UPPSALA
UNIVERSITET

U.U.D.M. Project Report 2019:28

Anomaly Detection in Seasonal ARIMA Models

Filip Örneholm

Examensarbete i matematik, 30 hp
Handledare: Rolf Larsson
Examinator: Denis Gaidashev
Juni 2019

Department of Mathematics
Uppsala University

Anomaly Detection in Seasonal ARIMA Models

Filip Örneholm

June 2019

Contents

Abstract

Anomaly detection is a broad subject pertaining to the identification of unexpected (or anomalous) patterns in data. When the data is in a time series format, a clear relationship to the field of change point detection can be found. Drawing inspiration from both change point detection and time series modelling, we employ a variation of the SR procedure on generated SARIMA time series. We found indications of usability, although a more thorough analysis of the proposed methods would be needed for a decisive conclusion.

1 Introduction

1.1 Anomaly Detection

Anomaly detection is a broad subject pertaining to the identification of unexpected (or anomalous) patterns in data. The type of data can range from spatial data in medical imaging and camera surveillance to sequential data in network security and temporal sensor readings. The distinction between what is considered normal data and anomalous data is likewise domain specific. Due to this fact, there is of course no universal anomaly detection technique. "Anomaly Detection" is therefore considered an umbrella term for a type of problem rather than anything else[1]. In this thesis, we will look at sequentially sampled time series data with the aim of detecting if and when it starts to deviate from what is expected. In particular, we will draw inspiration from the statistical area of *change point detection* to investigate the possibility of detecting changes in simulated *SARIMA* time series.

There are of course many articles discussing particular cases of anomaly detection, and oftentimes it comes down to finding *what* works in a given domain rather than *why*. While this is a reasonable approach in applied mathematics, especially when the goal is to provide some real life utility, it is also what motivates a theoretic dive into this subject. However, the aims of this thesis are two fold; (1) to bring forward and motivate usable statistics in a theoretical framework, and (2) to empirically try these statistics on the simulated data.
For an extensive overview of anomaly detection application domains and techniques, the reader is referred to [1]. However, it is by no means a prerequisite read for this thesis.

1.2 Generating the Data

One clear advantage of generating data is that it comes with the privilege of complete control over the test environment, which is extremely helpful in the context of establishing statistical theory. We will be generating one-dimesional time series data (denoted $x_1, x_2, ...$) with evenly spaced discrete time. Specifically, time series which aligns with the assumptions of the *Seasonal Autoregressive Integrated Moving Average* (SARIMA, def. 1.6) model. However, since the SARIMA model is an extension of other time series models, some definitions need to be covered before getting into it. We begin with notaion for the *backshift operator*, which will be used extensively.

Definition 1.1. The backshift operator is defined by

$$Bx_t = x_{t-1}.$$

We also extend it to powers and functions such that

$$B^k x_t = x_{t-k}$$

and

$$B^k f(t) = f(t-k).$$

1.2.1 Autoregressive (AR) and Moving Average (MA) models

Let w_t be a *white noise process* with variance σ^2, i.e. a series of random variables $w_1, w_2, ...$ which are iid $N(0, \sigma^2)$.

Definition 1.2. An **autoregressive model** of order p, abbreviated as $\mathbf{AR}(p)$, is of the form

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + w_t$$

where x_t is stationary and $\phi_1, \phi_2, ..., \phi_p$ are constants ($\phi_p \neq 0$). Equivalently, one can write

$$(1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p) x_t = w_t$$

or

$$\phi(B) x_t = w_t$$

Figure 1 (a) depicts an example of AR(1).

Definition 1.3. A **moving average model** of order q, abbreviated as $\mathbf{MA}(q)$, is of the form

$$x_t = w_t + \theta_1 w_{t-1} + \theta_2 w_{t-2} + ... + \theta_q w_{t-q}$$

where x_t is stationary and $\theta_1, \theta_2, ..., \theta_p$ are constants ($\theta_q \neq 0$). Equivalently, one can write

$$(1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q) w_t = x_t$$

or

$$\theta(B) w_t = x_t$$

Figure 1 (b) depicts an example of MA(1). If x_t follows an AR(p) model, it is a linear combination of the previous p observations $x_{t-p}, ..., x_{t_1}$ plus some random noise w_t. If x_t instead follows an MA(q) model, it is a linear combination of the previous q observations' random noises $w_{t-q}, ..., w_{t-1}$ plus it's own random noise w_t. Combining these simple models, we define the *ARMA(p,q)* model.

Definition 1.4. An **autoregressive moving average model** of order p, q abbreviated as $\mathbf{ARMA}(p, q)$, is of the form

$$x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q}$$

where x_t is stationary and $\phi_1, ..., \phi_p, \theta_1, ..., \theta_q$ are constants ($\phi_q, \theta_p \neq 0$). Equivalently, one can write

$$(1 - \phi_1 B - ... - \phi_p B^p) x_t = (1 + \theta_1 B + ... + \theta_q B^q) w_t$$

or

$$\phi(B) x_t = \theta(B) w_t$$

Figure 1 (c) depicts an example of ARMA(1,1). One drawback of using the ARMA model on real data is the assumption it makes about stationarity, i.e. that the mean is constant and the covariance between two points in the time series only depends on the distance between them. To tackle this, we can make use of differencing:

Definition 1.5. A process x_t is an **autoregressive integrated moving average model** of order p, d, q, abbreviated as **ARIMA**(p, d, q), if

$$\nabla^d x_t = (1 - B)^d x_t$$

is **ARMA**(p, q). Equivalently, one can write

$$\phi(B)(1 - B)^d x_t = \theta(B) w_t$$

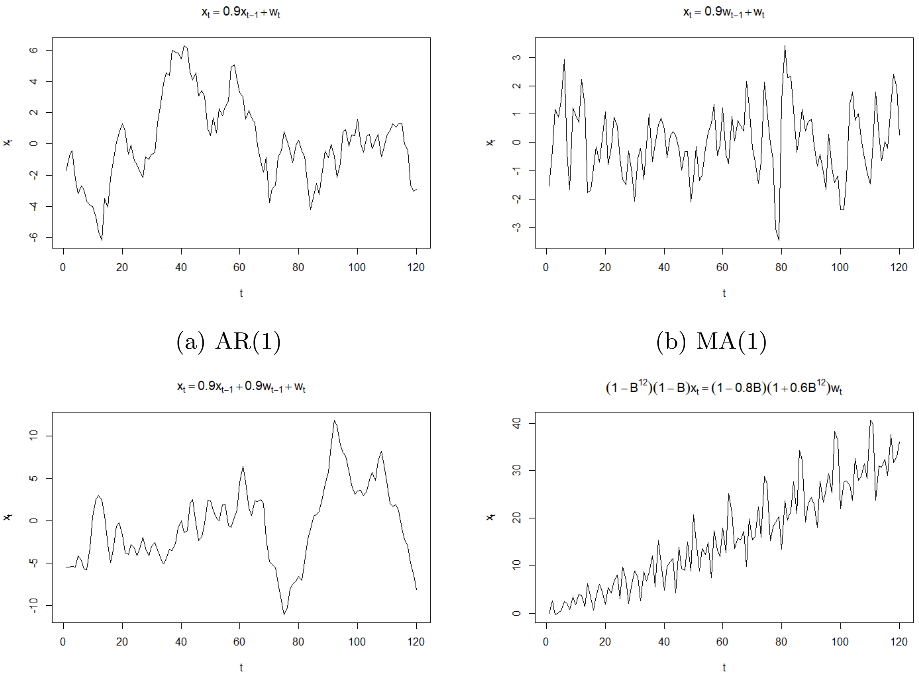
Oftentimes, real time series data contain some kind of seasonality. Whether it be the yearly trends of global temperature, or the vibrations of a machine during it's work cycle, an ARIMA model might still not cut it. We arrive at the final extension of the model.

Definition 1.6. The multiplicative **seasonal autoregressive integrated moving average** model, or **SARIMA** model is given by

$$\Phi_P(B^s)\phi(B)\nabla_s^D \nabla^d x_t = \Theta_Q(B^s)\theta(B) w_t$$

The general model is denoted as **SARIMA**$(p, d, q) \times (P, D, Q)_s$. The ordinary autoregressive and moving average components are represented by polynomials $\phi(B)$ and $\theta(B)$ of orders p and q, the seasonal autoregressive and moving average components by $\phi_P(B^s)$ and $\theta_Q(B^s)$ of orders P and Q, and ordinary and seasonal difference components by $\nabla^d = (1 - B)^d$ and $\nabla_s^D = (1 - B^s)^D$.

Figure 1 (a) depicts an example of SARIMA(0,1,1)$\times(0, 1, 1)_{12}$.



(a) AR(1)

(b) MA(1)

(c) ARMA(1,1)

(d) SARIMA(0,1,1)$\times$(0,1,1)$_{12}$

Figure 1: Four simulated processes

2 The Shiryaev-Roberts Statistic

Introduced by Shiryaev in 1961, the *Shiryaev-Roberts* (SR) procedure is a method of detecting change in the drift of a Brownian motion [4]. This procedure has since then been appropriated to accommodate so-called *change point problems*, i.e.problems of detecting a change in the state of a process. In a time series setting, observations are obtained one at a time. If the observations are in line with what is expected, the process may continue. If the state changes at some point, one wants to raise an alarm/stop the procedure as soon as possible. However, depending on the context, false alarms (i.e. alarms raised at a pre-change time) are ideally kept to a minimum. Thus, we can consider this an multi-objective optimization problem. Common characteristics of a detection policy are the *Average Run Length to False Alarm* (ARL2FA) and the *Average Delay to Detection* (AD2D). Let $X_1, X_2, ...$ be the series of random variables observed, and let $X_1, ..., X_v$ be iid f_0 and $X_{v+1}, ..$ iid f_1. f_0 and f_1 are assumed to be known and are, respectively, called the pre-change and post-change distributions. Furthermore, let $\mathbb{P}_k$ and $\mathbb{E}_k$ denote the probability and expectation when $v = k$. In the case where no change occurs ($v = \infty$), write $\mathbb{P}_\infty$ and $\mathbb{E}_\infty$. Then, the SR procedure calls for raising an alarm at time

$$N_A = \min\{n \geq 1 : R_n \geq A\},$$

where

$$R_n = \sum_{k=1}^{n} \frac{p(X_1, ..., X_n|v = k)}{p(X_1, ..., X_n|v = \infty)} = \sum_{k=1}^{n} \prod_{i=k}^{n} \frac{f_1(X_i)}{f_0(X_i)},$$

and A is chosen such that $\mathbb{E}_\infty N_A = S$, i.e. such that the expected stopping/alarm time is at least S, when no actual change ever occurs.

