Internship Report

Subject : An Analysis of Misspecification in Nonparametric Estimation of Autoregressive Processes

NECHNECH Idris

Research Supervisor :

FORTUNATI Stefano

SAMOVAR Lab of Institut Polytechnique de Paris (Telecom SudParis)

Date : 25/07/2025

Table of Contents

1 Introduction

In many estimation problems, observed data are modeled using parametric distributions whose structure depends on an unknown parameter vector. The objective is then to propose efficient estimators for these parameters and to evaluate their precision using theoretical tools such as the Cramér-Rao bound.
However, in real-world scenarios, the assumptions made about data distribution can be wrong. It then becomes necessary to broaden the classical framework to account for these misspecification errors. This report explores these two complementary approaches.
We first introduce the fundamentals of classical parametric statistics and statistics under misspecified models. We then consider concrete case studies using an autoregressive AR(1) model in both Gaussian and non-Gaussian contexts (specifically with noise following a Student's t-distribution). For this model, we implement and compare several estimators: the maximum likelihood estimator, the Newton estimator, M-estimators based on robust loss functions, and an R-estimator based on ranks. Their precision is investigated through theoretical bounds, comparing their Mean Squared Errors (MSE) to the Cramér-Rao bounds. The numerical implementations and simulations presented in this report were conducted in Python. The complete source code is available on my GitHub repository at the following address.

1.1 Parametric Statistics

Parametric statistics relies on the assumption that observed data follow a well-defined parametric probability distribution. Within this framework, model parameters are estimated using methods such as maximum likelihood or least squares. These estimators are then evaluated in terms of bias, variance, and efficiency, and their performance can be compared against theoretical bounds like the Cramér-Rao Bound (CRB), which provides a lower limit on the variance of an unbiased estimator.
Formally, let there be a parametric density $p(x;\theta)$ with $\theta \in \Theta \subset \mathbb{R}^d$,and an i.i.d. sequence of data $(X_i)_{i=1}^n$. The maximum likelihood estimator is given by:

$$\hat{\theta}_{\mathrm{ML}} = \arg max_{\theta\in\Theta} \sum_{i=1}^{n} \ln p(X_i;\theta)$$

The Cramér-Rao bound is then written as:

$$\mathrm{Var}(\hat{\theta}) \succeq CRB = 1/n \cdot I(\theta)^{-1} \quad \text{avec} \quad I(\theta) = \mathbb{E}\left[\nabla_\theta \ln p(X;\theta) \nabla_\theta \ln p(X;\theta)^\top\right]$$

1.2 Misspecified statistics

In an ideal setting, data strictly adhere to the probability distribution assumed by the model. In practice, however, this assumption is often too strong. For example, one might assume Gaussian noise when, in reality, it follows a t-distribution. This misspecification error, though sometimes unavoidable, impacts the performance of estimators and the validity of classical bounds like the CRB. Misspecified statistics specifically addresses these situations: it allows for the adaptation of estimation and evaluation tools by introducing more realistic bounds, such as the misspecified Cramér-Rao Bound (MCRB), which accounts for the discrepancy between the assumed model and the true data distribution [1].
Formally, if $p^*(x)$ denotes the true data-generating distribution and $p(x;\theta)$ is a misspecified model, then the score vector is generally no longer centered:

$$\mathbb{E}p^*[\nabla\theta \ln p(X;\theta)] \neq 0$$

We define:

$$A(\theta) = \mathbb{E}p^*\left[\nabla\theta^2 \ln p(X;\theta)\right], \quad B(\theta) = \mathbb{E}p^*\left[\nabla\theta \ln p(X;\theta) \nabla_\theta \ln p(X;\theta)^\top\right]$$

The misspecified Cramér-Rao Bound (MCRB) is then given by:

$$\mathrm{Var}(\hat{\theta}) \succeq MCRB = A(\theta)^{-1} B(\theta) A(\theta)^{-1}$$

It always holds that:

$$MCRB \succeq CRB$$

1.3 Maximum Likelihood Estimator (MLE)

The Maximum Likelihood Estimator (MLE) is one of the most fundamental tools in parametric statistics. It relies on the principle of adjusting model parameters to maximize the probability of observing the sampled data. In other words, it involves choosing the parameter θ that makes the observation $(X_1, \ldots, X_n)$ most plausible.
Formally, for a parametric family $(p(x;\theta))_{\theta\in\Theta}$, the maximum likelihood estimator is defined by: :

$$\hat{\theta}_{\mathrm{ML}} \in \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \ln p(X_i;\theta)$$

Under regularity conditions, the MLE is consistent, asymptotically normal, and asymptotically efficient, meaning it achieves the Cramér-Rao bound when the model is well-specified.

1.4 Newton Method

The Newton method is a widely used iterative numerical technique for computing, or at least approximating, estimators such as the maximum likelihood estimator. It relies on a local quadratic approximation of the function to be optimized, using the gradient and Hessian of the log-likelihood function. It aims to solve the same problem as the maximum likelihood estimator. We denote the log-likelihood function as:

$$L(\theta) = \sum_{i=1}^{n} \ell(X_i,\theta) = \sum_{i=1}^{n} \ln p(X_i;\theta).$$

The Newton method then constructs a sequence of iterations defined by:

$$\theta^{(k+1)} = \theta^{(k)} - \left[\nabla^2 L(\theta^{(k)})\right]^{-1} \nabla L(\theta^{(k)}).$$

where $\nabla L(\theta)$ is the gradient and $\nabla^2 L(\theta)$ the hessian of the function $L(\theta)$.
Under regularity conditions and proper initialization, the Newton method converges locally and rapidly to a critical point, often a local maximum or minimum.
However, this method requires the explicit (or approximated) computation of second derivatives, which can be computationally expensive for complex or high-dimensional models. Furthermore, global convergence is not guaranteed, and it may be necessary to use variants (such as the Newton-Raphson method with step size or quasi-Newton methods) to ensure numerical robustness.

1.5 M-estimators

M-estimation (for "maximum" or "minimization") generalizes maximum likelihood by replacing the log-likelihood with an arbitrary function. An M-estimator is then defined as the solution to the following optimization problem:

$$\hat{\theta}_M \in \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \rho(X_i,\theta)$$

where $\rho(x,\theta)$ is a loss (or contrast) function chosen according to the problem. This approach allows for the design of estimators that are more robust to model misspecification or outliers.
Maximum likelihood corresponds to a specific case of M-estimation, where one chooses $\rho(x,\theta) = -\log p(x;\theta)$. Thus, M-estimation provides a unified framework for many estimation methods and plays a crucial role when moving beyond classical assumptions.
Under certain conditions, M-estimators are consistent and asymptotically normal, but their asymptotic variance generally differs from that of MLEs. However, M-estimation relies solely on a criterion defined from observations, without necessarily assuming the existence of a probabilistic model. This makes it more flexible in cases of model misspecification.

1.6 R-estimators

R-estimation (for "rank") is a non-parametric estimation method that leverages the ranks of observations rather than their raw values. This approach grants R-estimation significant robustness to outliers and misspecification errors in the distributional model. Unlike Maximum Likelihood Estimation (MLE), which relies on strict distributional assumptions, or M-estimation, which uses an arbitrary loss function on values, R-estimation is based on the ranks of the residuals, making it particularly effective for non-Gaussian data or when the underlying distribution is unknown.
The fundamental idea behind R-estimation is to minimize an objective function based on the ranks of the residuals, for example:

$$\widehat{\theta}_R \in \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} a(R_i(\theta))\, r_i(\theta)$$

where $R_i(\theta)$ is the rank of the i-th residual $r_i(\theta) = Y_i - f(X_i;\theta)$, and $a(\cdot)$ is a score function that assigns weights to the ranks. Common examples of score functions include the scores of *Wilcoxon* or of *Van der Waerden*.
Under general conditions, R-estimators are consistent and asymptotically normal. Their asymptotic efficiency can be comparable to that of maximum likelihood estimators in certain cases, particularly for heavy-tailed distributions, and they often outperform mean-based estimators in the presence of outliers. R-estimation is therefore an attractive alternative when robustness is a priority and parametric assumptions are difficult to justify.

