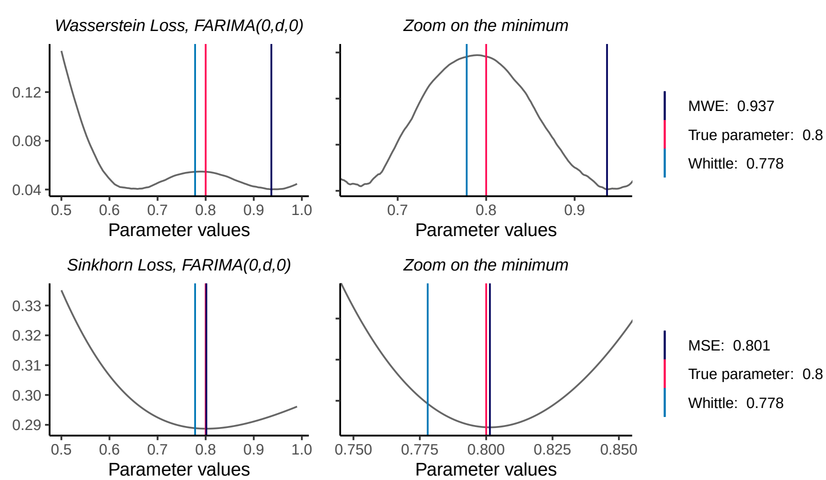


Figure 8: *Top: Wasserstein loss function of a FARIMA(0,d,0) process. Bottom: Sinkhorn loss function of the same process. The sample size is 1801.*

Following this, we are confronted to an important choice: which values to select for λ? As a reminder, when λ is very small, the Sinkhorn divergence approximates the Wasserstein distance. In order to choose the optimal lambda we suggest to implement classical machine learning techniques to perform model selection such as cross validation, leave-one-out, etc. For more information see for example Friedman et al. (2001) Chapter 7.

To illustrate, we randomly divide a time series into 2 groups C_1 (80%), C_2 (20%) also called folds. We treat the first group as train and the second group as validation/test. For a selection of λ, we estimate our parameter, the minimum Sinkhorn estimator (MSKE), on the train set and then use the corresponding $\hat{\theta}_{MSKE}$ to predict the time data of our test set. For simplicity, we demonstrate the procedure with an AR(1) process, $Y_t = \phi_1 Y_{t-1} + \epsilon_t$ where $\epsilon_t \sim N(0,1)$ and $\theta^* = \phi_1 = 0.6$. After estimating the parameter thanks to the train set, we substitute its value in $\hat{\epsilon}_t = Y_t - \hat{\theta}^*_{MSKE} Y_{t-1}$ where $t = 2, ..., l$. l is the length of the testing vector and depends on which ratios we choose to split our time process Y_t in our case $80\% - 20\%$. Then, we use the predictions of the error terms to compute the Mean Squared Error for a given λ:

$$MSE_\lambda \ of \ the \ test \ set = \frac{1}{l}\sum_{t=2}^{l} \hat{\epsilon}_t^2 = \frac{1}{l}\sum_{t=2}^{l}(Y_t - \hat{\theta}^*_{MSKE} Y_{i-t})^2$$

We repeat this method for several lambda values and plot the results on Figure 9. The minimum testing error is achieved when $\lambda = 0.1$ given an estimate $\hat{\theta}^*_{MSKE_{\lambda=0.1}} = 0.572$ which is close to the true parameter value $\theta^* = 0.6$.

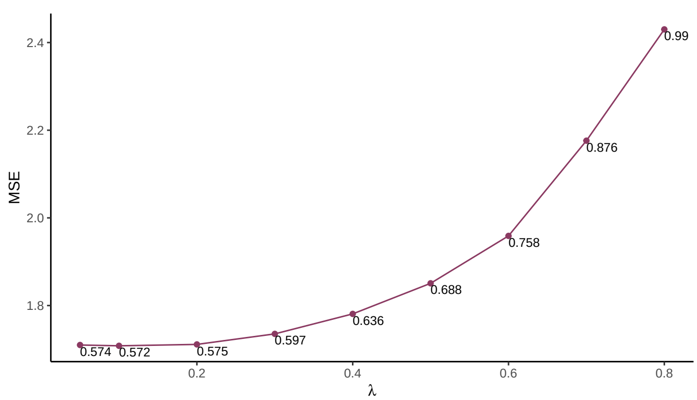


Figure 9: *Testing MSE vs lambda values for an AR(1) process. The sample size is 4001 and the true parameter value is 0.6. For this process, the minimum MSE is achieved when lambda = 0.1.*