When v has a geometric prior distribution $\mathbb{P}(v = k) = \rho(1 - \rho)^{k-1}$ $(k \geq 1, \rho \in (0, 1))$, it has been proven [3] that the SR procedure, as described above, minimizes the integral AD2D $= \sum_{k=1}^{\infty} \mathbb{E}_k(N - k)^+$. Using a geometric prior is fitting in cases when we assume that an event has some probability ρ of happening every time we make a new observation. The optimality properties of the SR procedure are, as with all detection procedures, highly dependent on the prior assumptions of the system which is observed. If particular knowledge about, for instance, the rate and time-dependency of the occurrence of anomalies is available, this information should naturally be incorporated in the detection procedure. [6]

3 SARIMA models with additive anomaly

Consider the following problem. We sequentially observe a SARIMA time series $x_1, x_2...$ with known parameters (i.e. $p, d, q, P, D, Q, s, \sigma^2$, $\phi(\cdot), \theta(\cdot)$, $\Phi_P(\cdot), \Theta_Q(\cdot)$). However, the time series we observe is affected by an *anomaly function* $g_v(t)$ at some time $v \in \mathbb{N}_{>0} \cup \infty$. Similarly to the change point problem presented in section 2, we want to detect v as soon as possible while keeping the false positives to a minimum. We begin with examining time series with *additive anomalies.*

Definition 3.1. The **additive anomaly function** is defined as

$$g_v^+(t) = \begin{cases} 0 & , t < v \\ g^+(t + 1 - v) & , t \geq v \end{cases}$$

where $g^+(t)$ is some known real-valued function and $g^+(1) \neq 0$.

In which case, we observe the time series

$$y_t = x_t + g_v^+(t) \tag{1}$$

We have, by def 1.6,

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \Theta_Q(B^s)\theta(B)w_t \tag{2}$$

$$\Leftrightarrow \Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d(y_t - g_v^+(t)) = \Theta_Q(B^s)\theta(B)w_t \tag{3}$$

$$\Leftrightarrow \Theta_Q(B^s)\theta(B)w_t + \Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d g_v^+(t) = \Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d y_t \tag{4}$$

Note that the r.h.s. of (4) is completely decided by $y_1, ..., y_t$. Hence, we can obtain observations of the random series

$$Z_t = \Theta_Q(B^s)\theta(B)w_t + \Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d g_v^+(t).$$

Since $(Z_t|v = k)$ is a linear combination of the normals $w_{t-(q+sQ)}, .., w_t$ plus the non-random expression

$$G_k(t) = \Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d g_k^+(t),$$

$(Z_t|v = k)$ is also normal distributed. Let $N_k(n) \sim (Z_1, ..., Z_n|v = k)$. We see that the covariances are not dependent on k:

$$\begin{aligned} Cov(Z_i, Z_j|v = k) &= E[(Z_i - G_k(i))(Z_j - G_k(j))] \\ &= E[(\Theta_Q(B^s)\theta(B)w_i)(\Theta_Q(B^s)\theta(B)w_j)] =: \mathbf{C}_{i,j} \end{aligned}$$

and that $N_k(n)$ is a multivariate normal, hence it's density is

$$\begin{aligned} f_{N_k(n)}(z_1, ..., z_n) &= \frac{1}{\sqrt{(2\pi)^n|\mathbf{C}|}} \exp\Bigg(- \frac{(\mathbf{z} - E[\mathbf{z}])^T\mathbf{C}^{-1}(\mathbf{z} - E[\mathbf{z}])}{2}\Bigg) \\ &= \frac{1}{\sqrt{(2\pi)^n|\mathbf{C}|}} \exp\Bigg(- \frac{(\mathbf{z} - \mathbf{G_k})^T\mathbf{C}^{-1}(\mathbf{z} - \mathbf{G_k})}{2}\Bigg), \end{aligned}$$

where $\mathbf{z} = (z_1, ..., z_n)$ and $\mathbf{G_k} = (G_k(1), ..., G_k(n))$. Note that $\mathbf{G}_\infty = \mathbf{0}$ and we find the SR statistic

$$\begin{aligned} R_n &= \sum_{k=1}^n \frac{f_{N_k(n)}(z_1, ..., z_n)}{f_{N_\infty(n)}(z_1, ..., z_n)} \\ &= \sum_{k=1}^n \exp\Bigg(\frac{1}{2}\big((\mathbf{z} - \mathbf{G}_\infty)^T\mathbf{C}^{-1}(\mathbf{z} - \mathbf{G}_\infty) - (\mathbf{z} - \mathbf{G_k})^T\mathbf{C}^{-1}(\mathbf{z} - \mathbf{G_k})\big)\Bigg) \qquad (5) \\ &= \sum_{k=1}^n \exp\Bigg(\frac{1}{2}\big(\mathbf{z}^T\mathbf{C}^{-1}\mathbf{G_k} + \mathbf{G_k}^T\mathbf{C}^{-1}\mathbf{z} - \mathbf{G_k}^T\mathbf{C}^{-1}\mathbf{G_k}\big)\Bigg) \end{aligned}$$

3.1 SR procedure on AR(p)

Applying the SR procedure on an AR(p) with an additive anomaly is very similar to the problem described in section 2. We have the AR(p) process x_t, the anomaly function $g_v^+(t)$ and observed time series y_t:

$$\begin{aligned} x_t &= \phi_1 x_{t-1} + ... + \phi_p x_{t-p} \\ y_t &= x_t + g_v^+(t) \\ \Rightarrow w_t + G_v(t) &= y_t - \phi_1 y_{t-1} - ... - \phi_p y_{t-p}. \end{aligned}$$

Thus, $\{Z_t \sim N(G_k(i), \sigma^2)\}_{t=1,2...}$ are mutually independent. Let $N_k(n) = (Z_1, ..., Z_n | v = k)$, we arrive at the SR statistic

$$\begin{aligned} R_n &= \sum_{k=1}^{n} \frac{f_{N_k(n)}(z_1, ..., z_n)}{f_{N_\infty(n)}(z_1, ..., z_n)} = \sum_{k=1}^{n} \prod_{i=1}^{n} \frac{f_{N(G_k(i),\sigma^2)}(z_i)}{f_{N(0,\sigma^2)}(z_i)} \\ &= \sum_{k=1}^{n} \prod_{i=k}^{n} \frac{f_{N(G_k(i),\sigma^2)}(z_i)}{f_{N(0,\sigma^2)}(z_i)} = \sum_{k=1}^{n} \prod_{i=k}^{n} \exp \frac{z_i^2 - (z_i - G_k(i))^2}{2\sigma^2} \end{aligned}$$

This procedure is exhibited by simulating an AR(3) process with $(\phi_1, \phi_2, \phi_3) = (0.5, 0.2, 0.15)$:

$$\begin{aligned} \phi(B)x_t &= w_t \\ \Leftrightarrow x_t &= 0.5x_{t-1} + 0.2x_{t-2} + 0.15x_{t-3} + w_t \end{aligned} \tag{6}$$

and $w_1, w_2, ...$ iid $N(0, 1)$. We add the anomaly

$$g_{300}^{+}(t) = \begin{cases} 0 & , t < 300 \\ \sin(\frac{\pi(t+1-300)}{100}) & , 300 \leq t < 400 \\ 0 & , t \geq 400 \end{cases}$$

We observe the following time series:

$$y_t = x_t + g_v^{+}(t) \tag{7}$$

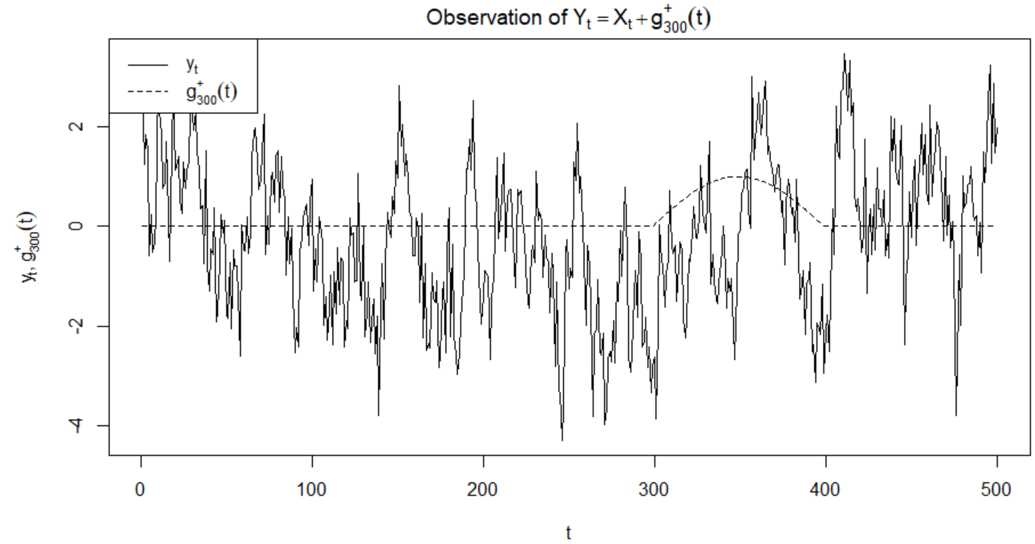


Figure 2: y_t and $g_{300}^{+}(t)$

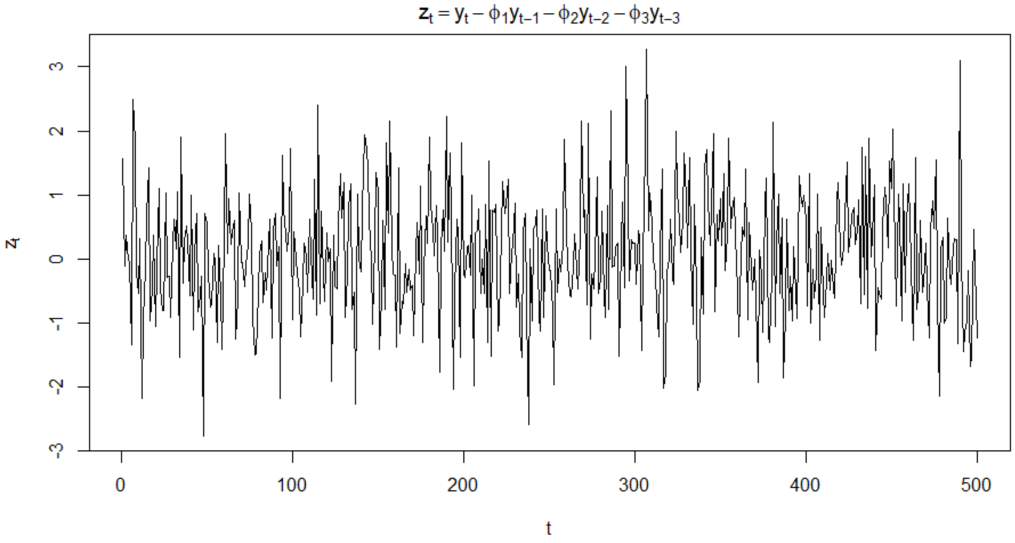


Figure 3: z_t

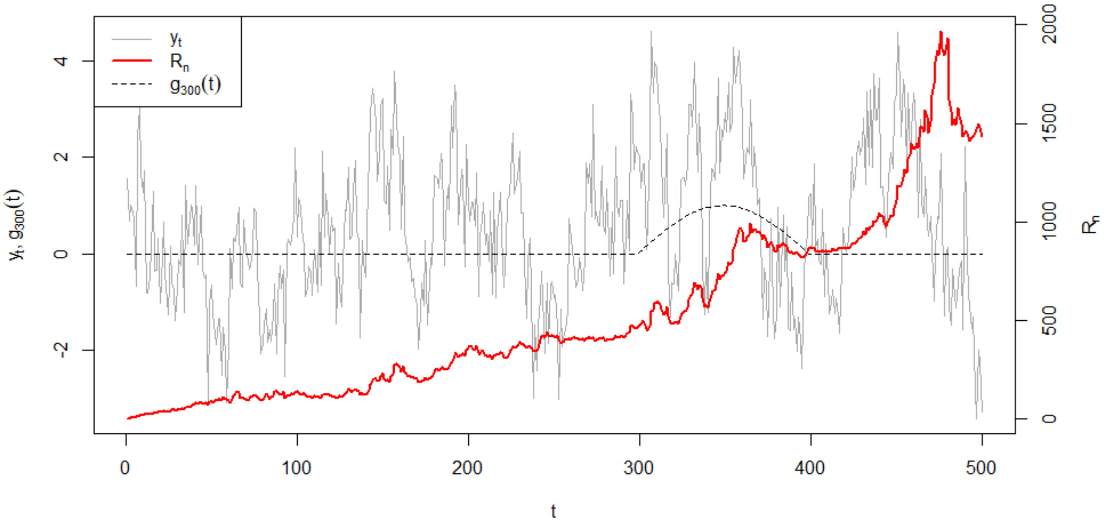


Figure 4: y_t, $g^+_{300}(t)$ and the SR statistic (R_n)

3.2 SR procedure on ARMA(p,q)

In this example we have an ARMA(p,q) with AR and MA coefficients $\phi_1, ..., \phi_p$ and $\theta_1, ..., \theta_q$:

$$\begin{aligned} &\phi(B)x_t = \theta(B)w_t \\ &\quad\Leftrightarrow x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q} \end{aligned} \tag{8}$$

and $w_1, w_2, ...$ iid $N(0, \sigma^2)$. We sequentially observe the time series

$$y_t = x_t + g_v^+(t) \tag{9}$$

where $g_v^+(t)$ is an additive anomaly function. Combining (8) and (9), we get

$$\begin{aligned} w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q} + &g_v(t) - \phi_1 g_v(t-1) - ... - \phi_p g_v(t-p) \\ =&y_t - \phi_1 y_{t-1} - ... - \phi_p y_{t-p}. \end{aligned} \tag{10}$$

So, from the r.h.s. of (10) we have $z_t = y_t - \phi_1 y_{t-1} - ... - \phi_p y_{t-p}$ as our observation of Z_t. Let $G_v(t) = g_v(t) - \phi_1 g_v(t-1) - ... - \phi_p g_v(t-p)$. Then, we conclude that from the l.h.s. of (10) that

$$Z_t \sim N(G_v(t), \sigma^2(1 + \theta_1^2 + ... + \theta_q^2)).$$

As before, let $N_k(n) \sim (Z_1, ..., Z_n | v = k)$. We have the covariance matrix $\mathbf{C}$, such that

$$\begin{aligned} \mathbf{C}_{i,j} &= E[(\theta(B)w_i)(\theta(B)w_j)] \\ &= E[(w_i + \theta_1 w_{i-1} + ... + \theta_q w_{i-q})(w_j + \theta_1 w_{j-1} + ... + \theta_q w_{j-q})] \end{aligned}$$

Since the covariance is symmetrical, and $i \neq j \Rightarrow E[w_i, w_j] = 0$, we let $j = i + h$, $h \geq 0$, and see that

$$\begin{aligned} \mathbf{C}_{i,i+h} =&E[(w_i + \theta_1 w_{i-1} + ... + \theta_q w_{i-q})(w_{i+h} + \theta_1 w_{i+h-1} + ... + \theta_q w_{i+h-q}) \\ =&\begin{cases} \sigma^2 \sum_{l=0}^{q-h} \theta_l \theta_{l+h} & , 0 \leq h \leq q \\ 0 & , h > q \end{cases} \end{aligned} \tag{11}$$

Now that we have $\mathbf{C}$ and $\mathbf{G_k}$, we can use (5) to calculate the SR statistic.

To illustrate this process, we simulate an ARMA(3,2) process with AR and MA components $(\phi_1, \phi_2, \phi_3) = (0.3, 0.2, 0.15)$ and $(\theta_1, \theta_2) = (0.4, 0.2)$:

$$\begin{aligned} &\phi(B)x_t = \theta(B)w_t \\ \Leftrightarrow x_t = 0.3x_{t-1} + 0.2x_{t-2} + &0.15x_{t-3} + w_t + 0.4w_{t-1} + 0.2w_{t-2} \end{aligned} \tag{12}$$

and $w_1, w_2, ...$ iid $N(0, 1)$. We add the anomaly $g = \sigma_w^2 = 1$ to this series at point $v = 150$, i.e.

$$g_{150}^+(t) = \begin{cases} 0 & , t < 150 \\ 1 & , t \geq 150, \end{cases}$$

and observe the following time series:

$$y_t = x_t + g_v^+(t) \tag{13}$$

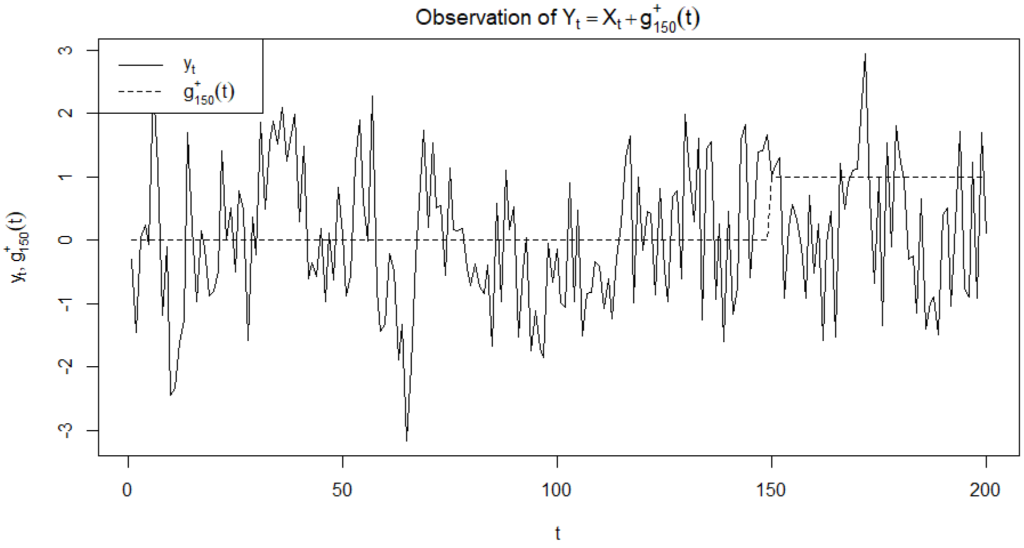


Figure 5: y_t and $g^+_{150}(t)$

Combining (12) and (13), we get

$$\begin{aligned} w_t + 0.4w_{t-1} + 0.2w_{t-2} + & g_v - 0.3g_v(t-1) - 0.2g_v(t-2) - 0.15g_v(t-3) \\ = & y_t - 0.3y_{t-1} - 0.2y_{t-2} - 0.15y_{t-3}. \end{aligned} \tag{14}$$

So, from the r.h.s. of (14) we $z_t = y_t - 0.3y_{t-1} - 0.2y_{t-2} - 0.15y_{t-3}$ as our observation of Z_t. We have $G_v(t) = g_v(t) - 0.3g_v(t-1) - 0.2g_v(t-2) - 0.15g_v(t-3)$ and conclude, from the l.h.s. of (14), that

$$\begin{aligned} Z_t \sim & N(G_v(t), \sigma^2(1 + \theta_1^2 + \theta_3^2 + \theta_2^2)) \\ = & N(G_v(t), 1 + 0.4^2 + 0.2^2) = N(G_v(t), 1.2) \end{aligned}$$

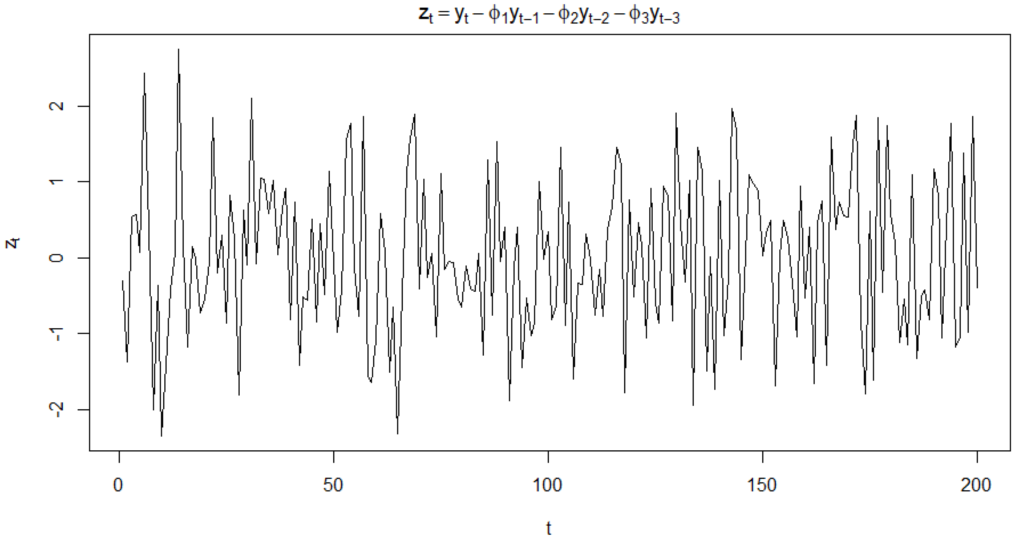


Figure 6: z_t

The SR statistic, R_n, can then be calculated using formula (5)

