Time Series Analysis

Olivier Wintenberger

2021–2022

Contents

Part I

Preliminaries

Chapter 1

Stationarity

We focus on discrete time processes $(X_t)_{t\in\mathbb{Z}}$ where t refers to time and X_t is a random variable (extensions to multivariate time series will also be considered hen possible). The random sequence (X_t) is built upon a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Observing $(X_1, \ldots, X_n)$ at times $t = 1, \ldots, n$, the classical statistical issue is to predict the future at time n: $X_{n+1}, X_{n+2}, \ldots$. In order to do so, we proceed in two steps; First inferring the dependence structure of the observations and second using the dependence structure in order to construct a prediction.

Remark. *To achieve the prediction objective, we have to assume a structure (a model) on (X_t) so that the information contained in $(X_1, \ldots, X_n)$ provides information on the future values of the process. We use the concept of* stationarity.

Definition 1. *The (possibly multivariate) process (X_t) is strictly (or strongly) stationary if for all $k \in \mathbb{N}$, the joint distribution of $(X_t, \ldots, X_{t+k})$ does not depend on $t \in \mathbb{Z}$.*

Hence, in order to predict the future at time n, one can subsample $(X_1, \ldots, X_n)$ in blocks of length k and use the fact that $(X_{n-k}, \ldots, X_{n+1})$ and $(X_1, \ldots, X_{k+2})$, $(X_2, \ldots, X_{k+3})$ are identically distributed... On the last blocks, the last values X_{k+2}, X_{k+3},... are observed so that one can assert the predictive power of the prediction of the last value of the block from the k first values within each block.

If the process is not likely to be stationary, one cannot rely on the observations to predict the future. In practice, one has to "stationarize" our observations first, ie pre-process the data so that stationarity is a reasonable assumption.

1.1 Data preprocessing

Let us assume that we observe data (D_t) indexed by the time t. Our aim is to find a reasonable transformation X_t of the data D_t such that (X_t) can be likely stationary. We will not discuss here potential preprocessing that are not specific to time series such as missing values, outliers,...

Consider that we are in the univariate case, otherwise the following treatment applies to each marginals *independently*. Most of the time series can be decomposed in three additive parts:

$$D_t = f(t) + S_t + X_t, \qquad t \geq 1, \tag{1.1}$$

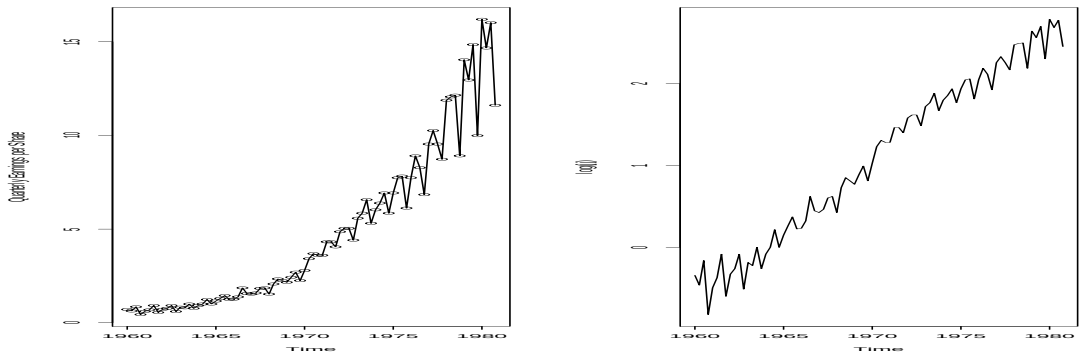

Figure 1.1: Econometrics data exhibiting an exponential (multiplicative) trend that turns into a linear (additive) trend after log-transform

where $f(t)$ is the trend part, i.e. a deterministic function f of the time t, S_t is a seasonal part with period $S_{t+T} = S_t$ for some period T and X_t is likely stationary. Of course the decomposition is not unique and it is a hard work to identify each components.

The additive form in (1.1) is completely artificial and chosen for its simplicity. For some data as for economics time series, a multiplicative form is much more natural. A log transformation is necessary to obtain the additional decomposition (1.1).

Example 1. *For economics data, it is reasonable to take into account an exponential trend due to the inflation. For time period $t = 1, \dots, n$ where the interest rate r is assumed to be fixed, the nominal price D_t is actually the real (deflated) price P_t times the inflation:*

$$D_t = P_t e^{rt}, \qquad t = 1, \dots, n.$$

Due to the presence of the exponential trend, this data cannot be stationary. By applying the log transform, we obtain

$$\log(D_t) = \log(P_t) + rt, \qquad t = 1, \dots, n.$$

The exponential trend is transformed in a linear additive trend that we will treat hereafter thanks to the additional decomposition (1.1)*. Figure 1 shows quarterly earnings per share for the U.S. company Johnson & Johnson from 1960 to 1980.*

1.1.1 Differencing

Let us treat the trend part $f(t)$ in the decomposition (1.1), assuming that the seasonality part is null $S_t \equiv 0$, $t \geq 0$. In what follows we will consider that $f(t)$ is a polynomial of the time t. The most common case is the one of linear trend as

$$f(t) = a_0 + b_0 t, \qquad t \geq 1,$$

where (a_0, b_0) are unknown coefficients. As a statistician, a natural approach is to treat this term as a linear model

$$D_t = a + bt + X_t, \qquad t \geq 1.$$

Then (X_t) is estimated from the residuals of the linear regression It is not the good approach as we will see on an example.