2 Processus AR(1)

Let $N \in \mathbb{N}$. We define the random vectors $(X_i)_{i\in[0,N]}$, such that $\forall i,\ X_i \in \mathbb{R}$ and

$$\forall n \in \{0,\dots,N-1\}, \quad X_{n+1} = \theta X_n + \varepsilon_{n+1} \tag{1}$$

with $\theta \in \mathbb{R}$ and ε_n a centered noise such that the $(\epsilon_i)_{i\in[0,N]}$ are i.i.d.
We further assume that $X_0 \sim \mathcal{N}(0,1)$.n the following, we only consider $\theta \in]0,1[$ in order to avoid divergences of our process.
Moreover, recursively, each $X_n$can be expressed solely as a function of the noise terms:

$$\forall n, \quad X_n = \theta^n X_0 + \sum_{j=1}^{n} \theta^j \varepsilon_{n-j} \tag{2}$$

2.1 Regarding the noise terms

We will focus on a Gaussian case and a Student case. To justify the comparison of the models, the variances of the noise terms must be equal. In both cases, the $(\epsilon_i)_{i\in[0,N]}$ have zero mean. As for their variance, it is given by:

$$\mathbb{V}[\varepsilon] = \sigma^2 \quad \text{si } \varepsilon \sim \mathcal{N}(0,\sigma^2) \tag{3}$$

$$\mathbb{V}[\varepsilon] = \sigma_\varepsilon^2 = \frac{\nu}{\nu-2}\frac{1}{\mu} \quad \text{si } \varepsilon \sim \mathcal{T}_\nu(0,\frac{1}{\mu}),\ \nu > 2 \tag{4}$$

We thus denote:

$$\frac{1}{\mu} = \sigma^2 = \bar{\sigma^2} \cdot (1-\theta^2) \tag{5}$$

where $\sigma^2$is the variance of the Gaussian noise.

2.2 Gaussian case

We assume that $\forall i, \varepsilon_n \sim \mathcal{N}(0,\ \sigma^2)$. For all $i \in [0, N-1]$, we have $X_{i+1}|X_i \sim \mathcal{N}(\theta X_i,\ \sigma^2)$
By the Markov property:

$$f(X_0, X_1, \ldots, X_n \mid \theta) = f(X_0) \cdot \prod_{i=0}^{N-1} f(X_{i+1} \mid X_i, \theta) \tag{6}$$

We have:

$$X_0 \sim \mathcal{N}(0,1) \quad \Rightarrow \quad \ln f(X_0) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}X_0^2$$

And as $\forall i,\ X_{i+1} \mid X_i \sim \mathcal{N}(\theta X_i, \sigma^2)$, we get:

$$\ln f(X_{i+1} \mid X_i, \theta) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(X_{i+1} - \theta X_i)^2$$

By summing :

$$\ln L(X_0, \ldots, X_n \mid \theta) = \ln f(X_0) + \sum_{i=0}^{N-1} \ln f(X_{i+1} \mid X_i, \theta) \tag{7}$$

Thus :

$$\ln L(\theta) = -\frac{1}{2}\ln(2\pi) - \frac{1}{2}X_0^2 - \frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=0}^{N-1}(X_{i+1} - \theta X_i)^2 \tag{8}$$

$$\frac{d}{d\theta}\ln L(\theta) = \frac{1}{\sigma^2}\sum_{i=0}^{N-1} X_i(X_{i+1} - \theta X_i) \tag{9}$$

$$\frac{d^2}{d\theta^2}\ln L(\theta) = -\frac{1}{\sigma^2}\sum_{i=0}^{N-1} X_i^2 \tag{10}$$

2.3 Estimation of the CRB

According to (2):

$$X_n = \theta^n X_0 + \sum_{j=1}^{n} \theta^j \epsilon_{n-j}$$

Since the ε_k are independent from each other and since $X_0 \sim \mathcal{N}(0,1)$ is independent of the ε_k, we have :

$$\mathbb{E}[X_n] = \theta^n \cdot \mathbb{E}[X_0] = 0 \tag{11}$$

And:

$$\mathbb{V}[X_n] = \sum_{i=0}^{\infty} \theta^{2i} \cdot \sigma^2 = \frac{\sigma^2}{1-\theta^2} \quad \text{because the process is stationary} \tag{12}$$

Thus:

$$\mathbb{E}[X_n^2] = \mathbb{V}[X_n] = \frac{\sigma^2}{1-\theta^2} = \bar{\sigma^2} \tag{13}$$

Finally:

$$CRB(\theta_0) = \frac{-1}{\mathbb{E}\left[\left(\frac{d^2}{d\theta^2}\ln L(\theta)\right)\right]} = \frac{\sigma^2}{\sum_{i=0}^{N-1}\mathbb{E}[X_i^2]} = \frac{1-\theta^2}{N} \tag{14}$$

2.3.1 Estimation of the MCRB

We consider the true distribution of the noise: they are centered t-distributed with a scale parameter $\frac{1}{\mu}$, i.e. $\forall i, \varepsilon_i \sim t_\nu(0, \frac{1}{\mu})$
For the misspecified case, we estimate, at fixed θ and $\sigma^2 = \frac{1}{\mu}$, the quantities:

$$A(\theta) = \mathbb{E}\left[\left(\frac{d^2}{d\theta^2}\ln L(\theta)\right)\right] = -\frac{1}{\sigma^2}\sum_{i=0}^{N-1}\mathbb{E}[X_i^2] \tag{15}$$

$$B(\theta) = \mathbb{E}\left[\left(\frac{d}{d\theta}\ln L(\theta)\right)^2\right] = \frac{1}{\sigma^4}\,\mathbb{E}\left[\left(\sum_{i=0}^{N-1} X_i(X_{i+1} - \theta X_i)\right)^2\right] = \frac{1}{\sigma^4}\,\mathbb{E}\left[\left(\sum_{i=0}^{N-1} X_i\epsilon_i\right)^2\right] \tag{16}$$

We thus calculate:

$$\begin{aligned} A(\theta) &= \frac{-\sigma_\varepsilon^2}{\sigma^2}\cdot\frac{N}{1-\theta^2} \\ &= \frac{\nu}{\nu-2}\cdot\frac{N}{1-\theta^2} \end{aligned} \tag{17}$$

Regarding $B(\theta)$, it can be estimated by the Monte Carlo method, but an explicit expression can be obtained under a fairly realistic assumption:
We assume that $\forall, i \geq j$, ε_i is independent of X_j.
We thus calculate:

$$\mathbb{E}\left[\left(\sum_{i=0}^{N-1} X_i\varepsilon_i\right)^2\right] = \mathbb{E}\left[\sum_{i=0}^{N-1}(X_i\varepsilon_i)^2 + \sum_{i\neq j}(X_i\varepsilon_i)(X_j\varepsilon_j)\right] \tag{18}$$

We have $\mathbb{E}[X_i^2\varepsilon_i^2] = \mathbb{E}[X_i^2]\,\mathbb{E}[\varepsilon_i^2]$, et $\mathrm{Var}(\varepsilon_i) = \sigma_\varepsilon^2 = \frac{\nu}{\nu-2}\cdot\sigma^2$.
Moreover, if $i > j$, then ε_i is independent of X_i, X_j, and ε_j.
So $\mathbb{E}[X_iX_j\varepsilon_j\varepsilon_i] = \mathbb{E}[X_iX_j\varepsilon_j]\,\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\sum_{i\neq j}(X_i\varepsilon_i)(X_j\varepsilon_j)] = 0$
Finally:

$$B(\theta) = \mathbb{E}[\sum_{i=0}^{N-1}(X_i\varepsilon_i)^2] = N\cdot\mathbb{E}[X_i^2]\cdot\mathbb{E}[\varepsilon_i^2] \tag{19}$$

$$= (\frac{\nu}{\nu-2})^2\cdot\frac{N}{1-\theta^2} \tag{20}$$

Then we calculate:

$$MCRB(\theta_0) = \frac{B(\theta_0)}{A(\theta_0)^2} = \frac{1-\theta_0^2}{N} \tag{21}$$

We observe that we always have:

$$MCRB(\theta_0) = CRB(\theta_0) \quad \forall\ \nu \in]2;+\infty[\tag{22}$$

2.3.2 Maximum Likelihood Estimator (MLE)

According to (9):

$$\frac{d}{d\theta}\ln L(\theta)\bigg|_{\theta=\theta_{\mathrm{ML}}} = \frac{1}{\bar{\sigma}^2}\sum_{n=0}^{N-1} X_n\left(X_{n+1} - \theta_{\mathrm{ML}}X_n\right) = 0$$

Therefore:

$$\theta_{\mathrm{ML}} = \frac{\sum_{n=0}^{N-1} X_nX_{n+1}}{\sum_{n=0}^{N-1} X_n^2} \tag{23}$$

2.4 Student's t case

Recall that the noise terms $(\epsilon_i)_i$ follow a centered Student's t distribution withνdegrees of freedom and a scale parameter $\frac{1}{\mu}$. In this case, we have that $X_{i+1} \mid X_i \sim t_\nu(\theta X_i, \frac{1}{\mu})$.