5 Results

5.1 Monte Carlo Simulations

Our criterion for evaluating the performance of each of the estimators is, as it is often the case in machine learning, the Mean Squared Error (MSE)

$$MSE(\hat{\theta}^*, \theta^*) = \frac{1}{mt} \sum^{mt} (\hat{\theta}^* - \theta^*)^2 \approx \mathrm{Var}(\hat{\theta}^*) + \mathrm{Bias}^2(\hat{\theta}^*)$$

where mt is the number of Monte Carlo simulations, i.e. the number of simulated processes. The MSE represents the bias-variance trade-off which typically emerges in statistics when it comes to model selection.

5.1.1 Long-memory Process

Firstly, we simulate $mt = 400$ stationary FARIMA$(0, d, 0)$ processes of size $n = 3201$ and $H = 0.8$ ($d = 0.3$) according to

$$(1 - L)^{0.3} Y_t = \epsilon_t.$$

For each process, we compute the Whittle's estimator $\hat{\theta}^*_{WH}$, the minimum Wasserstein estimator $\hat{\theta}^*_{MWE}$, the mean of the minimum Wasserstein estimators $\hat{\theta}^*_{MMWE}$, the minimum semidiscrete Wasserstein estimator $\hat{\theta}^*_{MSWE}$, the minimum weighted Wasserstein estimator $\hat{\theta}^*_{MWWE,\ k}$ and the minimum Sinkhorn estimator $\hat{\theta}^*_{MSKE,\ \lambda}$. Figure 10 reports the results.

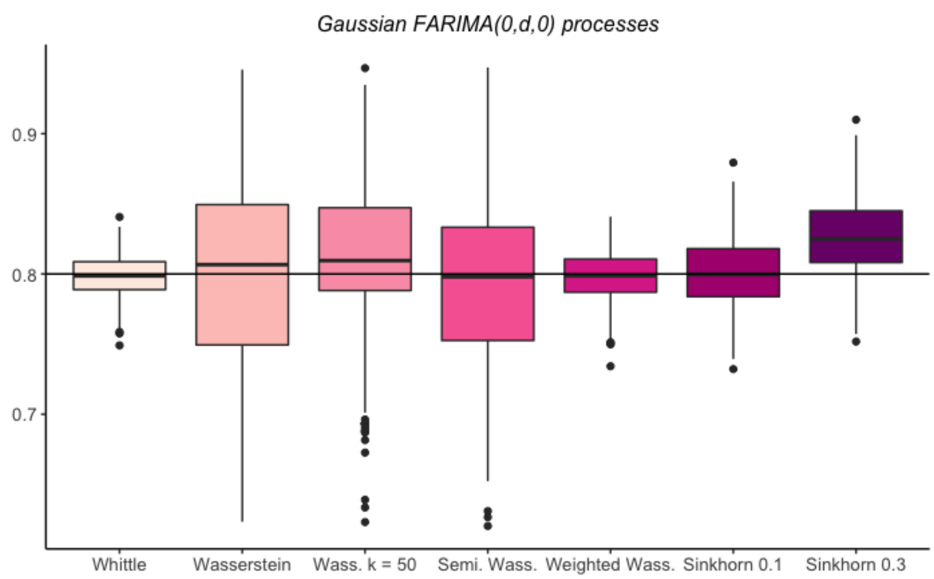


Figure 10: *Boxplots of all the estimators presented during this thesis. The sample size of the 400 simulated FARIMA(0,d,0) is 3201.*

We note that the minimum Wasserstein estimator has very high variance due to the problems mentioned earlier. As expected, by using either the mean of the minimum Wasserstein estimators or the minimum semidiscrete Wasserstein estimator, we can reduce the variance. Our minimum weighted Wasserstein estimator is similar to Whittle's estimator in terms of Mean Squared Error. Indeed, both new estimators have small variance and no bias. The density of these two estimators is in Figure 11. Both distributions are centered around the true parameter and have similar shape. Nevertheless, the minimum weighted Wasserstein estimator has larger tails, i.e larger variance.