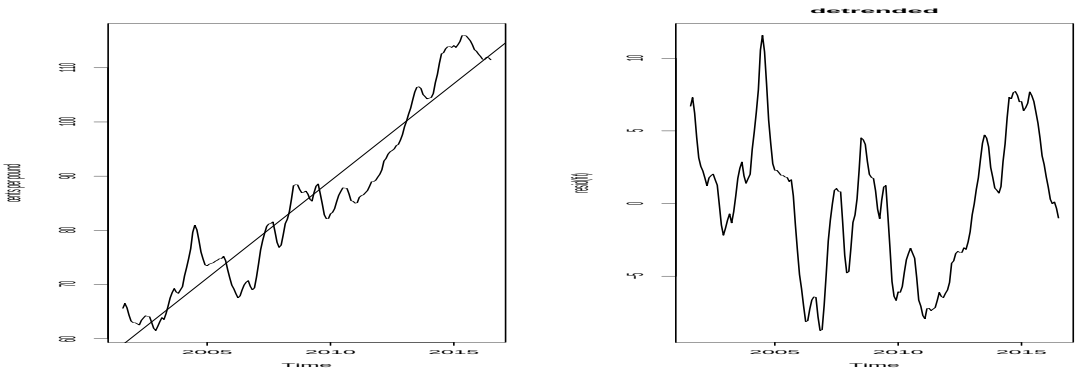

Figure 1.2: Estimation of the stationary component in presence of a linear trend thanks to linear regression on the time.

Example 2. *Let us regress the price of chicken in cents on the unit of time from 2001 to 2016 (note that on such short periods economic prices can be reasonably linear trended as the inflation $e^{rt} \sim 1 + rt$ when rt is small). Then X_t is taken as the residuals from the linear regression, see Figure 2.*

Let us introduce the important notion of filtration $(\mathcal{F}_t)$, which is a sequence of increasing σ-algebras. The events in $\mathcal{F}_t$ represent the available information at time t. A natural way to describe a filtration is to introduce a noise.

Definition 2. *A Strong White Noise (SWN) is some independent and identically distributed (i.i.d.) sequence (Z_t) observed at time t such that $\mathbb{E}[Z_0] = 0$ and $Var(Z_0) < +\infty$ (possibly multi-dimensional).*

A SWN generates the natural filtration $\mathcal{F}_t = \sigma(Z_t, Z_{t-1}, \dots)$ which corresponds formally to the information available at time t. The prediction at time n cannot use any information from the future Z_{n+1}, Z_{n+2},... The SWN (Z_t) is an *unpredictable* sequence; for instance, the best prediction for Z_{n+1} for the quadratic risk given the past is $\mathbb{E}[Z_t \mid Z_{t-1}, Z_{t-2}, \dots] = 0$. It corresponds to the classical i.i.d. setting studied in any basic course in statistics. In such i.i.d. settings, more interesting problems than prediction are usually treated (estimation and tests).

Let (Z_t) be a SWN (usually not observed).

Definition 3. *The process (X_t) is* non-anticipative *relatively to the SWN (Z_t) if $X_t \in \sigma(Z_t, Z_{t-1}, \dots)$. The process (X_t) is* invertible *if $X_t \in \sigma(Z_t, X_{t-1}, \dots)$.*

Notice that an invertible process is non-anticipative. The invertibility is the most important notion related to filtration in time series analysis. It means there is an incompressible random error in the prediction of X_{n+1} due to the lack of information Z_{n+1}, unknown and unpredictable at time n. It is fundamental to avoid degenerate situations (and not reasonable in our random setting) where one can predict the future from past observations. We say that the data are adapted to the filtration and respects the flow of information.

Example 3 (2, continued)**.** *Assume that (D_t) is non-anticipative with respect to (Z_t). Estimating the coefficients (a_0, b_0) thanks to the linear regression on $(D_1, \dots, D_n)$, one obtains coefficients $(\widehat{a}_0(D_1, \dots, D_n), \widehat{b}_0(D_1, \dots, D_n))$. Thus the residuals*

$$\widehat{X}_t = D_t - \widehat{a}_0(D_1, \dots, D_n) - \widehat{b}_0(D_1, \dots, D_n)t, \qquad 1 \le t \le n,$$

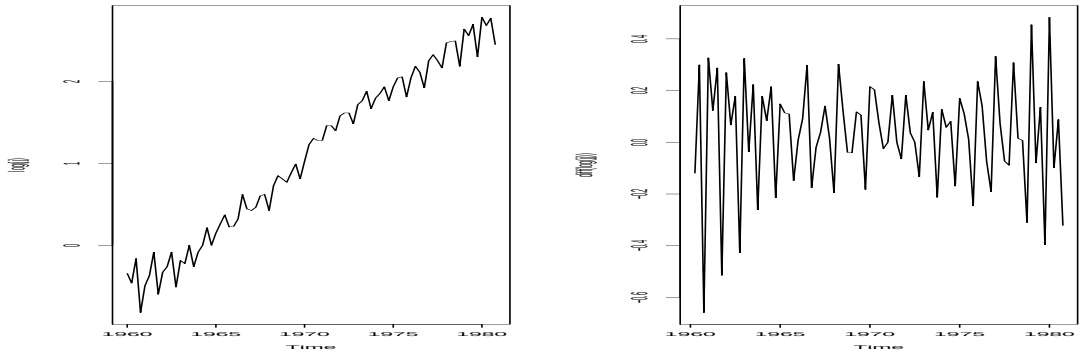

Figure 1.3: A linear (additive) trend is removed thanks to differencing. The obtained time series has a mean behavior constant in time but exhibits some heteroscedastic behavior, i.e. a non constant variance.

constitute an anticipative sequence because they depend on the future D_s and thus Z_s for $s > t$. Such preprocessing transformation does not respect the flow of information. It is likely that $(\widehat{a}_0(D_1,\dots,D_n) - \widehat{b}_0(D_1,\dots,D_n)t)$ overfits the data $(D_1,\dots,D_n)$. It usually biases the predictive power analysis and requires additional care, usually treated via a penalization procedure.