The conditional density is then given by:

$$\ln f(X_{i+1} \mid X_i, \theta) = \frac{\sqrt{\mu} \cdot \Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \mu \frac{(X_{i+1} - \theta X_i)^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{24}$$

The likelihood is thus given, by a calculation similar to that performed in the Gaussian case:

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{X_0^2}{2}\right) \cdot \prod_{i=0}^{N-1} \frac{\sqrt{\mu} \cdot \Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \mu \frac{(X_{i+1} - \theta X_i)^2}{\nu}\right)^{-\frac{\nu+1}{2}} \tag{25}$$

And :

$$\ln L(\theta) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}X_0^2 + N \log\left(\frac{\sqrt{\mu} \cdot \Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\right) - \frac{\nu+1}{2} \sum_{i=0}^{N-1} \log\left(1 + \mu \frac{(X_{i+1} - \theta X_i)^2}{\nu}\right) \tag{26}$$

We now calculate the gradients:

$$\frac{d}{d\theta} \ln L(\theta) = (\nu+1)\frac{\mu}{\nu} \sum_{i=0}^{N-1} \frac{X_i \left(X_{i+1} - \theta X_i\right)}{1 + \frac{\mu}{\nu}\left(X_{i+1} - \theta X_i\right)^2} \tag{27}$$

$$\frac{d^2}{d\theta^2} \ln L(\theta) = -(\nu+1)\frac{\mu}{\nu} \sum_{i=0}^{N-1} X_i^2 \cdot \frac{1 - \frac{\mu}{\nu}\left(X_{i+1} - \theta X_i\right)^2}{\left(1 + \frac{\mu}{\nu}\left(X_{i+1} - \theta X_i\right)^2\right)^2} \tag{28}$$

2.4.1 Estimation of the CRB

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{d^2}{d\theta^2} \ln L(\theta)\right] \tag{29}$$

$$= (\nu+1) \sum_{i=0}^{N-1} \mathbb{E}\left[X_i^2 \cdot \frac{\varepsilon_i^2 - \frac{\mu}{\nu}}{\left(\varepsilon_i^2 + \frac{\mu}{\nu}\right)^2}\right] \tag{30}$$

$$= (\nu+1) \sum_{i=0}^{N-1} \mathbb{E}[X_i^2] \cdot \mathbb{E}\left[\frac{\varepsilon_i^2 - \frac{\mu}{\nu}}{\left(\varepsilon_i^2 + \frac{\mu}{\nu}\right)^2}\right] \tag{31}$$

$$= \frac{N \cdot (v+1)v}{(v-2)\mu} \cdot \frac{\gamma}{1-\theta^2} \quad \text{avec} \quad \gamma = \mathbb{E}\left[\frac{\varepsilon_i^2 - \frac{\mu}{\nu}}{\left(\varepsilon_i^2 + \frac{\mu}{\nu}\right)^2}\right] \tag{32}$$

Thus:

$$CRB(\theta_0) = -\frac{1}{\mathcal{I}(\theta)} = \frac{1-\theta_0^2}{N} \cdot \frac{-(v-2)\mu}{(v+1)v} \cdot \frac{1}{\gamma} \tag{33}$$

Where:

$$\gamma = \int_{-\infty}^{\infty} \frac{x^2 - \frac{\mu}{\nu}}{\left(x^2 + \frac{\mu}{\nu}\right)^2} f_t(x)\,dx \quad \text{avec} \quad f_t(x) = \frac{\sqrt{\mu} \cdot \Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{\mu}{\nu}x^2\right)^{-\frac{\nu+1}{2}} \tag{34}$$

We can also directly calculate the expression of the Cramér-Rao bound, which allows us to have an explicit expression of γ:
The calculations in Appendix B of *Lange et al (1989)* [2] provide a closed-form expression of the Fisher

information matrix for a generalized t-distribution, and by matching the coefficients in the $CRB(\theta)$, we obtain:

$$\gamma = \frac{-\mu}{\nu + 3} \tag{35}$$

and

$$CRB(\theta_0) = \frac{1 - \theta_0^2}{N} \cdot \frac{(v-2)(v+3)}{(v+1)v} \tag{36}$$

2.4.2 Approximation of the maximum likelihood estimator by the Newton method

By using (27):

$$\frac{d}{d\theta} \ln L(\theta)\bigg|_{\theta=\theta_{\mathrm{ML}}} = (\nu+1)\frac{\mu}{\nu} \sum_{i=0}^{N-1} \frac{X_i \left(X_{i+1} - \theta_{\mathrm{ML}} X_i\right)}{1 + \frac{\mu}{\nu}\left(X_{i+1} - \theta_{\mathrm{ML}} X_i\right)^2}$$

We cannot compute this value numerically, so we will estimate θ_{ML} using the Newton method:

$$\theta^{(k+1)} = \theta^{(k)} - \frac{\frac{d}{d\theta} \ln L(\theta^{(k)})}{\frac{d^2}{d\theta^2} \ln L(\theta^{(k)})} \quad avec \quad \theta^{(0)} = \frac{\sum_{i=0}^{N-1} X_i X_{i+1}}{\sum_{i=0}^{N-1} X_i^2} \tag{37}$$

the value that would have been obtained via the maximum likelihood estimator for Gaussian data.

2.5 M-estimators

For our first-order autoregressive process, the use of the maximum likelihood estimator is compromised when in practice the distribution of the (ε_i) is unknown, and M-estimation is a good alternative. This has been studied by *Martin and Jong (1977)* [3] for the first-order autoregressive process. Indeed, if $|\theta| < 1$, the M-estimator of θ, denoted θ_M, is defined by the minimization problem:

$$\min_{\theta} \sum_{i=1}^{N} L(X_i - \theta X_{i-1}), \tag{38}$$

where $L : \mathbb{R} \to \mathbb{R}$ is a real-valued, even, and differentiable function, with derivative $\psi = L'$.
Si ψ est bornée, vérifie $t\psi(t) > 0$ pour tout $t \neq 0$ et $\psi'(0) = 1$, alors ce problème est équivalent à la résolution de :

$$\sum_{i=1}^{N} X_{i-1} \psi(X_i - \theta X_{i-1}) = 0. \tag{39}$$

The M-estimator θ_M is particularly robust when $\psi = \psi_H$, where ψ_H is the Huber function (1981), defined by :

$$\psi_H(t) = \begin{cases} t & \text{si } |t| \leq c, \\ c \cdot \operatorname{sgn}(t) & \text{si } |t| > c, \end{cases} \tag{40}$$

where $c > 0$ is a positive constant.
We can also choose $\psi = \psi_T$, where ψ_T is Tukey's biweight function, defined by:

$$\psi_T(t) = \begin{cases} t \cdot \left(1 - \left(\frac{t}{c}\right)^2\right)^2 & \text{si } |t| \leq c, \\ 0 & \text{si } |t| > c. \end{cases} \tag{41}$$

We denote by $\theta_{M,H}$ and $\theta_{M,T}$ the M-estimators associated respectively with the Huber and Tukey loss functions.

he Huber function combines the properties of least squares estimators (for small values of t) and least absolute deviations (for big values of t). It is simple, continuous, and provides good robustness while maintaining good asymptotic efficiency under a normal distribution. However, it is less efficient in the presence of highly influential extreme data.
The Tukey function, on the other hand, is even more robust, as it completely nullifies the influence of outliers, thus preventing their impact. However, it may cause convergence issues because it is non-convex [4].

2.6 R-estimators

A process $(X_t)_{t\in\mathbb{Z}}$ is an ARMA(p,q) (autoregressive moving-average) model if it satisfies the following equation:

$$X_t = \sum_{i=1}^{p} A_i X_{t-i} + \varepsilon_t + \sum_{j=1}^{q} B_j \varepsilon_{t-j}, \quad t \in \mathbb{Z} \tag{42}$$

Where :

- ε_t is a centered white noise, often assumed i.i.d. with constant variance σ^2;
- $p \geq 0$ is the order of the autoregressive (AR) part;
- $q \geq 0$ is the order of the moving average (MA) part;
- $A_1, \ldots, A_p$ are the AR coefficients;
- $B_1, \ldots, B_q$ are the MA coefficients.

We denote this model as ARMA(p, q), and we observe that our AR(1) model is in particular an ARMA$(1, 0)$.