The use of the Sinkhorn distance also seems promising but depends on the λ parameter. When λ is equal to 0.1, the minimum Sinkhorn estimator is well centered around the true parameter and has a reasonable variance. Here, we do not choose λ by cross-validation because of the time needed for computation. We compute the Sinkhorn distance where the value of λ is determined in advance. It is expected that by performing a selection of lambda parameters we will obtain even better results.

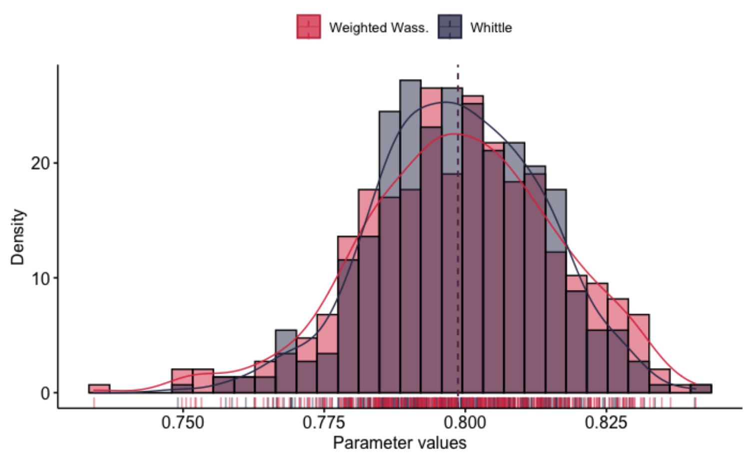


Figure 11: *Distribution of the Whittle's estimator and the weighted Wasserstein estimator estimated with the 400 simulations of Figure xx.*

	$MSE(\hat{\theta}^*, \theta^*)$
Distribution	**Gaussian**
$\hat{\theta}^*_{Whittle}$	0.00021
$\hat{\theta}^*_{MWE}$	0.00436
$\hat{\theta}^*_{MMWE,\ k=50}$	0.00324
$\hat{\theta}^*_{MSWE}$	0.00418
$\hat{\theta}^*_{MWWE}$	0.00030
$\hat{\theta}^*_{MSKE,\ \lambda=0.1}$	0.00061
$\hat{\theta}^*_{MSKE,\ \lambda=0.3}$	0.00144

Table 2: *Mean Squared Errors of Figure 10.*

5.1.2 Heavy-tailed Distribution

Mikosch et al. (1995) showed that the fatter the tails of the innovation distributions, the faster the Whittle's estimator converges to the true parameter value. Regarding the Wasserstein loss function, it becomes smooth and concave when the error distribution is heavy-tailed (see Figure 12), even for small sample sizes. Therefore, the Whittle's estimator and the MWE are usually very close in value.

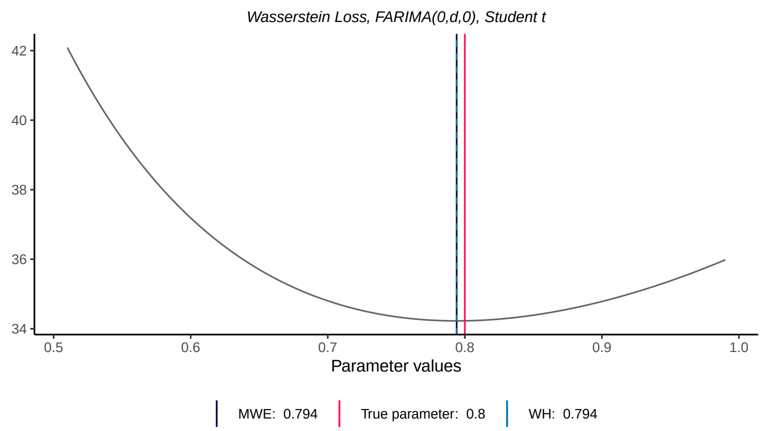


Figure 12: *Wasserstein loss function of a FARIMA(0,d,0) process distributed according to a Student t distribution with degree of freedom equal to 2. The sample size is equal to 3001.*

We simulate again $mt = 400$ FARIMA(0,d,0) processes with a Student t underlying distribution with degree of freedom equal to 2. On Figure 13, we note that all estimators (apart from those based on the Sinkhorn distance) have a considerably smaller variance than when the process is Gaussian and, consequently, a smaller MSE (see Table 3).