Preprocessing that respects the flow of information are based on the difference operator:

Definition 4. *The lag (or backshift) operator L is defined as $L((D_t)) = (D_{t-1})$ for any data D_t, $t \in \mathbb{Z}$. The difference operator $\Delta = Id - L$ with Id the identity over $\mathbb{R}^{\mathbb{Z}}$ is defined so that $\Delta D_t = D_t - D_{t-1}$, $t \geq 1$.*

In our case $D_t = a + bt + X_t$, applying the difference operator, we obtain

$$\Delta D_t = D_t - D_{t-1} = b + \Delta X_t, \qquad t \geq 1.$$

If (X_t) is stationary, then $b + \Delta X_t$ is also stationary and applying the difference operator stationarizes the linear trended data (D_t). Notice that the flow of information is preserved; if (D_t) is non-anticipative with respect to SWN (Z_t) so is (ΔD_t).

Example 4 (1, continued)**.** *On economics data, the log transformed data $\log(D_t) = \log(P_t) + rt$ exhibit a linear trend. Applying the difference operator, we obtain $\Delta\log(D_t) = \log(P_t/P_{t-1}) + r$ which is reasonably stationary. Neglecting the influence of the interest rate, one calls the obtained process $X_t = \Delta\log(D_t)(*100)$ the log-ratios.*

Example 5 (2, continued)**.** *We perform the difference operator on the chicken prices and compare it with the residuals of the linear regression. The residuals of the linear regression have a very smooth trajectory that seems simpler to predict than the difference ΔD_t, see Figure 4. It is due to the use of future observations to calculate the residuals at any time $1 \leq t \leq n$. The non respect of the flow of information may lead to overconfident predictions and then overfitting.*

The trend component $f(t)$ can be much more complicated than a simple linear dependence on time. We will treat any polynomial trend thanks to multiple differencing; consider a polynomial trend of degree 2

$$D_t = a_0 + b_0 t + c_0 t^2 + X_t, \qquad t \geq 1,$$

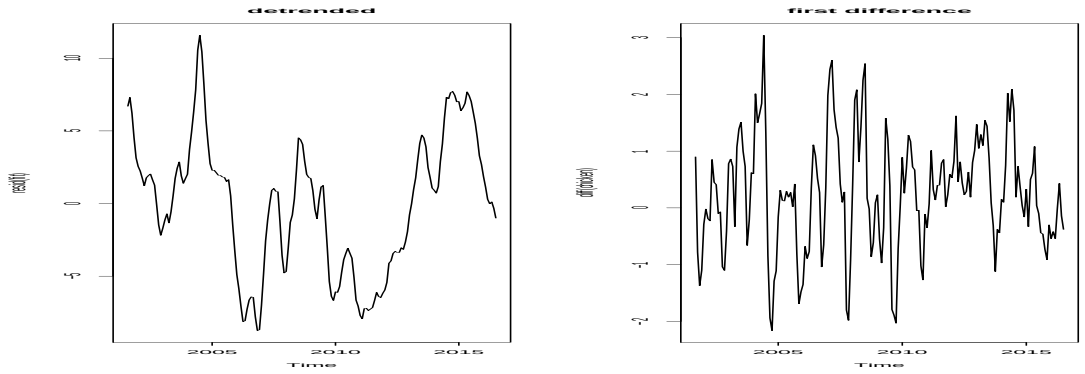

Figure 1.4: The differencing data are more variable then the residuals of the linear regression. There are also more reasonably stationary.

then, differencing once, we obtain a linear trend

$$\Delta D_t = b_0 + c_0(2t-1) + \Delta X_t = b_0 - c_0 + 2c_0 t + \Delta X_t, \qquad t \geq 1.$$

If (X_t) is stationary, so is (ΔX_t) and we are back to the linear trended case and we stationarize ΔD_t by differencing:

$$\Delta(\Delta D_t) = \Delta^2 D_t = D_t - 2D_{t-1} + D_{t-2} = 2c_0 + \Delta^2 X_t, \qquad t \geq 1.$$

In this case $(2c_0 + \Delta^2 X_t)$ corresponds to the stationarized version of (D_t). By a recursive argument, we see that we can treat any polynomial trend by successive differencing. Successive applications of the difference operator respect the flow of the information. Moreover, it is simple to come back to the orignal data D_t by the inverse operator, called *integration*:

$$\Delta D_t = X_t \qquad \Longleftrightarrow \qquad D_t = D_{t-1} + X_t\,,$$

(D_t) being called the integrated version of (X_t). Denoting $X_t = \Delta^2 D_t$, assuming that it is stationary so that we can construct a predictor $\widehat{X}_{n+1}$ then

$$\widehat{D}_{n+1} = \widehat{X}_{n+1} + 2D_n - D_{n-1}.$$

Let us treat the seasonal part S_t in the decomposition (1.1), assuming that the trend part is null $f(t) = 0$. Notice that in practice it is not a restriction; the previous discussion on removing the trend part is extendable in presence of a seasonal component $S_t \neq 0$. Thus, one can always assume that successive differencing of the data removed the trend part and one applies the seasonality decomposition that follows.