J.Allal, A.Kaaouach and D. Paindaveine(2001) [5] studied R-estimation for ARMA processes. By rewriting the ARMA(p, q) model in this way:

$$A(L)X_t = B(L)\varepsilon_t, \quad t \in \mathbb{Z} \tag{43}$$

where L is the lag operator, $A(L) := 1 - \sum_{i=1}^{p} A_i L^i$ and $B(L) := 1 + \sum_{i=1}^{q} B_i L^i$.
The parameter to estimate is

$$\theta_0 = (A_1, \ldots, A_p, B_1, \ldots, B_q) \in \mathbb{R}^{p+q}$$

Let $X^{(n)} := (X_1^{(n)}, \ldots, X_n^{(n)})$ be an observed series of length n, and let $H^{(n)}(\theta_0)$ be the hypothesis under which $X^{(n)}$ is generated by model (42). We denote by $R_t^{(n)}(\theta_0)$ the rank of the residual $Z_t^{(n)}(\theta_0)$ among the set $Z_1^{(n)}(\theta_0), \ldots, Z_n^{(n)}(\theta_0)$, where:

$$Z_t^{(n)}(\theta_0) := \frac{A(L)}{B(L)} X_t^{(n)}, \quad t = 1, \ldots, n \tag{44}$$

We then consider the autocorrelation coefficient of order k :

$$r_k^{(n)}(\theta_0) := \frac{1}{n-k} \sum_{t=k+1}^{n} \left[J_1\left(\frac{R_t^{(n)}(\theta_0)}{n+1}\right) J_2\left(\frac{R_{t-k}^{(n)}(\theta_0)}{n+1}\right) - m^{(n)} \right] \Big/ \sigma_k^{(n)} \tag{45}$$

where J_1 et J_2 are score functions, and $m^{(n)}$, $\sigma_k^{(n)}$ are normalization constants (their exact value is specified in [6]) such that $r_k^{(n)}(\theta_0)$ is centered and standardized under the hypothesis $H^{(n)}(\theta_0)$.
We then define the vector of rank statistics:

$$\sqrt{n}\, T_{J_1,J_2}^{(n)}(\theta_0) := \begin{pmatrix} \sum_{k=1}^{n-1} \sqrt{n-k}\, \psi_k^{(1)}(\theta_0)\, r_k^{(n)}(\theta_0) \\ \vdots \\ \sum_{k=1}^{n-1} \sqrt{n-k}\, \psi_k^{(p+q)}(\theta_0)\, r_k^{(n)}(\theta_0) \end{pmatrix} \tag{46}$$

With:

$$\left\{\psi_t^{(1)},\dots,\psi_t^{(p+q)}\right\}_{t\in\mathbb{Z}} \text{ a system of solutions of } A(L)B(L)\psi_t = 0,\ t\in\mathbb{Z}. \tag{47}$$

We then set:

$$\Delta^{(n)}_{J_1,J_2}(\theta_0) := \sqrt{n}\, M(\theta_0)^T \left(C_\psi(\theta_0)^T\right)^{-1} T^{(n)}_{J_1,J_2}(\theta_0) \tag{48}$$

and we define the R-estimator as:

$$\hat{\theta}^{(n)} := \arg\min_\theta \left\|\Delta^{(n)}_{J_1,J_2}(\theta)\right\| \tag{49}$$

where $\|\cdot\|$ is any norm on $\mathbb{R}^{p+q}$.
The matrix $M(\theta_0)$ is defined by:

$$M(\theta_0) := \begin{pmatrix} 1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ g_1 & 1 & \cdots & & h_1 & 1 & \cdots & \\ \vdots & & \ddots & & \vdots & & \ddots & \\ g_{p-1} & \cdots & \cdots & 1 & h_{q-1} & \cdots & \cdots & 1 \\ g_p & \cdots & \cdots & g_1 & h_q & \cdots & \cdots & h_1 \\ \vdots & & & \vdots & \vdots & & & \vdots \\ g_{p+q-1} & \cdots & \cdots & g_q & h_{p+q-1} & \cdots & \cdots & h_p \end{pmatrix} \tag{50}$$

Finally, $C_\psi(\theta_0)$ and $W^2_\psi(\theta_0)$ are the matrices whose elements are respectively: $\psi_i^{(j)}$ and $\sum_{t=1}^{\infty}\psi_t^{(i)}\psi_t^{(j)}$
In our case, there is a single parameter to estimate $\theta_0 = A_1$ and we denote by $R_t^{(n)}(\theta_0)$ the rank of the residual $\varepsilon_t(\theta_0)$ among the set $\{\varepsilon_1(\theta_0),\dots,\varepsilon_n(\theta_0)\}$
The AR(1) case is also much simpler since we are now in dimension 1 ($p = 1$ and $q = 0$, hence $p+q = 1$), and the matrices thus become scalars:

$$M(\theta_0) = 1 \tag{51}$$

$$C_\psi(\theta_0) = \psi_1^{(1)} = \psi_1 \tag{52}$$

Now, according to (47):

$$A(L)B(L)\psi_t = 0$$

With :

$$A(L) = 1 - A_1 L = 1 - \theta_0 L \tag{53}$$

$$B(L) = 1 \tag{54}$$

We then obtain the recurrence relation:

$$\psi_t = \theta_0\, \psi_{t-1} \tag{55}$$

And by taking the convention $\psi_0 = \alpha \in \mathbb{R}$, we have :

$$\psi_t = \alpha \cdot \theta_0^t \tag{56}$$

Hence:

$$\psi_1 = \alpha\cdot\theta_0 \quad \Rightarrow \quad C_\psi(\theta_0) = \alpha\cdot\theta_0 \tag{57}$$

Finally:

$$\Delta^{(n)}_{J_1,J_2}(\theta_0) = \frac{1}{\alpha \cdot \theta_0} \cdot \sum_{k=1}^{n-1} \sqrt{n-k}\, \psi^{(1)}_k(\theta_0)\, r^{(n)}_k(\theta_0) \tag{58}$$

$$= \sum_{k=1}^{n-1} \sqrt{n-k} \cdot \theta_0^{k-1} \cdot r^{(n)}_k(\theta_0) \tag{59}$$

And the R-estimator is then:

$$\theta_R := \arg\min_{\theta} \sum_{k=1}^{N-1} \sqrt{N-k} \cdot \theta^{k-1} \cdot r^{(N)}_k(\theta) \tag{60}$$

We can also introduce an asymptotic R-estimator $\hat{\theta}^{(n)}_{\mathrm{R}}$ following the Hájek–Le Cam approach, which satisfies:

$$\Delta^{(n)}_{J_1,J_2}\left(\hat{\theta}^{(n)}\right) = o_p(1). \tag{61}$$

It is written as:

$$\hat{\theta}^{(n)}_{\mathrm{R}} = \tilde{\theta}^{(n)} + \frac{1}{\sqrt{n}}\, \hat{c}^{-1}(J_1, J_2, g)\, \Gamma^{-1}(\tilde{\theta}^{(n)})\, \Delta^{(n)}_{J_1,J_2}(\tilde{\theta}^{(n)}), \tag{62}$$

with $\tilde{\theta}^{(n)}$ a preliminary $\sqrt{n}$-consistent estimator of θ_0, for example the maximum likelihood estimator θ_{ML}, and $\hat{c}^{-1}(J_1, J_2, g)$ an estimator of $c^{-1}(J_1, J_2, g)$, a constant depending on the noise density g of ε_n and the score functions J_1 and J_2. Its exact value (given in [5]) is not important to us because in our study we consider not knowing the noise density. However, another explicit value of $c(J_1, J_2, g)$ can be calculated via [7]:

$$c^{-1}(J_1, J_2, g)\, \Gamma^{-1}(\theta) = \mathcal{I}^{-1}(\theta) = n \cdot CRB(\theta) \tag{63}$$

Thus:

$$c(J_1, J_2, g) = \frac{1}{n} \cdot \Gamma^{-1}(\theta) \cdot CRB^{-1}(\theta) \tag{64}$$

With :

$$\Gamma(\theta_0) := M(\theta_0)^T\, C_\psi^{\top -1}(\theta_0)\, W^2_\psi(\theta_0)\, C_\psi^{-1}(\theta_0)\, M(\theta_0), \tag{65}$$

And in our case :

$$\Gamma(\theta_0) = \frac{W^2_\psi(\theta_0)}{C_\psi^{-2}(\theta_0)} \quad \text{avec} \quad W^2_\psi(\theta_0) = \sum_{t=1}^{\infty} \psi^{(1)}_t(\theta_0)^2 = \frac{\alpha^2 \cdot \theta_0^2}{1-\theta_0^2}. \tag{66}$$