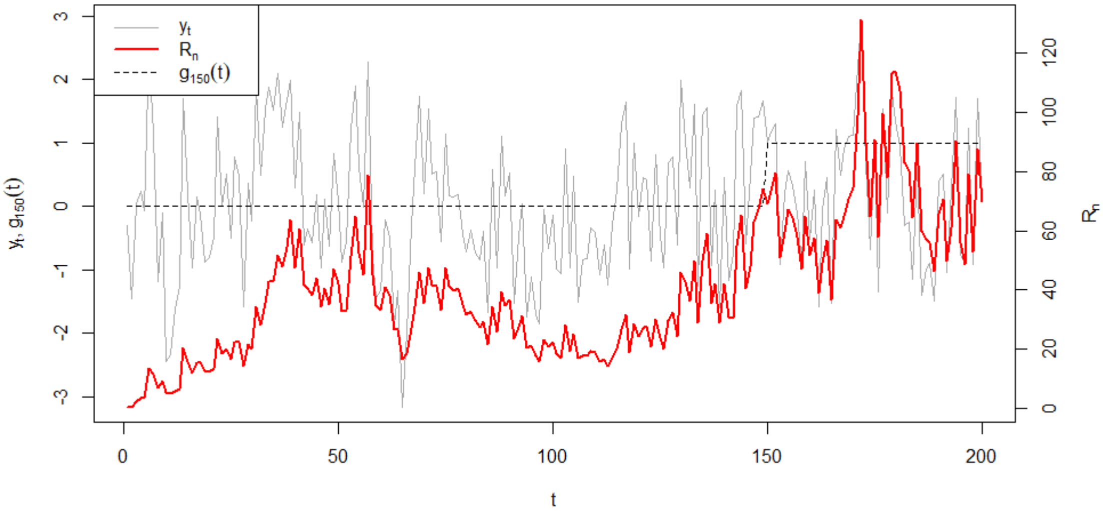


Figure 7: y_t, $g^+_{150}(t)$ and the SR statistic (R_n)

4 SARIMA models with multiplicative anomaly

Other than an additive anomaly, one can conceive of a number of other ways a time series could be interfered with. For instance, if the measurements $y_1, y_2, ...$ are the power output of a collection of generators one would expect a proportional decrease (rather than a flat decrease) in the event of one generator breaking down. Events like this motivate the following problem formulation.

Definition 4.1. The **multiplicative anomaly function** is defined as

$$g_v^*(t) = \begin{cases} 1 & ,t < v \\ g^*(t+1-v) & ,t \geq v \end{cases}$$

where $g^*(t)$ is some known real-valued function and $g^*(1) \neq 1$.

Given that $x_1, x_2, ...$ is SARIMA with known parameters, we have

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d x_t = \Theta_Q(B^s)\theta(B)w_t$$

and the observations

$$y_t = x_t g_v^*(t).$$

We denote $y_t/g_v^*(t) = [\frac{y}{g_v^*}]_t$ as a single entity, such that $B^i[\frac{y}{g_v^*}]_t = [\frac{y}{g_v^*}]_{t-i}$ and write

$$\Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d\big[\frac{y}{g_v^*}\big]_t = \Theta_Q(B^s)\theta(B)w_t. \tag{15}$$

Arriving at a corresponding SR statistic, as in the previous sections, is not possible in this case. Doing so would require a multiplicative inverse to remove $g_v^*(\cdot)$ from the l.h.s of (15). Such element does not exist. Instead, a similar but not identical approach is used. First of all, as before, we note that $Z_t := \Theta_Q(B^s)\theta(B)w_t$ is a sum of normals and therefore also normal. Define $N(n) \sim (Z_1, ..., Z_n)$. Obviously, $E[N(n)] = \mathbf{0}$ and we have the multivariate normal density

$$f_{N(n)}(z_1, ..., z_n) = \frac{1}{\sqrt{(2\pi)^n|\mathbf{C}|}}\exp\left(-\frac{\mathbf{z}^T\mathbf{C}^{-1}\mathbf{z}}{2}\right)$$

where $\mathbf{C}$ is the covariance matrix of $N(n)$. Instead of formulating conditional probabilities, as is needed for the SR statistic, we define $z_t(k) := \Phi_P(B^s)\phi(B)\nabla_s^D\nabla^d\big[\frac{y}{g_k^*}\big]_t$ and the statistic

$$R_n^* = \sum_{k=1}^{n}\frac{f_{N(n)}(z_1(k), ..., z_n(k))}{f_{N(n)}(z_1(\infty), ..., z_n(\infty))} = \sum_{k=1}^{n}\exp\frac{\mathbf{z}(\infty)^T\mathbf{C}^{-1}\mathbf{z}(\infty) - \mathbf{z}(\mathbf{k})^T\mathbf{C}^{-1}\mathbf{z}(\mathbf{k})}{2} \tag{16}$$

We call this the *non-conditional SR* (ncSR) statistic.

4.1 ncSR procedure on ARMA(p,q)

If x_t is an ARMA(p,q) series, then we have

$$\begin{aligned} &\phi(B)x_t = \theta(B)w_t \\ &\quad\Leftrightarrow x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q} \end{aligned} \tag{17}$$

and $w_1, w_2, ...$ iid $N(0, \sigma^2)$. We sequentially observe the time series

$$y_t = x_t g_v^*(t) \tag{18}$$

where $g_v^*(t)$ is a multiplicative anomaly function. Combining and rearranging (17) and (18), we get

$$\left[\frac{y}{g_v^*}\right]_t - \phi_1 \left[\frac{y}{g_v^*}\right]_{t-1} - ... - \phi_p \left[\frac{y}{g_v^*}\right]_{t-p} = w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q}$$

Using the covariance matrix $\mathbf{C}$ defined by (11), and

$$\mathbf{z(k)} = \left((\left[\frac{y}{g_k^*}\right]_1 - \phi_1 \left[\frac{y}{g_k^*}\right]_0 - ... - \phi_p \left[\frac{y}{g_k^*}\right]_{1-p}), ..., (\left[\frac{y}{g_k^*}\right]_n - \phi_1 \left[\frac{y}{g_k^*}\right]_{n-1} - ... - \phi_p \left[\frac{y}{g_k^*}\right]_{n-p}) \right),$$

we can compute the ncSR statistic as defined by (16).

In figures 8, 9 and 10 we see an example instance of of y_t, z_t and ncSR. We have an anomaly at time $v = 250$, i.e.

$$g_{250}^*(t) = \begin{cases} 1 & , t < 250 \\ \frac{3}{4} & , t \geq 250, \end{cases}$$

and an ARMA(3,2) process x_t with AR and MA components $(\phi_1, \phi_2, \phi_3) = (0.3, 0.2, 0.15)$ and $(\theta_1, \theta_2) = (0.4, 0.2)$:

$$\begin{aligned} \phi(B)x_t &= \theta(B)w_t \\ \Leftrightarrow x_t &= 0.3x_{t-1} + 0.2x_{t-2} + 0.15x_{t-3} + w_t + 0.4w_{t-1} + 0.2w_{t-2}, \end{aligned} \tag{19}$$

$w_1, w_2, ...$ iid $N(0, 1)$. We observe the following time series:

$$y_t = x_t g_{250}^*(t) \tag{20}$$

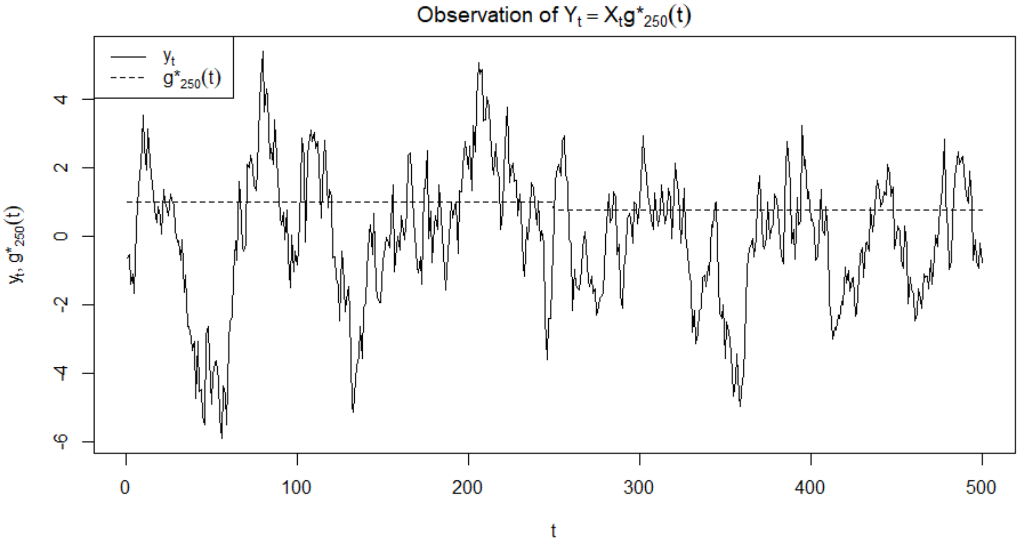


Figure 8: y_t and $g_{250}^*(t)$

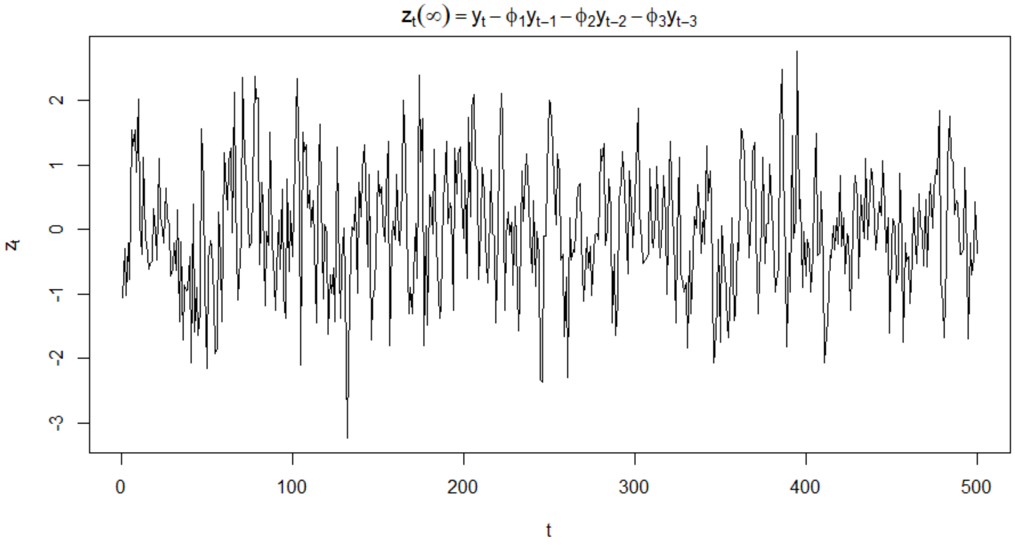


Figure 9: z_t

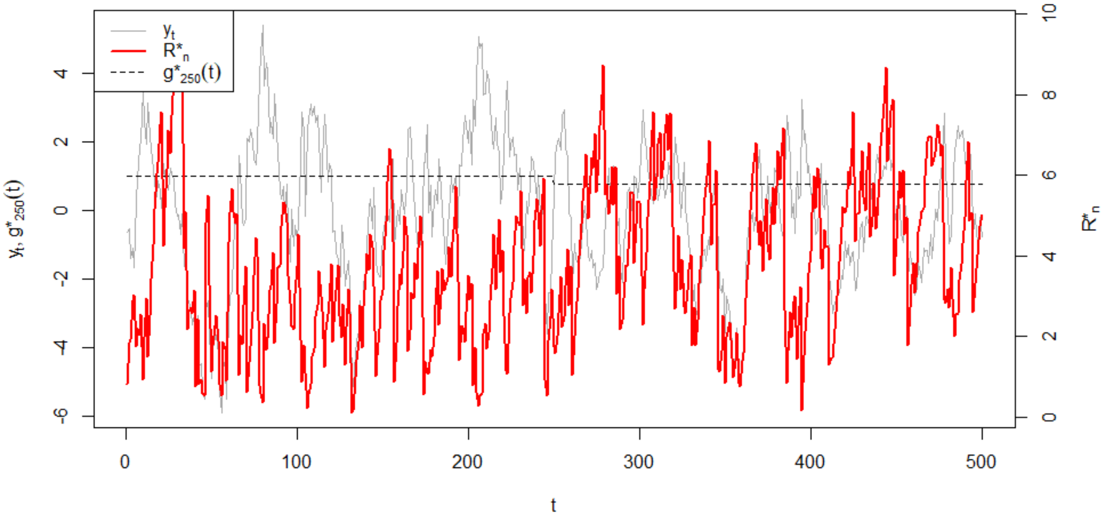


Figure 10: y_t, $g^*_{250}(t)$ and the ncSR statistic (R^*_n)

Figures 11, 12 and 13 depict the same procedure, but with the anomaly function

$$g^*_{150}(t) = \begin{cases} 1 & , t < 150 \\ \frac{5}{4} & , t \geq 150, \end{cases}$$

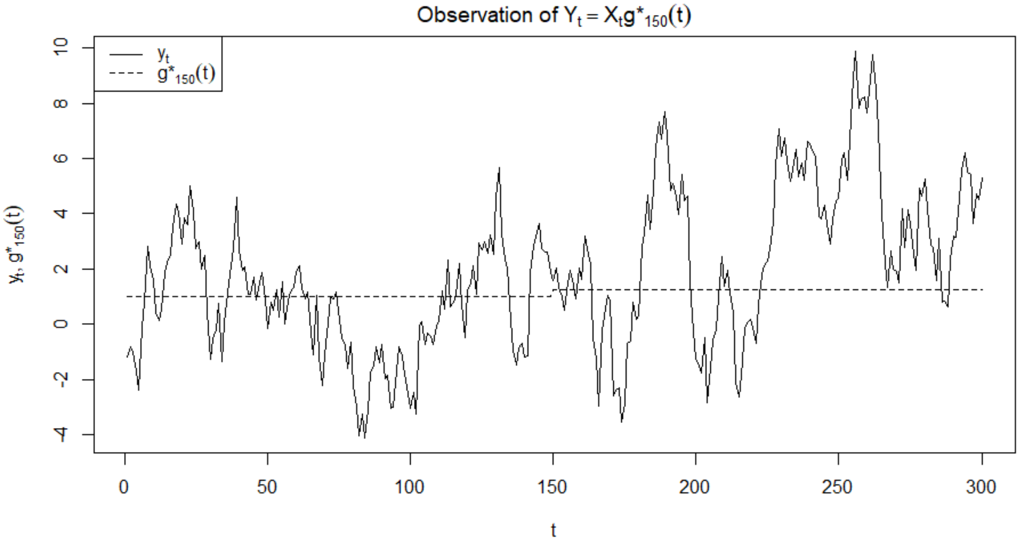


Figure 11: y_t and $g^*_{150}(t)$

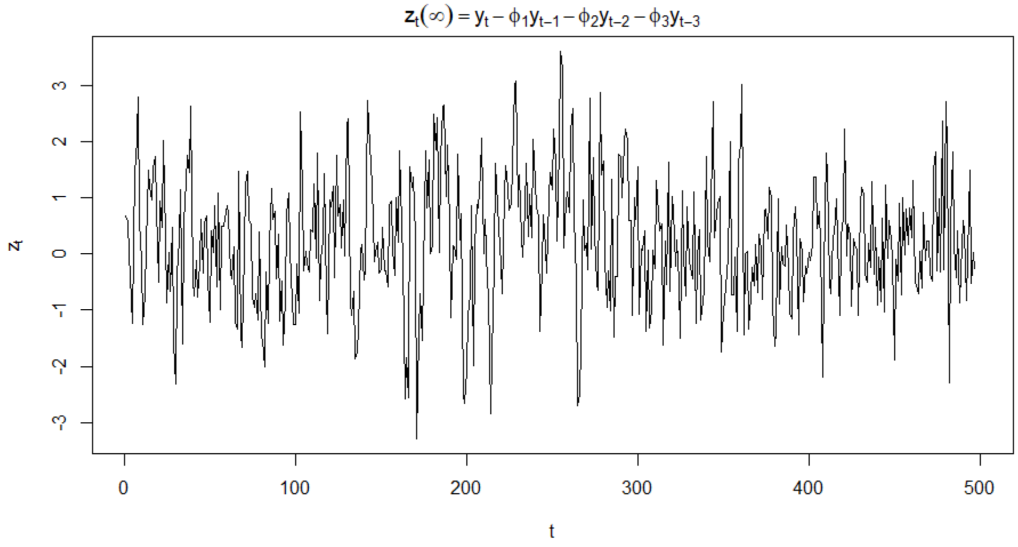


Figure 12: z_t

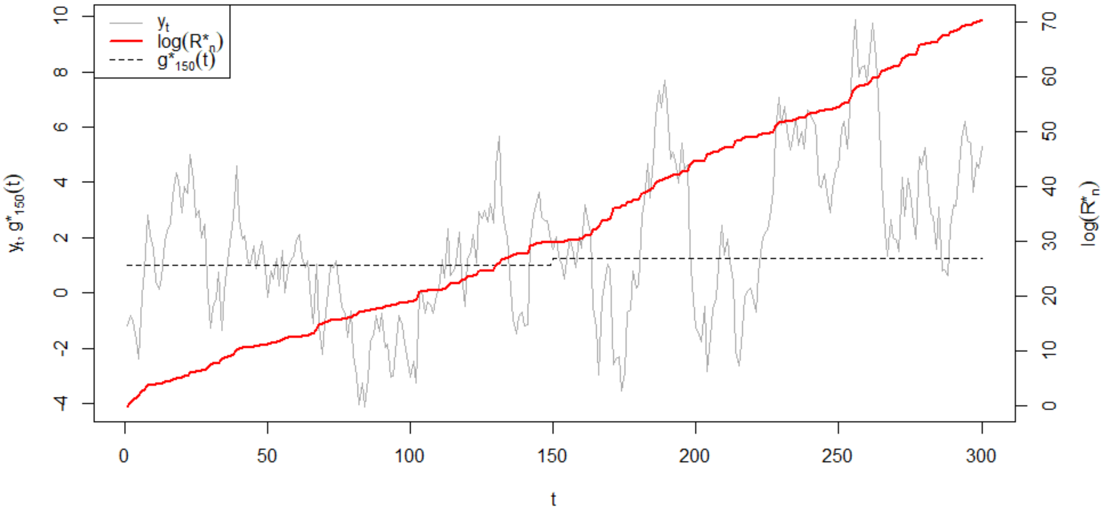


Figure 13: y_t, $g^*_{150}(t)$ and the ncSR statistic (R^*_n)

5 Performance

A natural question arises: are any of these procedures of use? More than anything else, this thesis is an exploration of the connection between change point detection ideas - the SR procedure, specifically - and SARIMA time series modelling. It is still worthwhile to investigate whether or not any utility can be found in this endeavor. Since there's a plethora of variables which affect the performance of these procedures, such as the anomaly function and SARIMA parameters, we will have to focus on only a few cases. First up, we consider the SR procedure on an ARMA(3,2) series.

5.1 SR procedure: additive anomaly on an ARMA(3,2)

We have an ARMA(3,2) process with $(\phi_1, \phi_2, \phi_3) = (0.5, 0.2, 0.15)$, $(\theta_1, \theta_2) = (0.4, 0.2)$, $w_1, w_2, ...$ iid $N(0, 1)$ and the anomaly function

$$g_v^+(t) = \begin{cases} 0 & , t < v \\ 1 & , t \geq v \end{cases}$$

To illustrate how the SR statistic behaves over time, 100 simulations of this ARMA process are simulated with the anomaly and 100 without the anomaly (same seeds).