As $S_{t+T} = S_t$, knowing the period T the seasonal coefficients $(S_j)_{1\leq j\leq T}$ are easily estimated by the empirical mean

$$\widehat{S}_{kT+j} = \frac{T}{n} \sum_{1\leq t=kT+j\leq n} D_t, \qquad 1 \leq j \leq T, 1 \leq k \leq n/T.$$

The preprocessing $D_t - \widehat{S}_j$ for $t = Tk + j$ breaks the flow of information. Thus, there is a risk of overfitting and it should not be used.

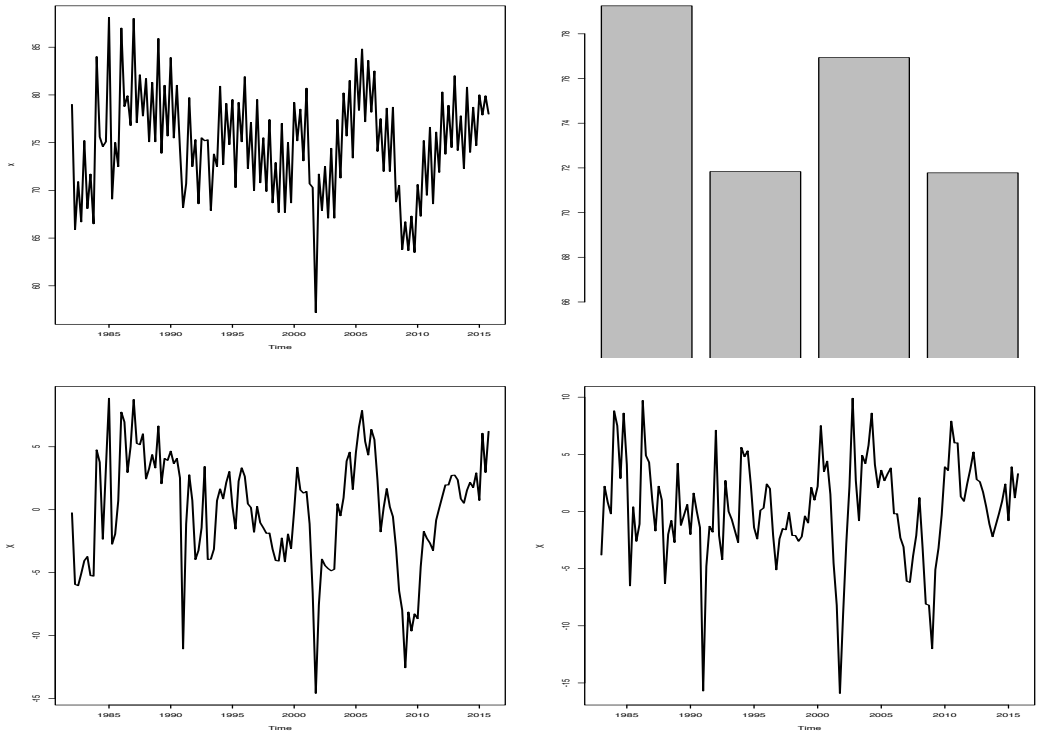

Figure 1.5: The orignal data (D_t), the 4 seasonal coefficients $(\widehat{S}_j)_{1\leq j\leq 4}$, the seasonally adjusted time series $(X_t = D_t - S_t)$ and the differenced data $(X_t = D_t - D_{t-T})$.

Example 6 (Figure 6)**.** *Consider the quarterly occupancy rate of Hawaiian hotels from 2002 to 2016 (top left). There is a strong suspicion of a seasonality of period $T = 4$. Thus, one can compute the 4 seasonal coefficients (top right). One can notice that the spring and autumn coefficients are equals, thus one could suspect a shorter (preferable) period $T = 2$. However, it is not the case as the winter and summer coefficients (the busy seasons) are significantly different. The seasonally adjusted time series $(X_t = D_t - S_t)$ (bottom left) and the differenced data $(X_t = D_t - D_{t-T})$ (bottom right) have similar patterns but the differecing is less smooth.*

Instead, we will use differencing:

Definition 5. *The difference operator of order T is $\Delta_T = Id - L^T$.*

If (D_t) has a seasonal component of period T such that $S_t = S_{t+T}$ then $\Delta_T D_t = D_t - D_{t-T} = X_t - X_{t-T} = \Delta_T X_t$ is stationary.

Moreover note that the differenced stationary time series $D_t = X_t$ is centered: $\mathbb{E}[\Delta D_t] = \mathbb{E}[\Delta X_t] = 0$. Thus by applying an extra time the difference operator, one can always consider that the pre-processed data are stationary and centered.

Differencing is very popular since the seminal work of Box and Jenkins (2011). It is a pre-process widely used on time series. His first merit is to avoid any use of the future for pre-processing the past and thus to reduce the risk of overfitting. Another merit is that we have $\Delta_T \Delta^k = \Delta^k \Delta_T$ thus the treatment of a polynomial trend and seasonal components can be done in any order. The main drawback of this approach is that the determination

of k and T is made visually and not rigorously. Any preprocessing should be doe with great care.

It seems that there is no limit in the differencing process: the more you difference and the more you are likely stationary. However, there is a caveat. Consider for instance one observes a SWN (D_t) in $\mathbb{R}$ with finite variance σ^2. Then $\Delta D_t = D_t - D_{t-1}$ is also stationary, so it is tempting to erroneously difference the observations. However, $\mathrm{Var}(\Delta D_t) = 2\sigma^2 > \mathrm{Var}(D_t)$ and the variance of the differencing process is larger than the original one. More generally, preprocessing by successive differencing should stop when it increases the variance, i.e. when $\mathrm{Var}(\Delta X_t) > \mathrm{Var}(X_t)$ and X_t is centered. Then (X_t) is considered as the stationary version of the data.