Hence:

$$\Gamma(\theta_0) = \frac{1}{1-\theta_0^2} \tag{67}$$

Regarding the estimation of $\hat{c}(J_1, J_2, g)$, we use the fact that, under hypotheses A.1 and A.2 formulated in [5], and for any $c > 0$, under $\mathcal{H}^{(n)}(\theta_0)$, as $n \to \infty$:

$$\sup_{\|\tau^{(n)}\| \le c} \left\| \Delta^{(n)}_{J_1,J_2}\left(\theta_0 + \frac{1}{\sqrt{n}}\tau^{(n)}\right) - \Delta^{(n)}_{J_1,J_2}(\theta_0) + c(J_1, J_2, g)\, \Gamma(\theta_0)\, \tau^{(n)} \right\| = o_p(1). \tag{68}$$

Hence, for sufficiently big n, we have:

$$\hat{c}(J_1, J_2, g) = \frac{|\Delta^{(n)}_{J_1,J_2}\left(\theta_0 + \frac{1}{\sqrt{n}}\tau^{(n)}\right) - \Delta^{(n)}_{J_1,J_2}(\theta_0)|}{\Gamma(\theta_0)\, |\tau^{(n)}|} \tag{69}$$

$$= \frac{|\Delta^{(n)}_{J_1,J_2}\left(\theta_0 + \frac{1}{\sqrt{n}}\tau^{(n)}\right) - \Delta^{(n)}_{J_1,J_2}(\theta_0)|}{|\tau^{(n)}|} \cdot (1-\theta^2) \tag{70}$$

And our estimator is:

$$\hat{\theta}_{\mathrm{RO}} = \tilde{\theta} + \frac{|\tau|}{\sqrt{N}} \cdot \frac{\Delta^{(N)}_{J_1,J_2}(\tilde{\theta})}{|\Delta^{(N)}_{J_1,J_2}\left(\theta_0 + \frac{1}{\sqrt{N}}\tau\right) - \Delta^{(N)}_{J_1,J_2}(\theta_0)|} \tag{71}$$

3 Simulations and comparison

3.1 Cramér-Rao Bounds

Figure 1 compares the true Cramér-Rao bound (CRB) calculated assuming correctly that the noise terms are t-distributed, with the true CRB calculated assuming the noise terms are Gaussian. The misspecified Cramér-Rao bound (MCRB) being equal to the Gaussian CRB, the curves overlap.
The two curves converge towards a similar value as $\nu \to +\infty$, which is consistent: a Student's t-distribution with many degrees of freedom tends towards a normal distribution. Therefore, in this case, assuming Gaussian noise is a good approximation.
For low values of ν, the Student CRB is much larger than the Gaussian CRB. This shows that the Gaussian approximation is poor. Furthermore, the Student CRB exhibits great instability for values of ν close to 2, with visible peaks. This reflects the sensitivity of the Gaussian model to the violation of its assumptions when the noise is not truly normal.
Moreover, a logarithmic scale is used to highlight the order-of-magnitude differences between the CRB and the MCRB.

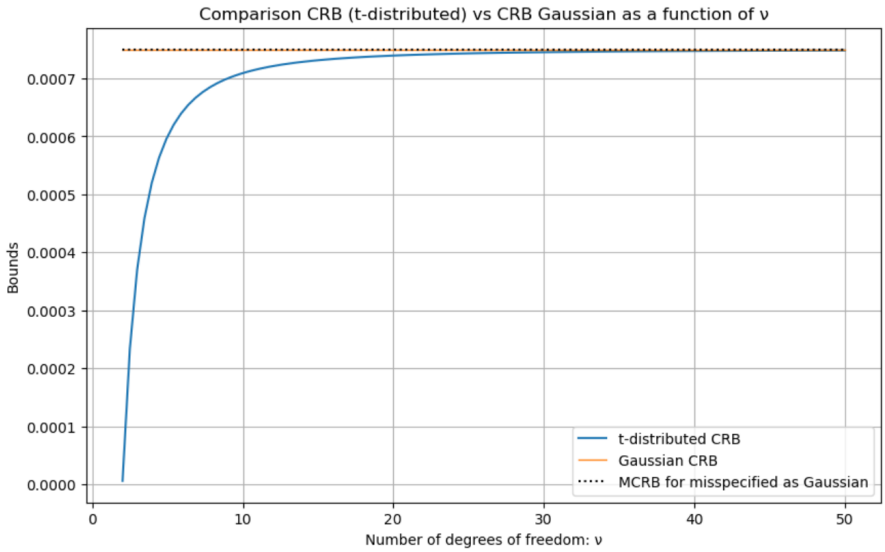

Figure 1: Comparison of Student CRB and Gaussian CRB as a Function of the Degrees of Freedom (Parameters: $N = 1000, \bar{\sigma}^2 = 1, \theta = 0.5$)

We show the same curve (Figure 2) for degrees of freedom $\nu > 15$ in order to use a normal scale, and we highlight the difference relative to $\nu = 50$. This difference of $1{,}76 \cdot 10^{-6}$ confirms the fact that $CRB_{\mathrm{Student}} \underset{\nu \to +\infty}{\longrightarrow} CRB_{\mathrm{Gaussian}}$.

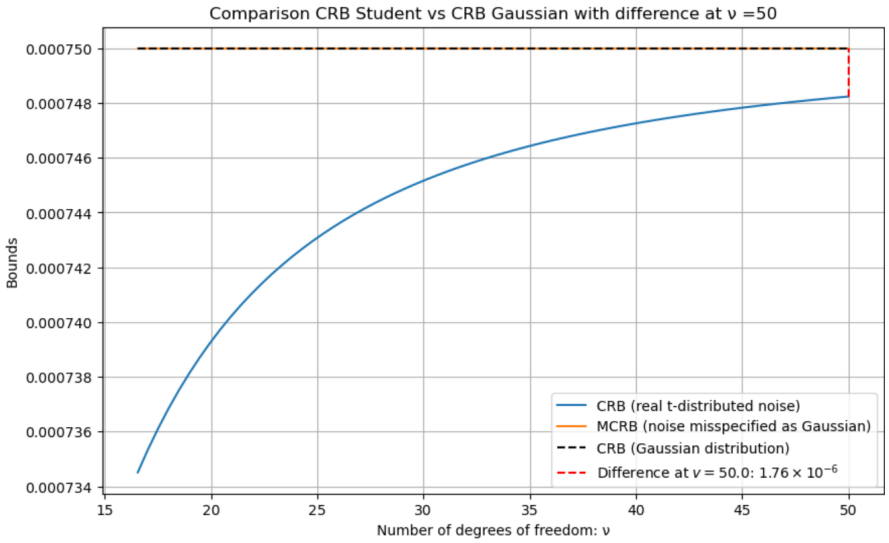

Figure 2: Focus on $\nu > 15$ (Parameters: $N = 1000, \bar{\sigma}^2 = 1, \theta = 0.5$)

3.2 Evaluation of estimators

To compare the different estimators and evaluate the performance of each model and verify which estimator is the best, we will study the mean squared errors (MSE) for the maximum likelihood estimator in the Gaussian case and the estimator obtained from the Newton method in the Student case. We will also compare these two estimators to the best estimator derived from M-estimation, whose performance varies both according to the choice of the loss function and the constant c, as well as the one-step R-estimator.
To compare them, we use the following formula, after generating multiple estimates of θ which we denote, for the maximum likelihood estimator, the Newton method estimator, the M-estimator, and the R-estimator respectively, $(\theta_{ML}^{(i)})_i$, $(\theta_{NT}^{(i)})_i$, $(\theta_{M}^{(i)})_i$ and $(\theta_{R}^{(i)})_i$:

$$MSE(\theta) = \frac{1}{N}\sum_{i=1}^{N}(\theta^{(i)} - \theta)^2 \quad with \quad \forall i, \quad \theta^{(i)} = \theta_{ML}^{(i)} \ , \ \theta_{NT}^{(i)} \ , \ \theta_{M}^{(i)} \ or \ \theta_{R}^{(i)} \tag{72}$$

We therefore compare the performance of the estimators via their mean squared error as a function of the number of observations N, the parameter θ, and the degrees of freedom ν.