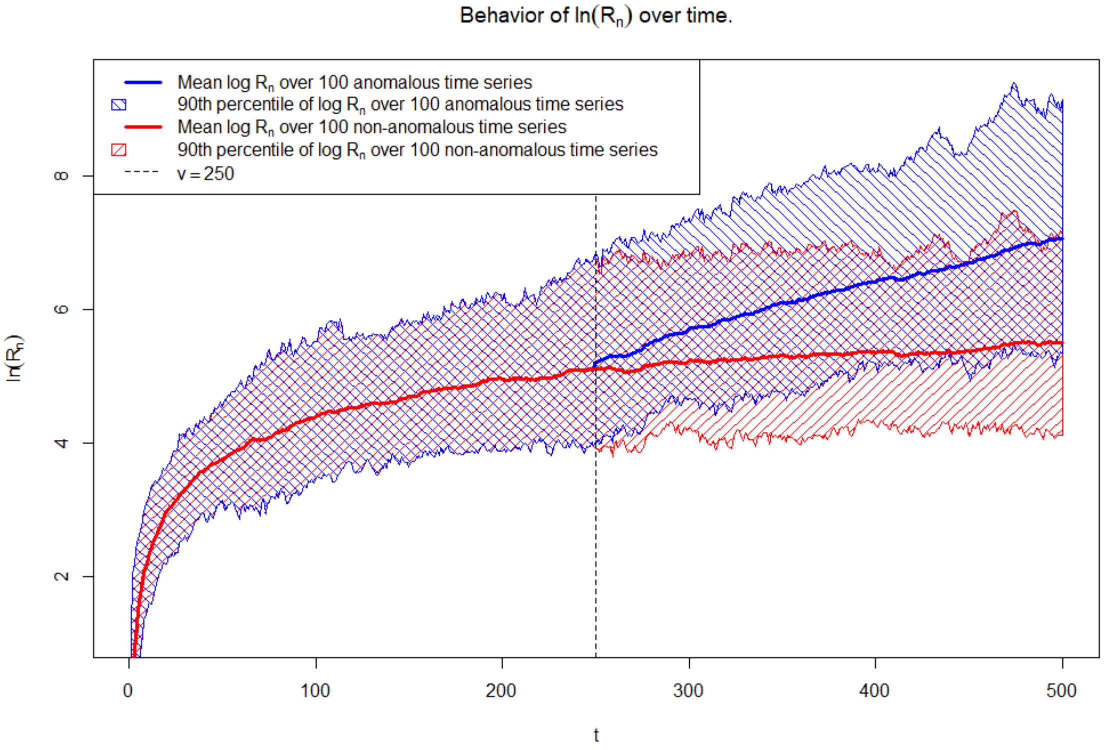
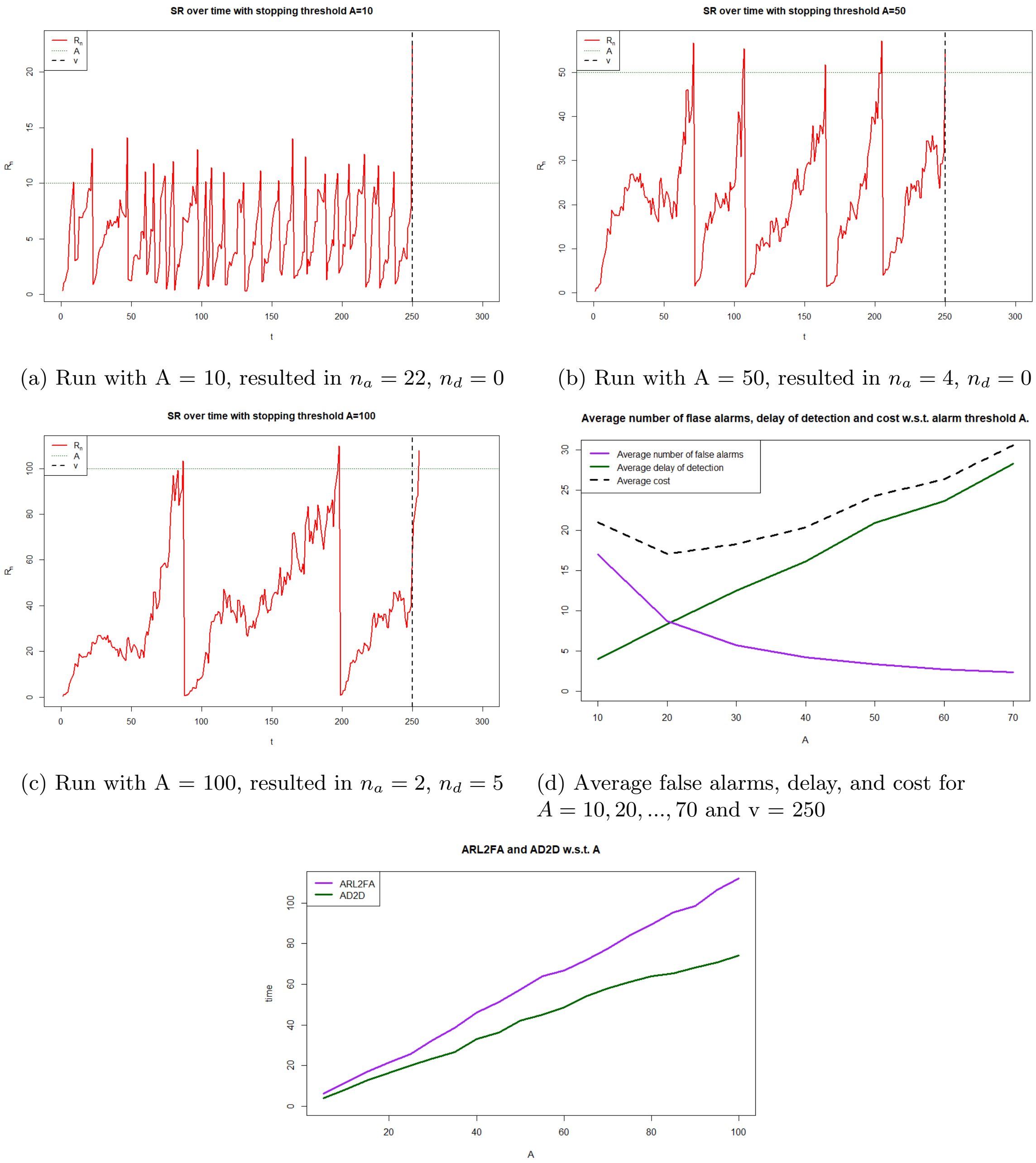


Figure 14: Behavior of SR on ARMA(3,2) with g = 1 for $v = \infty$ and$v = 250$.

While Figure 14 shows that R_n indeed reacts on the anomaly, it does not exhibit how the SR statistic would actually be used. A more realistic scenario is that, in the event of an alarm at time t_a, we either conclude that an anomaly has occurred at time $v \leq t_a$, or that it is a false alarm. In the first case, one would seize the sampling in order to address the anomaly. In the latter case, the SR procedure would be repeated with starting point $t_a + 1$. We can associate this procedure with two costs: the cost of raising a false alarm $(t_a < v)$, and the cost of having a delayed alarm $(t_a > v)$. Let $\hat{R}$ be a completed realisation of the SR procedure, n_f be the number of false alarms raised during it's run and n_d be the delay between the alarm raised after v and v itself.

$$c_{\hat{R}} = cn_f + n_d \tag{21}$$

where c is some positive constant. Since we are not considering a real-life example here, we let $c = 1$. Figure 15 depicts some runs and properties w.r.t. threshold A.



(a) Run with A = 10, resulted in $n_a = 22$, $n_d = 0$

(b) Run with A = 50, resulted in $n_a = 4$, $n_d = 0$

(c) Run with A = 100, resulted in $n_a = 2$, $n_d = 5$

(d) Average false alarms, delay, and cost for $A = 10, 20, ..., 70$ and v = 250

(e) ARL2FA and AD2D for $A = 10, 20, ..., 70$

Figure 15: Properties of SR on ARMA(3,2) with g = 1.

5.2 ncSR procedure: multiplicative anomaly on ARMA(3,2)

We have an ARMA(3,2) process with $(\phi_1, \phi_2, \phi_3) = (0.5, 0.2, 0.15)$, $(\theta_1, \theta_2) = (0.4, 0.2)$ and $w_1, w_2, ...$ iid $N(0, 1)$. We consider first the multiplicative anomaly function

$$g_v^*(t) = \begin{cases} 1 & , t < v \\ 0.75 & , t \geq v \end{cases}$$

To illustrate how the ncSR statistic behaves over time, 100 simulations of this ARMA process are simulated with the anomaly and 100 without the anomaly (same seeds). See Figure 16.

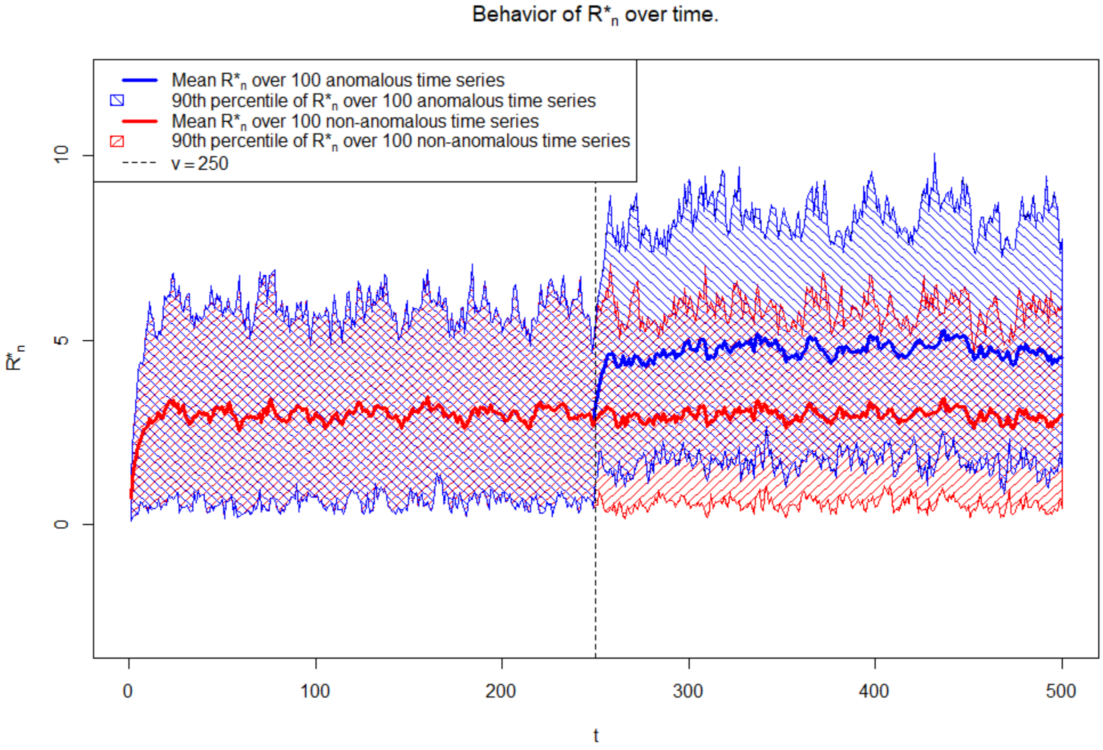
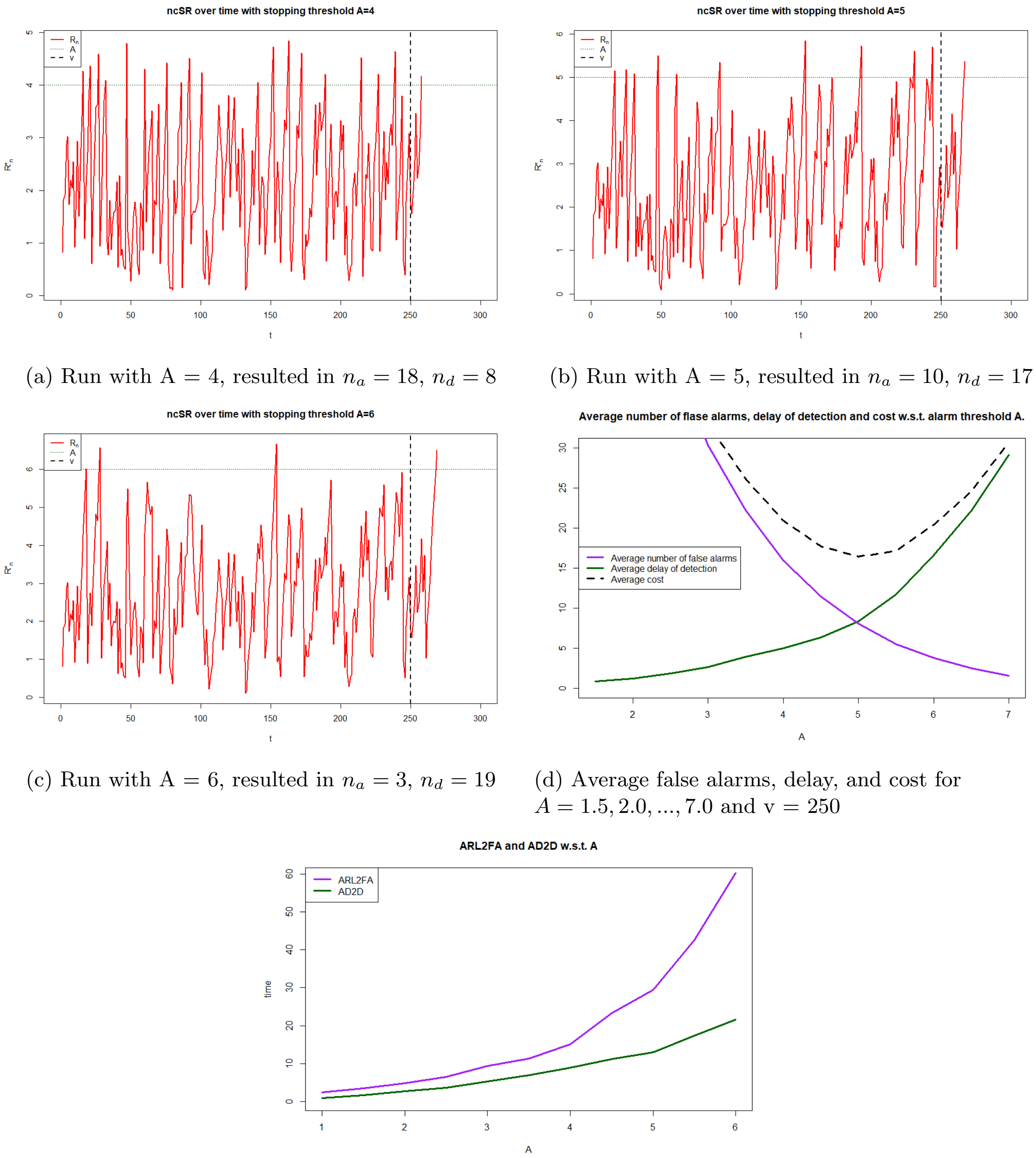