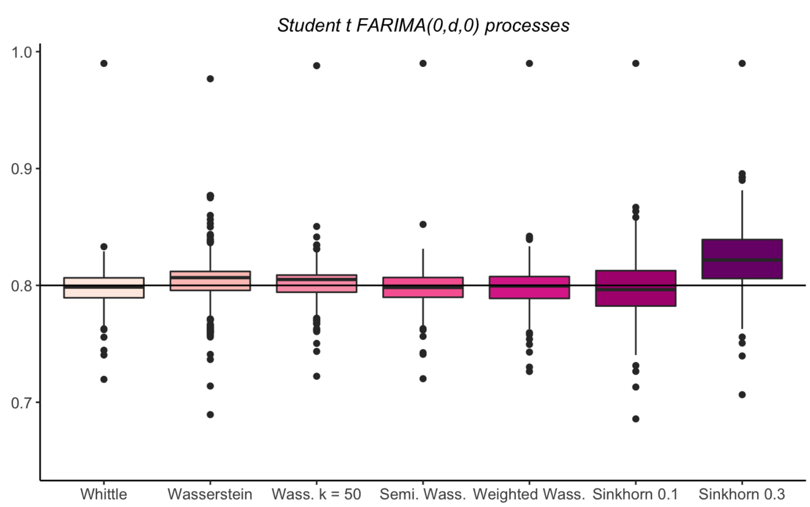


Figure 13: *Boxplots of all the estimators presented during this thesis. The sample size of the 200 simulated FARIMA(0,d,0) is 3201 and the underlying distribution is a Student t with degree of freedom equal to 2.*

5.1.3 Skewed and Heavy-tailed Distribution

Let us now focus on the case where the distribution of the innovation terms remains heavy-tailed but on top of that skewed. The skew t distribution was recently developed by Azzalini and Capitanio (2003). It is related to a standard skew normal random variable T and a random variable M following a chi-squared distribution with ν degree of freedom by the equation:

$$R = \frac{T}{\sqrt{\frac{M}{\nu}}}.$$

Then the linear transformation $X = \mu + \sigma R$ has a skew-t distribution with parameters μ, σ, α, and ν and the corresponding notation $ST(\mu = 0, \sigma = 1, \gamma, \nu)$ to denote the skew t random variable X. For example, we consider the underlying distribution of the process as a skew t distribution with degree of freedom equal to 2 and skewness parameter γ equal to 4. The corresponding density function is represented on Figure 14).

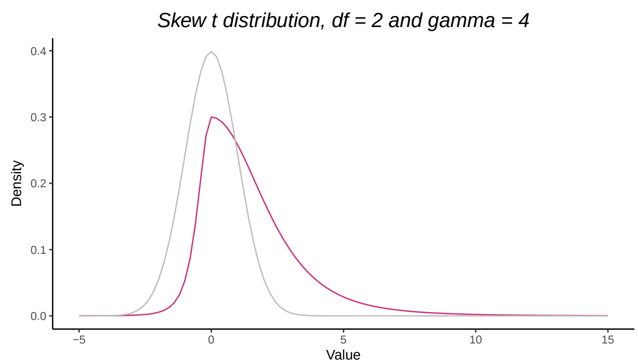


Figure 14: *Skew t distribution with degree of freedom = 2 and gamma = 4 (pink) and standard normal distribution (grey).*

On Figure 15, we compute the Wasserstein loss function in the case of skew t distributed FARIMA(0,d,0) process. The loss remains smooth and concave, as it is the case for a Student t underlying distribution, but the parameter minimizing the loss function has a much larger value than the true parameter ($0.945 > 0.8$). The same effect occurs for the Whittle's estimator, we also overestimate the parameter value (see Figure 15).