3.2.1 Comparison of the M-estimators

Figure 3 plots $M = 1000$ Monte Carlo simulations of the MSE calculation for the M-estimators $\theta_{M,H}$ and $\theta_{M,T}$ as a function of $c \in$, with their respective minima shown as dashed lines:

$min\ MSE(\theta_{M,H}) = 0.00039$, for $c = 0.798$
$min\ MSE(\theta_{M,T}) = 0.00053$, for $c = 4.672$

OWe therefore observe that the M-estimator based on the Huber loss function is the best, and this result is explained by the nature of the simulated data, which follow a Student's t-distribution with a moderate number of degrees of freedom ($\nu = 3$).
Indeed, the Student's t-distribution tends towards a normal distribution as the degrees of freedom increase, which means the generated data are close to a Gaussian distribution. This closeness to normality makes the Huber loss function particularly well-suited, since it combines quadratic sensitivity to small residuals (like the maximum likelihood estimation under the Gaussian assumption) with linear robustness against outliers.

Therefore, the Huber function achieves an effective compromise between efficiency and robustness, which explains its good performance in this context where the data are 'quasi-normal' but may still contain some outliers. Moreover, the M-estimator based on Tukey's method is very sensitive to initialization, and it is therefore common to obtain poorer performance.
However, we observe that depending on the degrees of freedom ν, the best estimator varies considerably. One possible explanation is that the maximum likelihood estimator provides a good starting point, and the M-estimator derived from Tukey's loss function therefore remains effective.
In the following, we will denote by θ_M the M-estimator $\theta_{M,H}$.

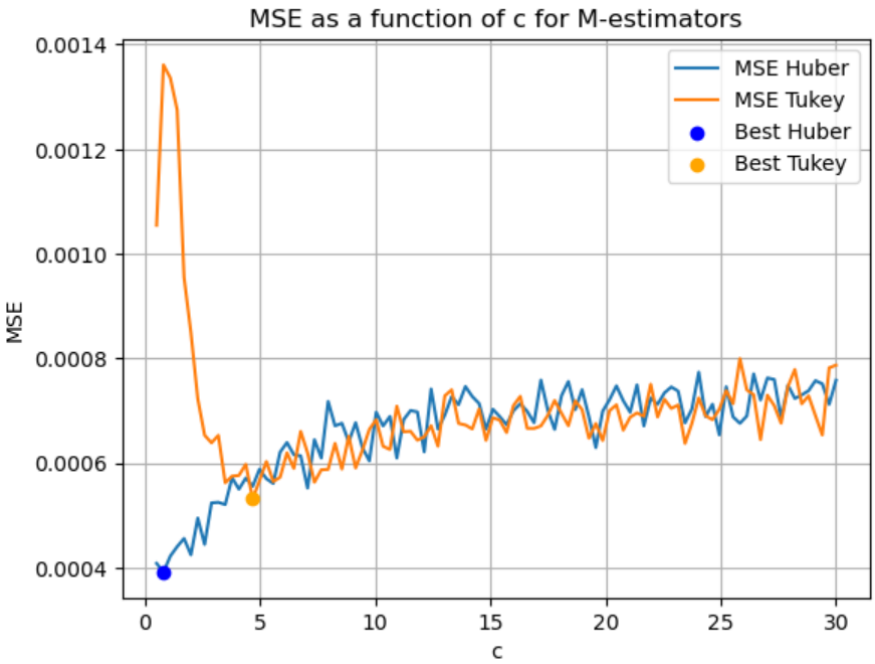

Figure 3: MSE of the M-estimators as a function of c (Parameters: $\bar{\sigma}^2 = 1, \theta = 0.5, \nu = 3$)

Degrees of freedom	Best c Huber	MSE Huber	Best c Tukey	MSE Tukey	Best loss function
2.00	1.48	1.57e-04	4.10	2.12e-04	Huber
5.11	2.79	5.21e-04	5.41	5.77e-04	Huber
8.22	4.10	5.72e-04	5.41	5.26e-04	Tukey
11.33	5.09	5.23e-04	5.74	5.83e-04	Huber
14.44	6.40	5.92e-04	3.78	6.11e-04	Huber
17.56	3.45	5.42e-04	3.45	6.16e-04	Huber
20.67	7.71	6.22e-04	7.05	6.06e-04	Tukey
23.78	7.38	5.77e-04	8.36	5.63e-04	Tukey
26.89	7.38	5.97e-04	6.07	5.54e-04	Tukey
30.00	3.45	6.33e-04	6.40	5.86e-04	Tukey

Table 1: Best c and MSE for the Huber and Tukey estimators as a function of the degrees of freedom v

3.2.2 Comparison of the R-estimators

Regarding the R-estimators, the chosen score functions J_1 and J_2 are the Van Der Waerden scores:

$$J_1(t) = J_2(t) = \phi(t)^{-1} \tag{73}$$

where ϕ is the cumulative distribution function of a standard normal distribution.
The computation time of the R-estimators is very long due to their high complexity. To distinguish between them, we note that the one-step R-estimator has clear asymptotic properties and that its computation time is slightly lower than the other. Moreover, as the simulations in [5] show, their respective MSEs are very close. The simulation was indeed performed, as in their work, with:

$$(\varepsilon_n)_{n\in[0,N]} \sim \mathcal{N}(0,1), \theta = 0.8, N = 500 \text{ et } M = 300 \text{ Monte Carlo repetitions} \tag{74}$$

We obtain:

$$MSE(\theta_R) = 750 \cdot 10^{-6} \tag{75}$$

$$MSE(\theta_{R,one-step}) = 845 \cdot 10^{-6} \tag{76}$$

Whereas they obtained:

$$MSE(\theta_{R'}) = 735 \cdot 10^{-6} \tag{77}$$

$$MSE(\theta_{R,one-step'}) = 752 \cdot 10^{-6} \tag{78}$$

For all the reasons mentioned above, the simulations were performed with the one-step R-estimator $\theta_{R,one-step}$, which we denote more simply as θ_R.

3.2.3 Global comparison

All the MSE curves were plotted with a number of Monte Carlo simulations $M = 100$ except for the one in Figure 5, where $M = 10000$.

We compare the performance of the estimators via their mean squared error as a function of the number of observations N which varies between 100 and 10000 (Figure 4).
We observe that the MSE decreases as N increases; the larger the sample size, the more precise the estimators are, according to the law of large numbers.
The estimator θ_{NT} obtained from the Newton method has performance similar to that of the M-estimator θ_M, and their performances are slightly better than that of the maximum likelihood estimator θ_{ML} from the misspecified case. This is expected because the MSEs converge toward their respective Cramér-Rao bounds, with the MSE of the M-estimator converging to the Student CRB, as in the well-specified case. Indeed, for M-estimation, the actual observations are considered and no misspecification is made.

When varying the degrees of freedom ν, we observe (Figure 5) that the MSE values are highly variable, with many visible peaks. The estimators θ_{ML} and θ_{NT} tend to the same value because when $\nu \to +\infty$, $CRB_{\text{Student}} \underset{\nu\to+\infty}{\longrightarrow} CRB_{\text{Gaussian}} = \frac{1-\theta^2}{N}$
On the other hand, θ_M tends toward a slightly higher value, indicating poorer performance. Indeed, we fixed $c = 0.798$ in the Huber loss function, but this value is optimal only for $\nu = 3$, and it is not optimal for varying ν, as indicated in Table 1. Hence the poorer performance.

When θ varies (Figure 6), we also observe that the MSEs decrease and converge towards their respective Cramér-Rao bounds.

Finally, we observe that even though the mean error tends to zero for all estimators, the estimator obtained from the Newton method and the M-estimator consistently remain slightly more performant across the range of sample sizes N and parameter values θ: the misspecification is justified, but the increasing amount of data brings robustness against it.