Figure 16: Behavior of ncSR on ARMA(3,2) with g = 0.75 for $v = \infty$ and$v = 250$.

In Figure 17 some test runs and properties w.r.t. threshold A are found (analogous to Figure 15).



(a) Run with A = 4, resulted in $n_a = 18$, $n_d = 8$

(b) Run with A = 5, resulted in $n_a = 10$, $n_d = 17$

(c) Run with A = 6, resulted in $n_a = 3$, $n_d = 19$

(d) Average false alarms, delay, and cost for $A = 1.5, 2.0, ..., 7.0$ and v = 250

(e) ARL2FA and AD2D for $A = 1.5, 2.0, ..., 7.0$

Figure 17: Properties of ncSR on ARMA(3,2) with g = 0.75.

Figures 18 and 19 illustrate the corresponding properties for the anomaly function

$$g_v^*(t) = \begin{cases} 1 & ,t < v \\ 1.25 & ,t \geq v \end{cases}$$

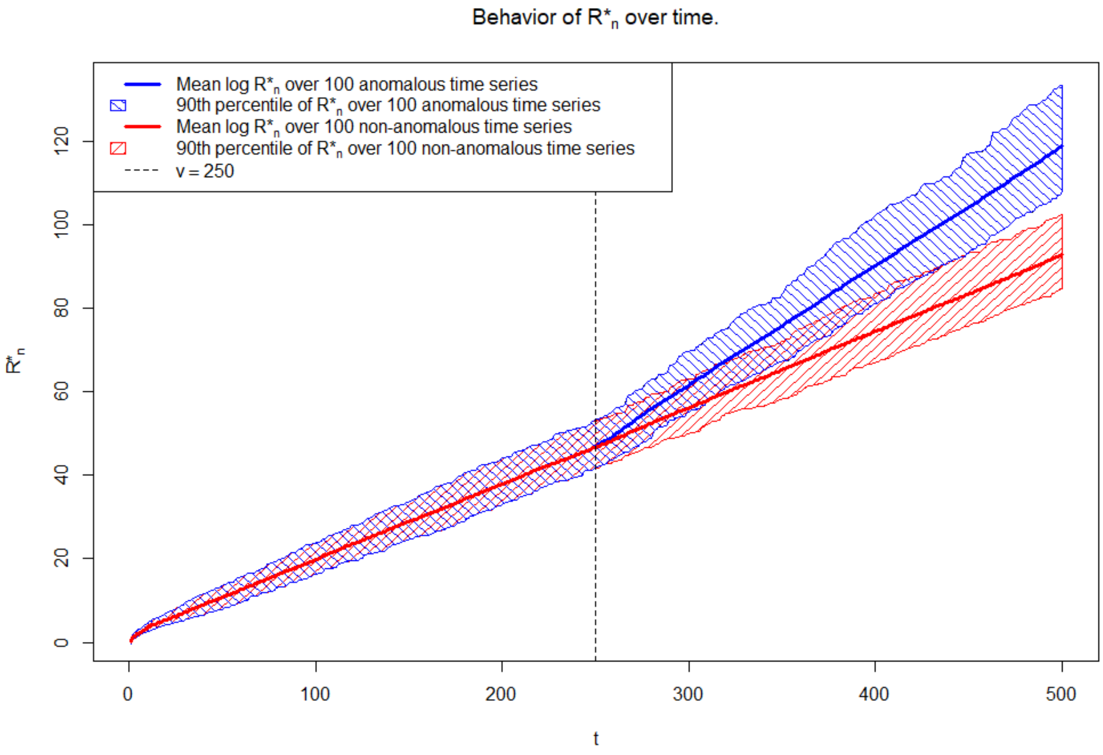
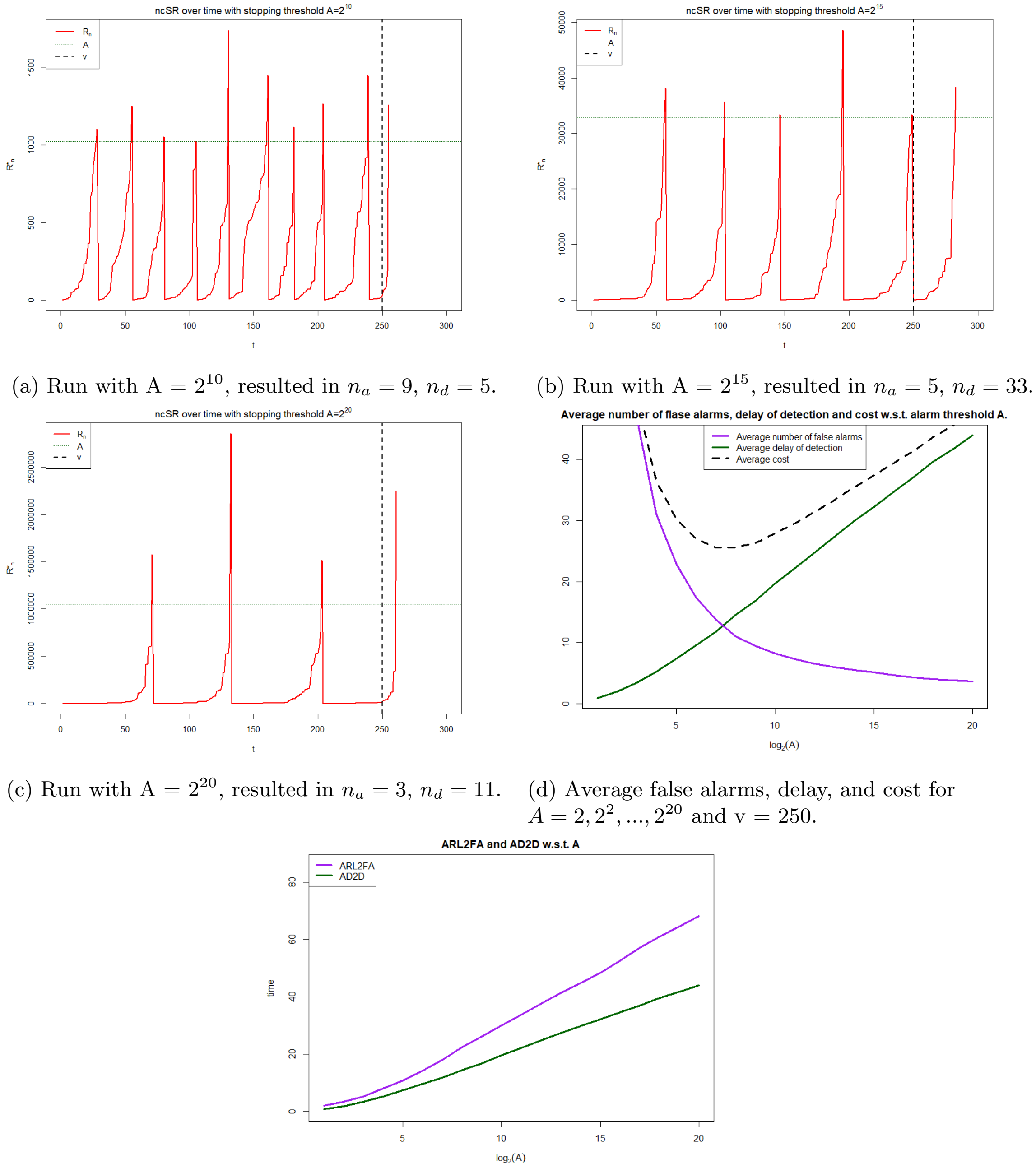


Figure 18: Behavior of ncSR on ARMA(3,2) with g = 1.25 for $v = \infty$ and $v = 250$.



(a) Run with A $= 2^{10}$, resulted in $n_a = 9$, $n_d = 5$.

(b) Run with A $= 2^{15}$, resulted in $n_a = 5$, $n_d = 33$.