1.2 Second order stationarity

We consider now that the preprocessing has been applied and that (X_t) is reasonably stationary and centered. Let us first consider that it is likely second order stationary which implies some homoscedasticity (not a lot of extreme values).

Definition 6. *The (possibly multivariate) time series (X_t) is second order stationary (or weakly stationary) if $\mathbb{E}[X_t]$ and $\mathbb{E}\left[X_t X_{t+k}^\top\right]$ exist and do not depend on t, for all $k \in \mathbb{N}$.*

Remark. *Strong stationarity combined with the existence of second order moments imply second order stationarity.*

1.2.1 Autocorrelations

Definition 7. *Let (X_t) be a centered second order stationary process (univariate). We define, for any $h \in \mathbb{Z}$:*

- *the* autocovariance function*:*

$$\gamma_X(h) = \mathrm{Cov}(X_t, X_{t+h}) = \mathrm{Cov}(X_0, X_h) = \mathbb{E}[X_0 X_h],$$

- *the* autocorrelation function*:*

$$\rho_X(h) = \rho(X_t, X_{t+h}) = \frac{\gamma_X(h)}{\gamma_X(0)}.$$

- *The* cross-covariance function*:*

$$\gamma_{XY}(h) = \mathrm{Cov}(X_t, Y_{t+h}) = \mathbb{E}[X_0 Y_h]$$

  for (Y_t) an auxiliary centered second order stationary process.

- *The* cross-correlation function*:*

$$\rho_{XY}(h) = \frac{\gamma_{XY}(h)}{\sqrt{\gamma_X(0)\gamma_Y(0)}}$$

  for (Y_t) an auxiliary centered second order stationary process.

The sequences $(\gamma_X(h))_{h\in\mathbb{Z}}$ or $(\rho_X(h))_{h\in\mathbb{Z}}$ completely determine the second order properties of a second order stationary process (X_t).

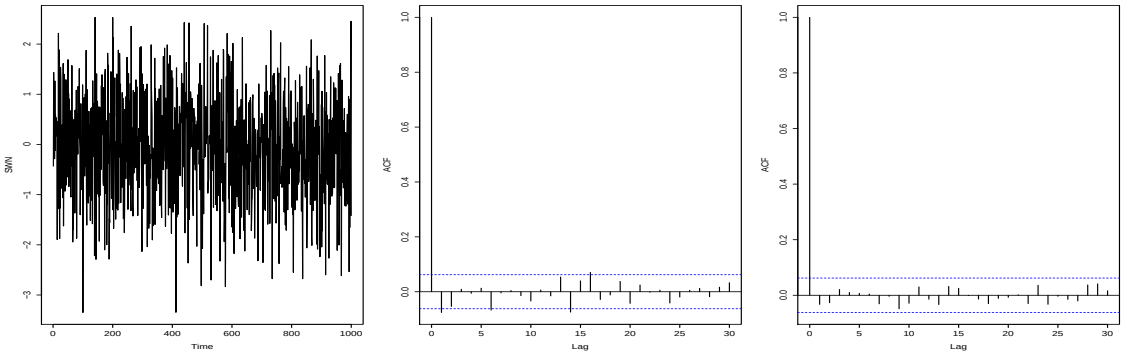

Figure 1.6: A trajectory and the corresponding ACF of a SWN and its squares

Remark.

- *We can restrict ourselves to* $\mathbb{N}$*, as* $\forall h \in \mathbb{Z}, \quad \gamma_X(h) = \gamma_X(-h)$.
- $\gamma_X(0) = Var(X_t)$ *and* $\rho_X(0) = 1$.

Example 7. *If* (X_t) *is a SWN with* $X_0 \sim P$*, then* (X_t) *is stationary and* $(X_t, \dots, X_{t+k}) \sim P^{\otimes(k+1)}$*. Moreover* $\gamma_X(0) = Var(X_t) = \sigma^2$ *exists and* X_t *is also weak-sense (second order) stationary and* $\gamma_X(h) = 0$ *for* $h \geq 1$*. We denote SWN*(σ^2)*.*

Definition 8. *A* (weak) white noise *is a second order stationary processus* (X_t) *such that:*

$$\mu_X = \mathbb{E}[X_t] = 0 \text{ and } \gamma_X(h) = \begin{cases} \sigma^2 \text{ if } h = 0 \\ 0 \text{ otherwise} \end{cases}$$

We denote $(X_t) \in WN(\sigma^2)$*.*

1.2.2 Linear time series

We have the following definition

Definition 9. *A time series is linear if it can be written as the output of a linear filter applied to a WN: let* (Z_t) *be WN and* (ψ_j) *be a linear filter, i.e. a series of deterministic coefficients such that* $\sum_{j\in\mathbb{Z}} \psi_j^2 < \infty$*, then* $X_t = \sum_{j\in\mathbb{Z}} \psi_j Z_{t-j}$*,* $j \in \mathbb{Z}$*, is a centered linear time series.*

We have to prove the existence of the infinite series $\sum_{j\in\mathbb{Z}} \psi_j Z_{t-j}$. Actually, it derives from the existence of second order moments which is a by-product of the following result:

Proposition. *Let* (Z_t) *be WN*(σ^2) *and* $\sum_{j\in\mathbb{Z}} \psi_j^2 < \infty$*, then* $X_t = \sum_{j\in\mathbb{Z}} \psi_j Z_{t-j}$*,* $j \in \mathbb{Z}$*, is a second order stationary time series satisfying*

$$\gamma_X(h) = \sigma^2 \sum_j \psi_{j+h}\psi_j \,.$$

Proof. By bilinearity:

$$\begin{aligned}\mathrm{Cov}\left(X_{t+h}, X_t\right) &= \mathrm{Cov}\left(\sum_j \psi_j Z_{t-j+h}, \sum_i \psi_i Z_{t-i}\right) \\ &= \sum_j \sum_i \psi_j \psi_i \mathrm{Cov}\left(Z_{t+h-j} Z_{t-i}\right) \\ &= \sum_j \sum_i \psi_j \psi_i \gamma_Z\left(h-j+i\right) \\ &= \sum_l \sum_j \psi_{j+l+h} \psi_j \gamma_Z(l) \\ &= \sigma^2 \sum_j \psi_{j+h} \psi_j < \infty\end{aligned}$$

where the finiteness follows by Cauchy-Schwartz inequality. In particular

$$\gamma_X(0) = \sigma^2 \sum_j \psi_j^2 = \mathbb{E}\left[\left(\sum_j \psi_j Z_{t-j}\right)^2\right] < \infty.$$

Moreover, by dominated convergence, the series $\sum_{|j|\geq k} \psi_j Z_{t-j}$ converges absolutely in $\mathbb{L}^2$ to $X_t = \sum_{j\in\mathbb{Z}} \psi_j Z_{t-j}$ that is finite a.s.. □

1.2.3 Hilbert spaces, projection and the Wold theorem

It is natural to consider projections in the Hilbert space $\mathbb{L}^2(\mathbb{P})$ when studying second order stationary time series (X_t). Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space.

Definition 10. *The set of all random variables $X : \Omega \longrightarrow \mathbb{R}$ such that $\mathbb{E}[X^2] = \int_\Omega X(w)^2 dP(w) < +\infty$ is denoted by $\mathcal{L}^2(\mathbb{P})$.*

Definition 11. *The inner product associated to $\|X\| = \sqrt{\mathbb{E}[X^2]}$ is $\langle X_1, X_2\rangle = \mathbb{E}[X_1 X_2]$.*

Proposition. *$\langle\cdot,\cdot\rangle$ has the following properties:*

- *Bilinearity: $\langle \alpha X_1, \beta X_2\rangle = \alpha\beta\langle X_1, X_2\rangle$*
- *Symmetric: $\langle X_2, X_1\rangle = \langle X_1, X_2\rangle$*
- *Non-negative: $\langle X, X\rangle \geq 0$ and $\langle X, X\rangle = 0$ iff $X = 0$ a.s.*
- *$\|\cdot\|$ is a seminorm: $\|X_1 + X_2\| \leq \|X_1\| + \|X_2\|$ and $\|\alpha X\| = |\alpha| \|X\|$.*

Definition 12. *We denote by $\mathbb{L}^2(\mathbb{P})$ the quotient space $\mathcal{L}^2(\mathbb{P})/\sim$ with $X \sim Y$ iff $X = Y$ a.s.*

Proposition. *$\left(\mathbb{L}^2(\mathbb{P}), ||\cdot||\right)$ is a Hilbert space, a vector space such that the norm induced by the inner product turns into a complete metric space.*

Definition 13. *Any $X, Y \in \mathbb{L}^2(\mathbb{P})$ are orthogonal if $< X, Y >= 0$ and are denoted $X \perp Y$. Two subsets $\mathcal{F}$ and $\mathcal{G}$ are orthogonal if $X \perp Y$ for any $X \in \mathcal{F}$ and $Y \in \mathcal{G}$.*

Theorem (Projection)**.** *Let L be a linear sub-space closed in $\mathbb{L}^2(\mathbb{P})$. Then for any $X \in \mathbb{L}^2(\mathbb{P})$ the minimizer of $Y \in L \to \|X - Y\|^2$ exists, is unique and is denoted $P_L(X)$. Moreover $P_L(X) \in L$ and $X - P_L(X) \perp L$ and these 2 relations characterize completely $P_L(X)$, the projection of X onto L.*

Notice that by orthogonality we have the Pythagorean theorem: for $Y \in L$

$$\|X - Y\|^2 = \|X - P_L(X)\|^2 + \|P_L(X) - Y\|^2.$$

The Projection theorem has nice probabilistic interpretations. For $\mathbb{P}$ being the distribution of the WN (Z_t), we identify the second order stationary time series (X_t) as measurable functions $X_t \in \mathbb{L}^2(\mathbb{P})$. Moreover $< X, Y >= \mathbb{E}[XY] = \mathrm{Cov}(X, Y)$ if X and Y are centered. Thus, being orthogonal means being uncorrelated.

Let $\mathcal{A}_0$ be a sub-σ algebra of $\mathcal{A}$ and let L be the set of r.v. that are $\mathcal{A}_0$-measurable and square integrable. Then L is a closed linear sub-space.

Definition 14. *The projection $P_L(X)$ is called the conditional expectation of X on $\mathcal{A}_0$ and is denoted $P_L(X) = \mathbb{E}[X \mid \mathcal{A}_0]$.*

When $\mathcal{A}_0$ is the σ algebra generated by some r.v. Y then we also write $\mathbb{E}[X \mid \mathcal{A}_0] = \mathbb{E}[X \mid Y]$. By the Theorem on the projection, we have that $\mathbb{E}[X \mid Y]$ is square integrable and that $\mathbb{E}[(X - \mathbb{E}[X \mid Y])h(Y)] = 0$ for any measurable and square integrable function h.