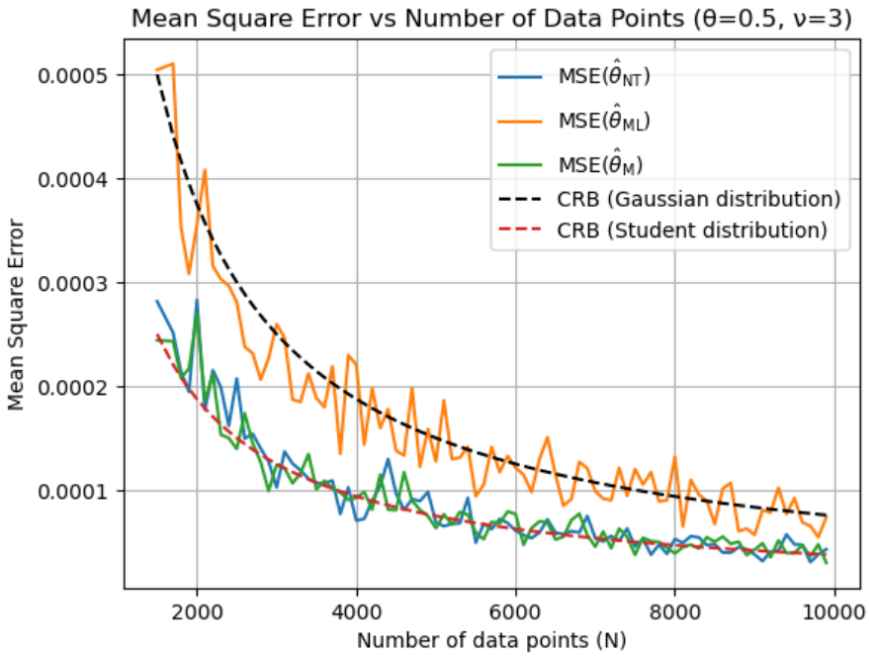

Figure 4: MSE as a function of N (Parameters: $\bar{\sigma}^2 = 1, \theta = 0.5, \nu = 3$)

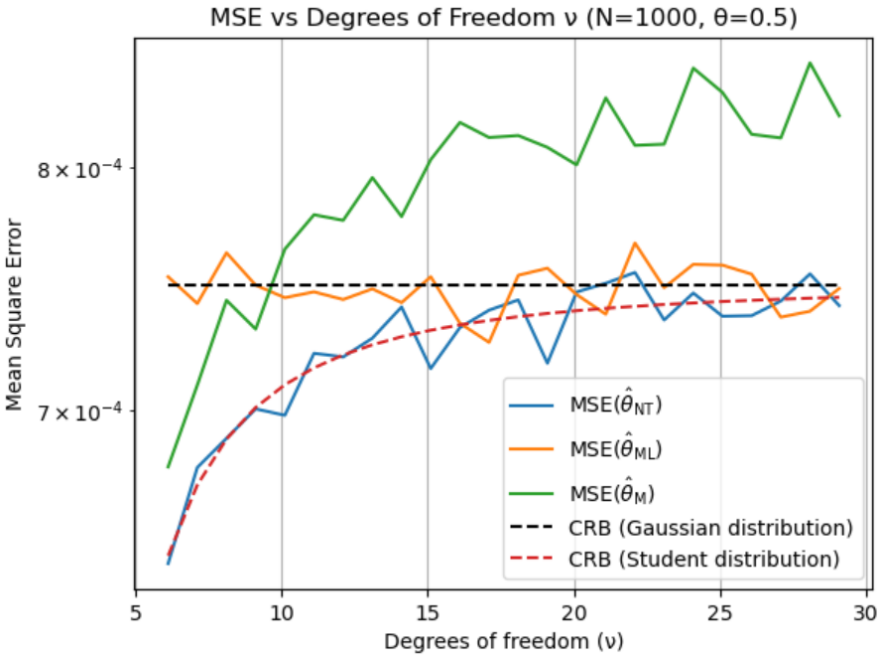

Figure 5: MSE as a function of ν (Parameters: $N = 1000, \bar{\sigma}^2 = 1, \theta = 0.5$)

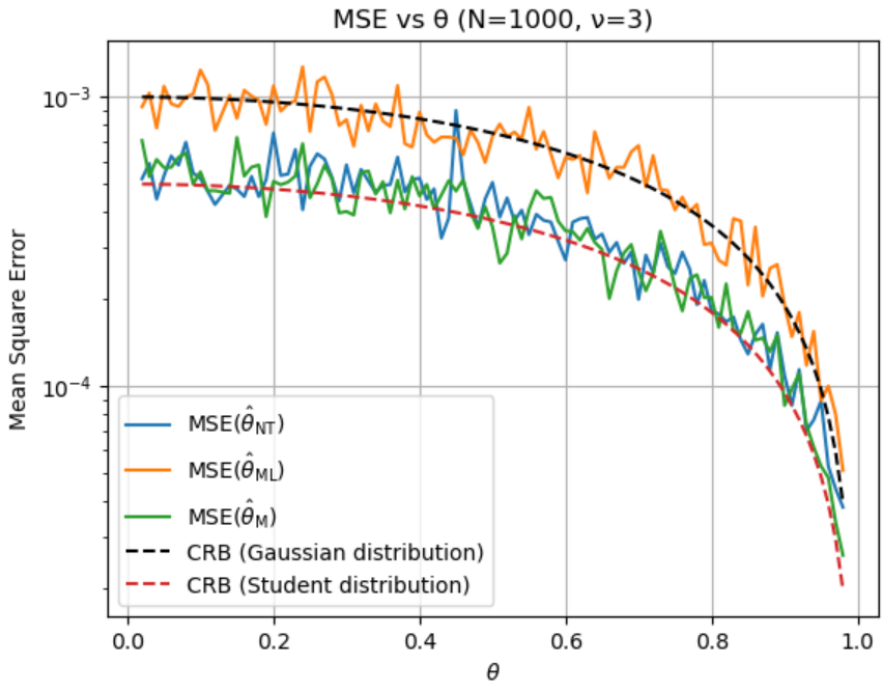

Figure 6: MSE as a function of θ (Parameters: $N = 1000, \bar{\sigma}^2 = 1, \nu = 3$)

In addition to the previous figures, we compared the performance of the one-step R-estimator in Tables 2 to 4. We did not graphically represent the MSE curve of this estimator because its simulation cost is particularly high, due to the necessity of re-estimating the entire set of ranks at each iteration.

N	MSE(θ_R)	MSE(θ_M)	MSE(θ_{NT})	MSE(θ_{ML})
2500	$2.540 \cdot 10^{-4}$	$1.579 \cdot 10^{-4}$	$1.713 \cdot 10^{-4}$	$1.950 \cdot 10^{-4}$
5000	$1.065 \cdot 10^{-4}$	$8.412 \cdot 10^{-5}$	$8.627 \cdot 10^{-5}$	$1.983 \cdot 10^{-4}$
10000	$5.173 \cdot 10^{-5}$	$3.304 \cdot 10^{-5}$	$3.561 \cdot 10^{-5}$	$7.790 \cdot 10^{-5}$

Table 2: MSE of the estimators as a function of N (Parameters: $\bar{\sigma}^2 = 1, \theta = 0.5, \nu = 3, M = 100$)

ν	MSE(θ_R)	MSE(θ_M)	MSE(θ_{NT})	MSE(θ_{ML})
3	$2.574 \cdot 10^{-3}$	$4.067 \cdot 10^{-4}$	$4.238 \cdot 10^{-4}$	$7.515 \cdot 10^{-4}$
10	$7.908 \cdot 10^{-4}$	$7.463 \cdot 10^{-4}$	$6.911 \cdot 10^{-4}$	$7.623 \cdot 10^{-4}$
29.1	$9.874 \cdot 10^{-4}$	$8.520 \cdot 10^{-4}$	$7.542 \cdot 10^{-4}$	$7.744 \cdot 10^{-4}$

Table 3: MSE of the estimators as a function of ν (Parameters: $N = 1000, \bar{\sigma}^2 = 1, \theta = 0.5, M = 5000$)

θ	MSE(θ_R)	MSE(θ_M)	MSE(θ_{NT})	MSE(θ_{ML})
0.1	$6.101 \cdot 10^{-3}$	$5.595 \cdot 10^{-4}$	$6.097 \cdot 10^{-4}$	$8.091 \cdot 10^{-4}$
0.5	$6.477 \cdot 10^{-4}$	$3.905 \cdot 10^{-4}$	$5.113 \cdot 10^{-4}$	$6.904 \cdot 10^{-4}$
0.9	$1.684 \cdot 10^{-4}$	$1.219 \cdot 10^{-4}$	$1.416 \cdot 10^{-4}$	$2.209 \cdot 10^{-4}$

Table 4: MSE of the estimators as a function of θ (Parameters: $N = 1000, \bar{\sigma}^2 = 1, \nu = 3, M = 100$)

We observe that the one-step R-estimator θ_R often has similar performance, but never better than that of the M-estimator θ_M and the Newton estimator θ_{NT}, although it outperforms the maximum likelihood estimator θ_{ML}. This is indeed due to the fact that in our simulations, we used $N = 1000$, which is too small for our approximation in (70) to be accurate, and thus our estimation of $c(J_1, J_2, g)$ is wrong. Moreover, for reasons of computational complexity, the estimator $\hat{c}(J_1, J_2, g)$ of $c(J_1, J_2, g)$

was constructed using only a single Monte Carlo simulation, resulting in a large variance (even with 10000 Monte Carlo simulations to construct $\hat{c}(J_1, J_2, g)$, one obtains a variance on the order of 0.3 for a value of $c(J_1, J_2, g) < 1$).
In summary, since our estimation of $c(J_1, J_2, g)$ is noisy, our R-estimator is also noisy and therefore does not provide better performance than the M-estimator.
Moreover, the R-estimator has a very large error for $\nu = 3$ and $\theta = 0.1$ (Tables 3 and 4), showing that its performance suffers when the tails are heavy and the process is highly stationary. The first error is related to the fact that the score functions J_1 and J_2 chosen for the estimation of $c(J_1, J_2, g)$ are optimal for the R-estimation of AR(1) processes with Gaussian noise; however, our noise has degrees of freedom $\nu = 3$, making it far from Gaussian, which explains the large error, whereas for large ν the performance is good.