We compare all the estimators when the underlying distribution is a skew t on Figure 16. All estimators are biased in the sense they overestimate the value of the parameter. Some are more biased than others and curiously the minimum Wasserstein estimator is the least biased but it has a greater variance than others estimators. Regarding the Mean Squared Error values listed in Table 3,

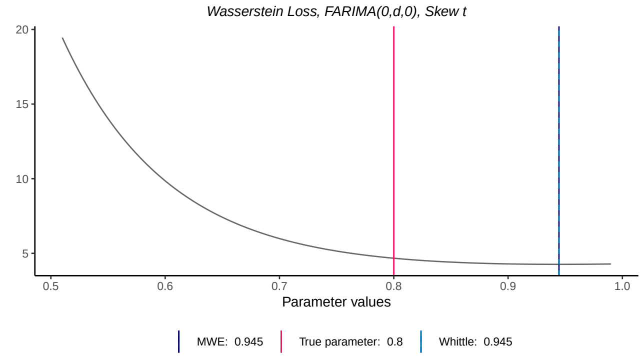


Figure 15: *Wasserstein loss function of a FARIMA(0,d,0) process distributed according to a skew t distribution with degree of freedom = 4 and gamma = 2. The sample size is equal to 3001.*

all our new estimators surpass the Whittle's estimator except the minimum weighted Wasserstein estimator. This gain in terms of MSE is mainly due to a gain in bias since most estimators have a similar variance to Whittle's. The estimator with the smallest MSE is the one obtained by means of the Sinkhorn distance when $\lambda = 0.1$.

In order to verify that this bias does not come from the fact that the error distribution is skewed we also simulated processes coming from a skew normal distribution (left and right skewed). Then, we reproduced the same boxplots and what we observe is that there is no bias. All these results are available on the Github link mentioned at the beginning of this thesis.

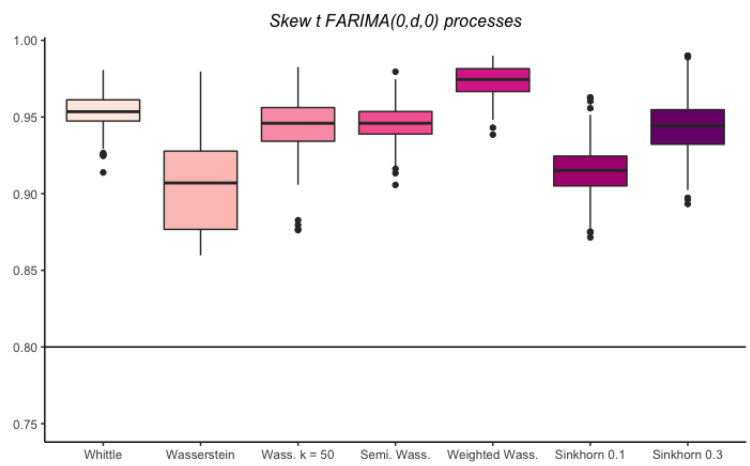


Figure 16: *Boxplots of all the estimators presented during this thesis. The sample size of the 200 simulated FARIMA(0,d,0) is 3201 and the underlying distribution is a skew t with df = 4 and gamma = 2.*

	$MSE(\hat{\theta}^*, \theta^*)$	$MSE(\hat{\theta}^*, \theta^*)$
Distribution	**Student t**	**Skew t**
$\hat{\theta}^*_{Whittle}$	0.000295	0.023688
$\hat{\theta}^*_{MWE}$	0.000565	0.012266
$\hat{\theta}^*_{MMWE,\ k=50}$	0.000294	0.021146
$\hat{\theta}^*_{MSWE}$	0.000305	0.021387
$\hat{\theta}^*_{MWWE}$	0.000354	0.030284
$\hat{\theta}^*_{MSKE,\ \lambda=0.1}$	0.000653	0.013515
$\hat{\theta}^*_{MSKE,\ \lambda=0.3}$	0.001260	0.020946

Table 3: *Mean Squared Error of Figure 13 and 16.*

5.1.4 Additive Outliers

In the presence of contamination in the time series (e.g. additive outliers). For example, in the case of Gaussian FARIMA(0, d, 0) some of our estimators (in particular, the ones based on weighted Wasserstein distance and/or on the Sinkhorn divergence) seem to overperform Whittle's estimator in terms of MSE. To demonstrate this propriety we simulate $mt = 200$ FARIMA$(0, d, 0)$ contaminated by occasional isolated outliers. The processes $\{Y_t\}$ are distributed according to

$$Y_t = (1 - W_t)\, X_t + W_t\, (c \cdot V_t)$$

where $W_t \sim Bern(p)$, $V_t \sim t_2$ and c = 10.