1.2.4 Best linear prediction

Let $X_1, \ldots, X_n$ be the n first observations of a second order stationary time series (X_t) that is centered.

Definition 15. *The best prediction at time n is $P_n(X_{n+1}) = \mathbb{E}[X_{n+1} \mid X_n, \ldots X_1]$. It is the measurable function f of the observation minimizing the quadratic risk (of prediction) $R_{n+1} = \mathbb{E}[(X_{n+1} - f(X_n, \ldots, X_1))^2]$.*

One can also think of the projection on the closed subset L of linear combinations of $X_1, \ldots, X_n$ called the span of the observations. One always has $L \subset \sigma(X_1, \ldots, X_n)$ and we define

Definition 16. *The best linear prediction at time n is $\Pi_n(X_{n+1}) = P_L(X_{n+1})$. It is the linear function f of the observation minimizing the quadratic risk (of prediction) $R^L_{n+1} = \mathbb{E}[(X_{n+1} - f(X_n, \ldots, X_1))^2]$.*

By definition, one has $R^L_{n+1} \geq R_{n+1}$ and $R^L_n \geq R^L_{n+1}$ because of the second order stationarity and the linearity of f

$$R^L_{n+1} = \mathbb{E}[(X_{n+1} - f(X_n, \ldots, X_2))^2].$$

Thus (R^L_n) is a converging sequence with non-negative limit denoted R^L_∞. Moreover $\Pi_n(X_{n+1}) = \theta_1 X_n + \cdots + \theta_n X_1$ and $\mathrm{Cov}(X_{n+1} - \Pi_n(X_{n+1}), X_k) = 0$ for all $1 \leq k \leq n$. Actually, the two last properties completely determine the best linear prediction. We can write down these equations in the matrix form, dividing by $\gamma_X(0)$:

Definition 17. *The system of n equations on the covariances defining the coefficients of the best linear prediction is called the Yule-Walker system and it is equal to*

$$\begin{pmatrix} \rho_X(0) & \cdots & \rho_X(n-1) \\ \vdots & \ddots & \vdots \\ \rho_X(n-1) & \cdots & \rho_X(0) \end{pmatrix} \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_n \end{pmatrix} = \begin{pmatrix} \rho_X(1) \\ \vdots \\ \rho_X(n) \end{pmatrix}$$

Denoting $\mathbb{X} = (X_n, \dots, X_1)$ one can write in a compact way the best linear predictor

$$\Pi_n(X_{n+1}) = \mathbb{X}\theta = \mathbb{X}\mathbb{E}[\mathbb{X}^\top \mathbb{X}]^{-1}\mathbb{E}[\mathbb{X}^\top X_{n+1}].$$

The Yule-Walker method is based on the compact formula, requiring to invert a covariance matrix at each step n. Other procedures can compute this explicit formula in an efficient way, i.e. avoiding to invert the covariance matrix of the observations $(X_1, \dots, X_n)'$.

The matrix of variance-covariance of the observation $(\rho_X(i-j))_{1\leq i,j\leq n}$ is a Toeplitz symmetric semi-definite matrix with diagonal dominant terms 1. It is not definite only if it exists a deterministic vector $u \neq 0$ in its kernel such that

$$0 = u^\top(\gamma_X(i-j))_{1\leq i,j\leq n}u = \mathbb{E}[u^\top \mathbb{X}\mathbb{X}^\top u] = \mathbb{E}[(u^\top \mathbb{X})^2] = 0,$$

where $\mathbb{X} = (X_1, \dots, X_n)^\top$. Thus $u^\top \mathbb{X} = 0$ a.s. and X_n expresses as a linear combination of the past values $X_1, \dots, X_{n-1}$. In particular $\Pi_{n-1}(X_n) = X_n$ a.s. and $R^L_{n+k} = 0$ for all $k \geq 0$. More generally, in any cases where $R_\infty = 0$ one says that the second order stationary time series (X_t) is *deterministic*. For instance $X_t = X$ for all $t \in \mathbb{Z}$, where X is a random variable, is deterministic. There are other example of deterministic (but random) time series:

Example 8. *Let A and B two random variables such that $Var(A) = Var(B) = \sigma^2$, $\mathbb{E}(A) = \mathbb{E}(B) = 0$ and $Cov(A, B) = 0$. Let $\lambda \in \mathbb{R}$. We define the following trigonometric sequence:*

$$X_t = A\cos(\lambda t) + B\sin(\lambda t)$$

Then (X_t) is weak-sense stationary as $\mu_X = \mathbb{E}(X_t) = 0$ and

$$\begin{aligned}
\gamma_X(h) &= Cov\left(X_t, X_{t+h}\right) \\
&= Cov\left[A\cos\left(\lambda t\right) + B\sin\left(\lambda t\right), A\cos\left(\lambda(t+h)\right) + B\sin\left(\lambda(t+h)\right)\right] \\
&= \cos\left(\lambda t\right)\cos\left(\lambda(t+h)\right)\sigma^2 + \sin\left(\lambda t\right)\sin\left(\lambda(t+h)\right)\sigma^2 \\
&= \sigma^2\cos(\lambda h)
\end{aligned}$$

Although A and B are random variables, the process (X_t) is deterministic.

1.2.5 The innovations and the Wold theorem

Let us introduce the following notion

Definition 18. *The innovation at time n is the error of linear prediction $I_n = X_n - \Pi_{n-1}(X_n)$.*

So, by definition the innovations are centered and their variances are equal to R^L_n. In general, the innovations are not stationary as R^L_n decreases with n. We have the following simple decomposition

Proposition. *The linear projection Π_{n+1} can be decomposed into the sum of two projection*

$$\Pi_{n+1} = \Pi_n + P_{I_{n+1}}, \qquad n \geq 1,$$

where $P_{I_{n+1}}$ is the projection on the linear span of the innovation I_{n+1}.