3.2.4 Focus on the R-estimator

The rather disappointing performance of the R-estimator with respect to ν might lead us to doubt it, and the following simulations aim to verify the asymptotic properties of the R-estimator and to show that it is as efficient as, or even more efficient than, the M-estimator. Its previous performance issues were solely due to an inaccurate estimation of $c(J_1, J_2, g)$ (itself caused by insufficient computational power).
We therefore perform new simulations of the so-called 'exact' one-step R-estimator, denoted $\theta_{R,o}$, in the sense that we use the true value of $c(J_1, J_2, g)$ calculated in (64).

N	MSE($\theta_{R,o}$)	MSE(θ_M)	MSE(θ_{NT})	EQM(θ_{ML})
2500	$2.279 \cdot 10^{-4}$	$1.556 \cdot 10^{-4}$	$1.330 \cdot 10^{-4}$	$2.925 \cdot 10^{-4}$
5000	$1.081 \cdot 10^{-4}$	$7.328 \cdot 10^{-5}$	$9.619 \cdot 10^{-5}$	$1.444 \cdot 10^{-4}$
10000	$6.154 \cdot 10^{-5}$	$4.259 \cdot 10^{-5}$	$4.167 \cdot 10^{-5}$	$6.457 \cdot 10^{-5}$

Table 5: MSE of the estimators as a function of N (Parameters: $\bar{\sigma}^2 = 1,\ \theta = 0.5,\ \nu = 3,\ M = 100$)

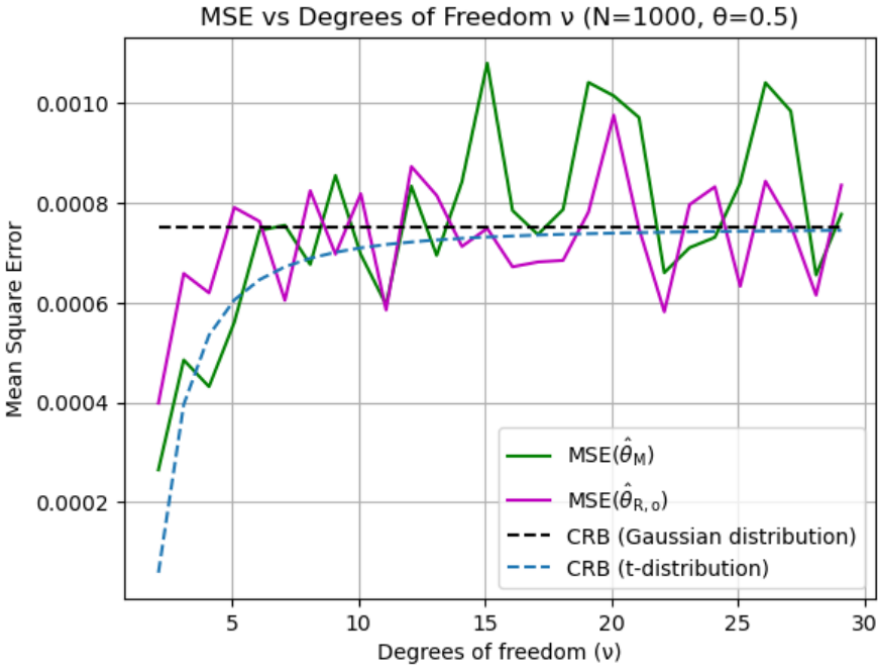

Figure 7: MSE of θ_M and $\theta_{R,o}$ as a function of ν (Parameters: $N = 1000, \bar{\sigma}^2 = 1, \theta = 0.5$)

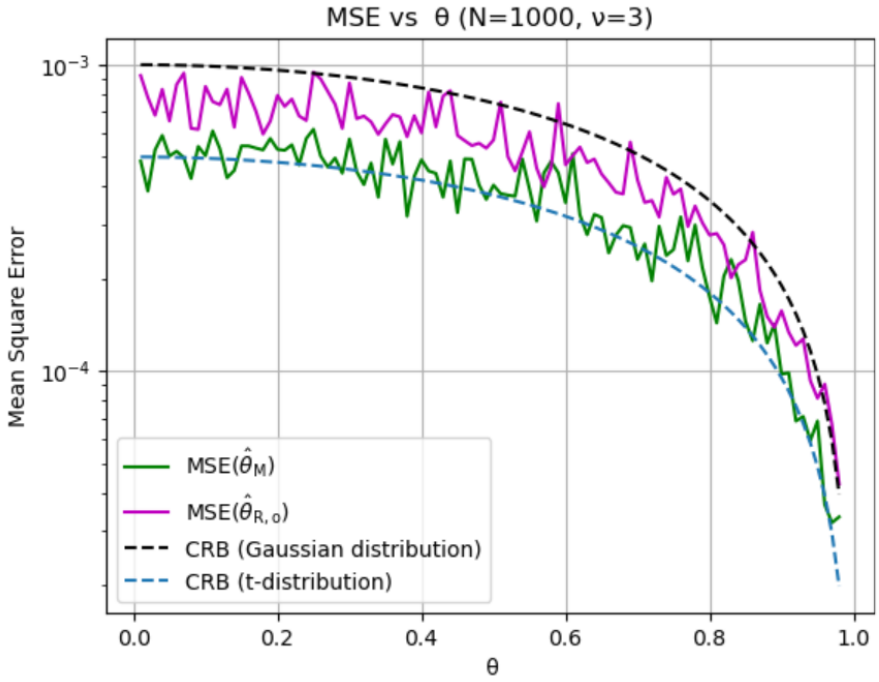

Figure 8: MSE of θ_M and $\theta_{R,o}$ as a function of θ (Parameters: $N = 1000, \bar{\sigma}^2 = 1, \nu = 3$)

Our results are indeed much better, with good results for variable N (Table 5): the MSE of $\theta_{R,o}$ is always close to those of θ_M and θ_{ML}, while being slightly higher than that of θ_{ML}. Figure 7 confirms our hypothesis of poor performance of $\theta_{R,o}$ for heavy tails due to a poor estimation of $c(J_1, J_2, g)$: here, we considered the "true" $c(J_1, J_2, g)$, and the R-estimator is always better than the M-estimator. This is explained, as previously mentioned, by the fact that the efficiency of θ_M depends on the value of c, which in turn depends on ν (Table 1). Figure 8 also attests to the good performance of the R-estimator, which is slightly less good than the M-estimator, but here we no longer observe MSE values higher than those of θ_{ML} for small values of θ.

References

[1] S. Fortunati, F. Gini, M. S. Greco, and C. D. Richmond, "Performance bounds for parameter estimation under misspecified models: Fundamental findings and applications," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 142–157, 2017.

[2] K. L. Lange, R. J. A. Little, and J. M. G. Taylor, "Robust statistical modeling using the t distribution," *Journal of the American Statistical Association*, vol. 84, no. 408, pp. 881–896, 1989.

[3] R. Martin and J. De Jong, "Asymptotic properties of robust generalized m-estimates for the first-order autoregressive parameter," technical memo, Bell Laboratories, Murray Hill, New Jersey, 1977.

[4] L. Denby and R. D. Martin, "Robust estimation of the first-order autoregressive parameter," *Journal of the American Statistical Association*, vol. 74, no. 365, pp. 140–146, 1979.

[5] J. Allal, A. Kaaouach, and D. Paindaveine, "R-estimation for arma models," *Journal of Nonparametric Statistics*, vol. 13, no. 6, pp. 815–831, 2001.

[6] M. Hallin and M. Puri, "Aligned rank tests for linear models with autocorrelated error terms," *Journal of Multivariate Analysis*, vol. 50, no. 2, pp. 175–237, 1994.

[7] L. Le Cam, "On the asymptotic theory of estimation and testing hypotheses," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* (J. Neyman, ed.), vol. 1, pp. 129–156, Berkeley, CA: University of California Press, 1956.