In Table 4, we report the ratio between the MSE of the Whittle's estimator and the minimum weighted Wasserstein estimator for different values of $p = 0, 0.001, 0.01, 0.05$ (when $p = 0$ the process is not contaminated by outliers). The results suggest that when the time series is contaminated, the minimum weighted Wasserstein estimator overperform the Whittle's estimator in terms of MSE since as soon as we introduce noise the ratio becomes greater than 1.

p	0	0.001	0.01	0.05
ratio	0.682	1.208	1.105	1.012

Table 4: *Mean Squared Error of the Whittle's estimator divided by the Mean Squared Error of the minimum weighted Wasserstein estimator. The number of simulated time series is equal to 200 with sample size equal to 3001.*

5.1.5 Short-memory Process

We also aim to demonstrate the performance of our estimators for short-memory processes. To do this, we simulate $mt = 400$ auto-regressive processes of order 2 according to:

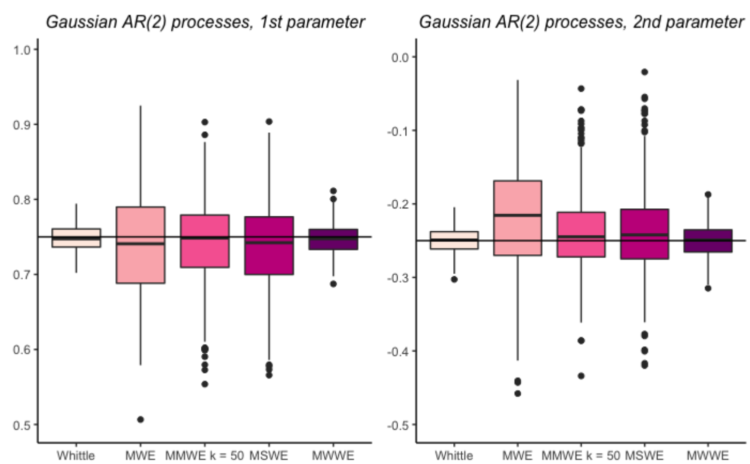
$$Y_t = 0.75Y_{t-1} - 0.25Y_{t-2} + \epsilon_t.$$

The processes are stationary since the three stationary conditions are met:

1. $\phi_2 < 1 + \phi_1$
2. $\phi_2 < 1 - \phi_1$
3. $\phi_2 > -1$

where $\phi_1 = 0.75$ and $\phi_2 = -0.25$.

We cannot include the Sinkhorn divergence in our comparison because the function used on R requires too much time to calculate this divergence and fails to converge. The results when $\theta^* \subset \mathbb{R}^2$ are on Figure 17 with corresponding MSE in Table 5. Again, we consider several distributions for ϵ_t: $\epsilon_t \sim N(0,1), \epsilon_t \sim t_2$ and $\epsilon_t \sim ST(\mu = 0, \sigma = 1, \gamma = 2, \nu = 4)$ (Gaussian, Student t and skew t).