(c) Run with A $= 2^{20}$, resulted in $n_a = 3$, $n_d = 11$.

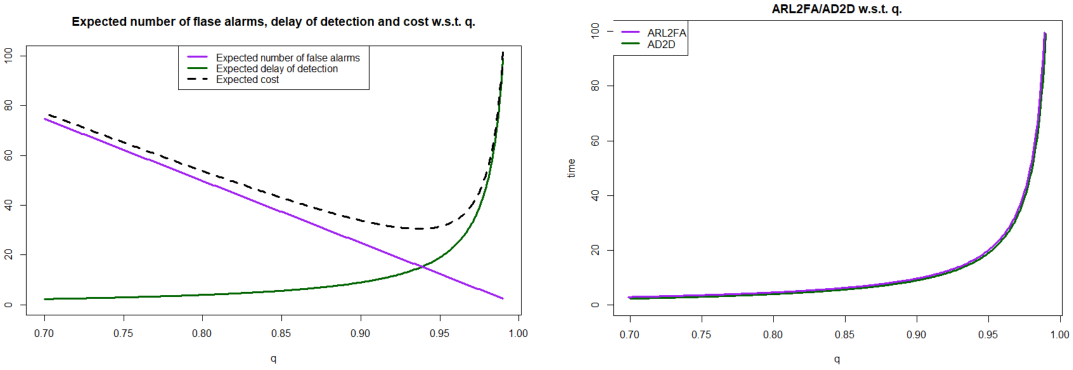
(d) Average false alarms, delay, and cost for $A = 2, 2^2, ..., 2^{20}$ and v $= 250$.

(e) ARL2FA and AD2D for $A = 2, 2^2, ..., 2^{20}$

Figure 19: Properties of ncSR on ARMA(3,2) with g $= 1.25$.

6 Discussion

The results presented in section 5 shows us that these methods do react to the introduced anomalies. One minimum bar to pass is that the ARL2FA should be higher than the AD2D. This was true for all examples considered. If this was not the case, these methods would be worse than useless. In deed, if we at every time step we independently raise an alarm with some probability $1 - q$, we have that $ARL2FA = AD2D$, as seen in Figure 20 (b). We would also witness a similar trade-off between n_f and n_d, if this mundane procedure were employed (compare Figure 20 (a) to Figures 15 (d), 17 (d) and 19 (d)).



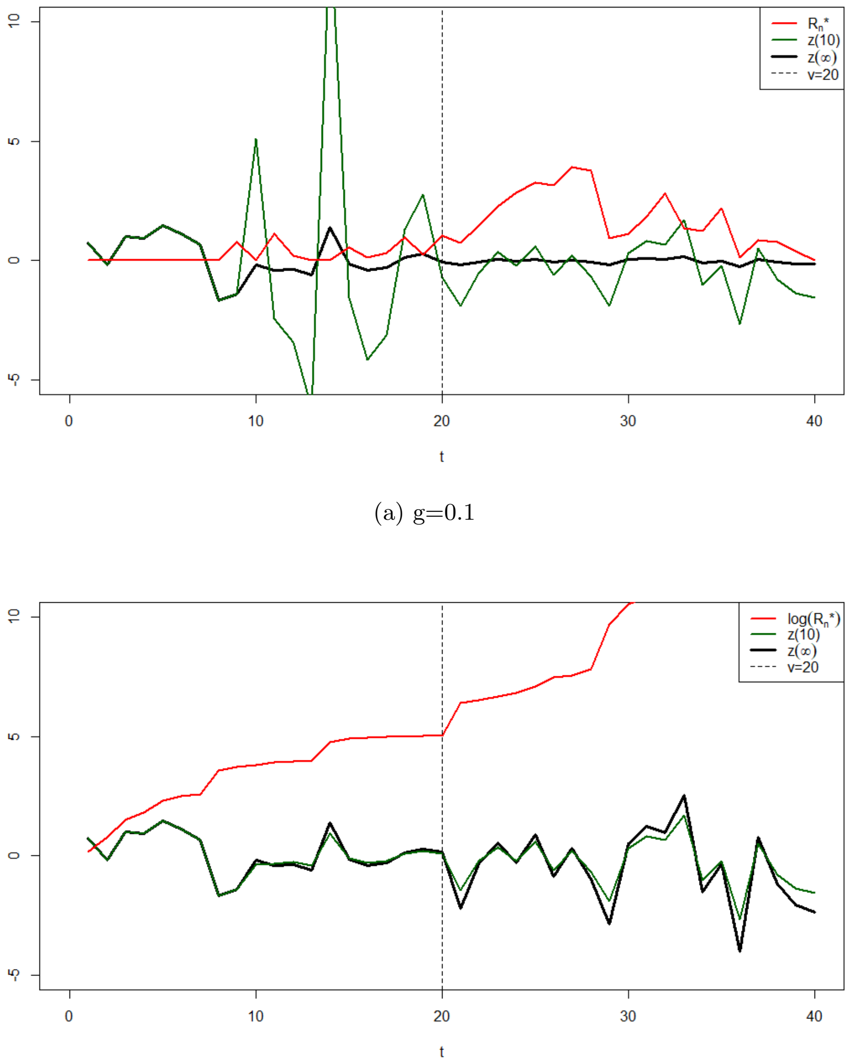
(a) Expected number of false alarms, delay of detection and cost for $v = 250$

(b) Expected ARL2FA/AD2A

Figure 20: Properties of the random alarm procedure with probability $1 - q$

However, note that the number of false alarms risen (n_f) depends on the anomaly time v. Since the whole point of these procedures is to detect v, one would probably not use the cost as described in (21) (Figure 15(d), 17(d) and 19(d)) to determine an appropriate stopping threshold. A more practical approach would be to choose a threshold which achieves an acceptable AD2D and/or an acceptable ARL2FA on historical/training data.

The conjecture was that the SR statistic would grow exponentially, especially if an anomaly is introduced. The results of the experiments supports this notion (Figure 14). The ncSR used on multiplicative anomaly behaves like a step function for $g < 1$ (Figure 16) and exponentially for $g > 1$ (Figure 18). While the ncSR statistic reacts to the anomaly, it is clear that it should not be interpreted the same way as the SR statistic. Instead of taking a likelihood ratio of an observed series $\mathbf{z}$ between distributions $N_k(n)$ and $N_\infty(k)$, like in the SR statistic (5), we are taking the likelihood ratio between different observations $\mathbf{z}(\mathbf{k})$ and $\mathbf{z}(\infty)$ from the same distribution $N(n)$ (16). This becomes a problem for the ncSR if we have a multiplicative anomaly. Consider the case where we have an anomaly $g_v^*(t)$ with $g(t) < 1$. The observations $z_t(\infty), ..., z_n(\infty)$ will be closer to zero compared to $z_t(k), ..., z_n(k)$ for $k \geq t$ (Figure 21 (a)). This means that, since $N(n)$ is multivariate normal with zero mean, we generally have $f_{N(n)}(z_1(\infty), ..., z_n(\infty)) > f_{N(n)}(z_1(k), ..., z_n(k))$. In the case where $g > 1$, we see the opposite relationship between $\mathbf{z}(\mathbf{k})$ and $\mathbf{z}(\infty)$ (Figure 21 (b)), i.e. $f_{N(n)}(z_1(\infty), ..., z_n(\infty)) < f_{N(n)}(z_1(k), ..., z_n(k))$. The ncSR is not taking the variance into account in a satisfactory way.



(a) g=0.1

(b) g = 1.5

Figure 21: $(z_1(10), ..., z_{40}(10)), (z_1(\infty), ..., z_{40}(\infty))$ and the ncSR statistic

7 Conclusion and Future work

Investigating the possibility of finding a truly analogous statistic of the SR when we have a multiplicative anomaly would be the author's first extension of this work, had there been more time. One idea to solve this would be trying to use the fact that a casual ARMA process x_t can be written as an MA process (with infinite MA-coefficients) [5]. As such one can write

$$x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q} = \sum_{j=0}^{\infty} \psi_j w_{t-j} \approx \sum_{j=0}^{m} \psi_j w_{t-j}, \tag{22}$$

for some arbitrarily large m, and $\psi_1, \psi_2, ...$ determined by $\theta(\cdot)$ and $\phi(\cdot)$. We then observe

$$y_t = x_t g_v^*(t). \tag{23}$$

(22) and (23) could then be used to acquire the approximation

$$y_t \approx g_v^*(t) \sum_{j=0}^{m} \psi_j w_{t-j}$$

which implies that y_t is approximately normal distributed. This allows us to formulate the SR statistic

$$R_n = \sum_{k=1}^{n} \frac{f_{N_k(n)}(y_1, ..., y_n)}{f_{N_\infty(n)}(y_1, ..., y_n)}.$$

where $N_k(n) = (Y_1, ..., Y_n | v = k)$. Since the SR procedure has proven optimality properties[3], whereas the ncSR is - as discussed in the previous section and witnessed in section 5 - not very suitable for this task, this route might be fruitful.

Another relevant extension would be to test how well these methods work on real data sets, where the SARIMA model has to be estimated before anything else. Clearly, this would result in more degrees of freedom in the procedure and less accurate results. Now, there are many ways to approach time series modelling in a SARIMA context; manual methods, automated methods, and everything in between [2][5]. Of course, different approaches to modelling will lead to different statistical properties. There is therefore no shortage of insights to be had in this regard.

Lastly, we have still not established whether or not these methods are practically usable. When it comes to these small examples presented in this thesis, computational time was not an issue. However, a thorough analysis of computational performance would be essential before any practical application. Also, while it is clear that these methods are better than useless (compare Figures 15, 17 and 19 to Figure 20), that is not a very difficult bar to pass. More research, both in a theoretical and empirical context, is needed before making any conclusive remarks regarding performance. For example, analysis of the theoretical expected ARL2FA and AD2D and comparisons to other change point detection procedures of time series. In other words, this thesis shows that there are many opportunities to investigate further.

References

[1] Banerjee A. Chandola, V. and V. Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 2009.

[2] Rob Hyndman et al. *Forecasting Functions for Time Series and Linear Models*, 2019.

[3] M. Pollak and A.G. Tartakovsky. Optimality properties of the shiryaev-roberts procedure. *Statistica Sinica*, 2009.

[4] A. N. Shiryaev. The problem of the most rapid detection of a disturbance in a stationary process. *Soviet Math. Dokl. 2, 795-799*, 1961.

[5] Robert H. ShumwayDavid S. Stoffer. *Time Series Analysis and Its Applications*. Springer, Cham, 2017.

[6] Alexander G. Tartakovsky and George V. Moustakides. Discussion on “quickest detection problems: Fifty years later” by albert n. shiryaev. *Sequential Analysis*, 29(4):386–393, 2010.