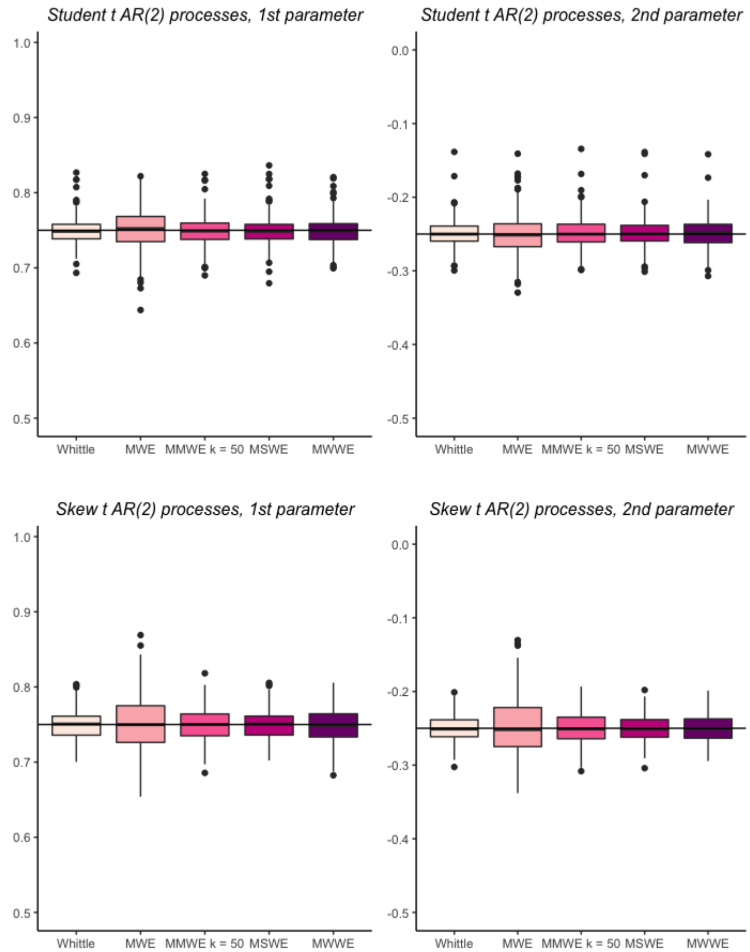


Figure 17: *Boxplots of the Whittle's estimator, MWE, MSWE, MMWE, MWWE for 400 AR(2) processes. The innovation terms densities are (in the order of apparition): Gaussian, Student t, Skew t. The left column is the first parameter (0.75) of the process, the right one is for the second parameter (-0.25).*

When the process is Gaussian, the MSE of the minimum weighted Wasserstein estimator is similar to Whittle's estimator. The other estimators have larger variance than Whittle's. As observed for processes with long-memory, when the tails of the error distributions are wider than those of the normal distribution, the minimum Wasserstein estimator converges to Whittle's estimator in terms of bias and variance. In the case of Student t innovation term, all estimators are relatively similar in terms of MSE (bias - variance). On the other hand, contrary to long-memory processes, we observe that the fact that the underlying distribution are skewed or not is not relevant during the estimation procedure and the estimator do not overestimate the true parameter. Indeed, the results for the Student t and the skew t distribution are very close.

$MSE(\hat{\theta}^*, \theta^* = \phi_i)$	ϕ_1	ϕ_2	ϕ_1	ϕ_2	ϕ_1	ϕ_2
Distribution	**Gaussian**		**Student t**		**Skew t**	
$\hat{\theta}^*_{Whittle}$	0.00028	0.00031	0.00027	0.00029	0.00033	0.00029
$\hat{\theta}^*_{MWE}$	0.00512	0.00703	0.00070	0.00066	0.00134	0.00143
$\hat{\theta}^*_{MMWE,\ k=50}$	0.00313	0.00317	0.00033	0.00038	0.00045	0.00044
$\hat{\theta}^*_{MSWE}$	0.00397	0.00407	0.00030	0.00033	0.00033	0.00030
$\hat{\theta}^*_{MWWE}$	0.00038	0.00045	0.00035	0.00038	0.00047	0.00038

Table 5: *Mean Squared Error of Figure 17.*

6 Conclusion

To conclude, we introduce, in this thesis, five new estimators that are based on minimum distance estimation. Our results suggest that we can outperform the state-of-the art estimation procedure when we are dealing with long-memory processes that have skewed underlying distributions. Moreover, it seems that our minimum weighted Wasserstein estimator can also be more efficient when the process is contaminated by occasional outliers. In the case of short memory processes, we have similar results to Whittle's estimator in terms of MSE. Through this thesis, we open the possibility for further research. Indeed, the weights are certainly not optimal and therefore would be subject to further investigation. As well as the choice of the regularization parameter when using the Sinkhorn divergence. The shape of the Wasserstein loss function and why the estimation procedure behaves better for certain vector Z_j also remains an opened question. We can also extend our research to other distances such as the energy distance:

$$D^2(F,G) = 2\int_{-\infty}^{\infty} (F(t) - G(t))^2 \, \mathrm{d}t.$$

Another important step is to compute the theory surrounding these estimators (consistency, robustness, etc.). To sum up, our results are promising and open the possibility of further researches.

References

Ambrosio, Luigi, and Nicola Gigli. 2013. "A User's Guide to Optimal Transport." In *Modelling and Optimisation of Flows on Networks*, 1–155. Springer.

Arjovsky, Martin, Soumith Chintala, and Léon Bottou. 2017. "Wasserstein Generative Adversarial Networks." In *International Conference on Machine Learning*, 214–23. PMLR.

Azzalini, Adelchi, and Antonella Capitanio. 2003. "Distributions Generated by Perturbation of Symmetry with Emphasis on a Multivariate Skew t-Distribution." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 65 (2): 367–89.

Bassetti, Federico, Antonella Bodini, and Eugenio Regazzini. 2006. "On Minimum Kantorovich Distance Estimators." *Statistics & Probability Letters* 76 (12): 1298–1302.

Basu, Ayanendranath, Hiroyuki Shioya, and Chanseok Park. 2019. *Statistical Inference: The Minimum Distance Approach*. Chapman; Hall/CRC.

Beran, Jan. 1994. *Statistics for Long-Memory Processes*. Vol. 61. CRC press.

Bernton, Espen, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. 2019. "On Parameter Estimation with the Wasserstein Distance." *Information and Inference: A Journal of the IMA* 8 (4): 657–76.

Brillinger, David R. 2001. *Time Series: Data Analysis and Theory*. SIAM.

Cuturi, Marco. 2013. "Sinkhorn Distances: Lightspeed Computation of Optimal Transport." *Advances in Neural Information Processing Systems* 26: 2292–2300.

Friedman, Jerome, Trevor Hastie, Robert Tibshirani, and others. 2001. *The Elements of Statistical Learning*. Vol. 1. 10. Springer series in statistics New York.

Kantorovich, Leonid V. 1942. "On the Translocation of Masses." In *Dokl. Akad. Nauk. USSR (NS)*, 37:199–201.

Kolouri, Soheil, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. 2017. "Optimal Mass Transport: Signal Processing and Machine-Learning Applications." *IEEE Signal Processing Magazine* 34 (4): 43–59.

Mikosch, Thomas, Tamar Gadrich, Claudia Kluppelberg, and Robert J Adler. 1995. "Parameter Estimation for ARMA Models with Infinite Variance Innovations." *The Annals of Statistics*, 305–26.

Monge, Gaspard. 1781. "Mémoire Sur La Théorie Des déblais Et Des Remblais." *Histoire de l'Académie Royale Des Sciences de Paris.*

Ni, Hao, Lukasz Szpruch, Magnus Wiese, Shujian Liao, and Baoren Xiao. 2020. "Conditional Sig-Wasserstein GANs for Time Series Generation." *arXiv Preprint arXiv:2006.05421.*

Panaretos, Victor M, and Yoav Zemel. 2020. *An Invitation to Statistics in Wasserstein Space.* Springer Nature.

Peyré, Gabriel, Marco Cuturi, and others. 2019. "Computational Optimal Transport: With Applications to Data Science." *Foundations and Trends in Machine Learning* 11 (5-6): 355–607.

Priestley, Maurice Bertram. 1981. *Spectral Analysis and Time Series: Probability and Mathematical Statistics.* 04; QA280, P7.

Santambrogio, Filippo. 2015. "Optimal Transport for Applied Mathematicians." *Birkäuser, NY* 55 (58-63): 94.

Tsay, Ruey S. 2005. *Analysis of Financial Time Series.* Vol. 543. John wiley & sons.

Villani, Cédric. 2009. *Optimal Transport: Old and New.* Vol. 338. Springer.

Whittle, Peter. 1953. "Estimation and Information in Stationary Time Series." *Arkiv för Matematik* 2 (5): 423–34.