Proof. The proof is based on the orthogonal decomposition of L_{n+1} the linear span of $(X_1, \dots, X_{n+1})$ as the linear span L_n of $(X_1, \dots, X_n)$ and the linear span of I_{n+1}. Indeed, by definition of $I_{n+1} \in L_{n+1}$ we have $I_{n+1} \perp L_n$. We conclude by a dimension argument, as the dimension of L_{n+1} is $n+1$ and so the orthogonal complement of L_n of dimension n is a span of dimension 1. □

In particular the innovations (I_n) are uncorrelated.

Let us describe the asymptotic behaviour of the innovations. To do so, it is useful to use a backward argument; one observes $(X_{-1}, \dots, X_{-n+1})$ and we try to predict X_0 for all $n \geq 1$. We denote $\Pi_{-n}(X_0)$ the corresponding best linear prediction. By second order stationarity, we have

$$R_n^L = \mathbb{E}[(X_0 - \Pi_{-n}(X_0))^2], \qquad n \geq 1.$$

Moreover $X_0 - \Pi_{-n}(X_0)$ is orthogonal to the span of $(X_{-1}, \dots, X_{-n+1})$. By orthogonality of $\Pi_{-n+k}(X_0)$ and $\Pi_n(X_0)$ with $\Pi_n(X_0) - X_0$ for $1 \geq k \geq n$ we have

$$\begin{aligned}\mathbb{E}[(\Pi_{-n+k}(X_0) - \Pi_{-n}(X_0))^2] &= R_{n-k}^L + R_n^L + 2\mathbb{E}[(\Pi_{-n+k}(X_0) - X_0)(\Pi_{-n}(X_0) - X_0)] \\ &= R_{n-k}^L + R_n^L - 2\mathbb{E}[X_0(\Pi_{-n}(X_0) - X_0)] \\ &= R_{n-k}^L - R_n^L.\end{aligned}$$

Thus as (R_n^L) is converging, it is a Cauchy sequence and so is $(\Pi_{-n}(X_0))$ in $\mathbb{L}^2(\mathbb{P})$. Thus $\Pi_{-n}(X_0)$ converges and one denotes $\Pi_\infty(X_0)$ its limit. Defining $I_\infty(X_0) = X_0 - \Pi_\infty(X_0)$, we have the identity

$$R_\infty^L = \mathbb{E}[I_\infty(X_0)^2].$$

Defining $\Pi_\infty(X_n)$ and $I_\infty(X_n)$ thanks to the lag operator $L^n \Pi_\infty(X_n) = \Pi_\infty(X_0)$, one can also check that

$$\mathbb{E}[(I_n - I_\infty(X_n))^2] = \mathbb{E}[(\Pi_{n-1}(X_n) - \Pi_\infty(X_n))^2] = R_n^L - R_\infty^L \to 0.$$

In particular $(I_n - I_\infty(X_n))$ converges in $\mathbb{L}^2$ to 0 and (I_n) converges in distribution to $I_\infty(X_0)$. Let us use this concept of limit innovation $I_\infty(X_n)$ in order to prove that any second order stationary time series (X_t) is the sum of a linear time series and a deterministic process:

Theorem (Wold)**.** *Let (X_t) be second order stationary. Then X_t is uniquely decompose as*

$$X_t = \sum_{j \geq 0} \psi_j I_\infty(X_{t-j}) + r_t$$

where

- $\psi_0 = 1$ *and* $\sum_{j \geq 0} \psi_j^2 < \infty$,
- $(I_\infty(X_{t-j}))$ *is a WN(R_∞^L),*
- $Cov(I_\infty(X_t), r_s) = 0$ *for all* t, $s \in \mathbb{Z}$.
- (r_t) *is deterministic.*

Proof. Let us show the 2 first assertions. By construction $(I_\infty(X_t))$ is a WN(R_∞^L). Let define

$$\psi_j = \frac{\mathbb{E}[X_t I_\infty(X_{t-j})]}{R_\infty^L}.$$

Then $\psi_0 = 1$ and $\sum_{j \geq 0} \psi_j I_\infty(X_{t-j})$ is the orthogonal projection of X_t on the span of $(I_\infty(X_{t-j}))_{j \geq 0}$ by the use of the previous Proposition and a recursive argument. Thus $\mathbb{E}[(\sum_{j \geq 0} \psi_j I_\infty(X_{t-j}))^2] = \sum_{j \geq 0} \psi_j^2 < \infty$. □

The Wolds's representation motivates the following definition

Definition 19. *The linear time series is* causal *iff $\psi_j = 0$, $j < 0$.*

Remark that from Wold's representation, any second order stationary time series that has no deterministic component admits a causal linear representation.

Part II

Models and estimation

Chapter 2

ARMA models

Assume that after preprocessing the data one obtains (X_t) that are second order stationary without deterministic component: by Wold's representation, (X_t) admits a causal linear representation

$$X_t = \sum_{j\geq 0} \psi_j Z_{t-j}, \qquad t \geq 1,$$

where (Z_t) is some WN(σ^2) and $\sum_j \psi_j^2 < \infty$. This linear setting motivates the use of the best linear prediction

$$\Pi_n(X_{n+1}) = \theta_1 X_n + \cdots + \theta_n X_1$$

as the associated error of prediction $I_{n+1} = X_{n+1} - \Pi_n(X_{n+1}) =: I_{n+1}(X_{n+1})$ converges in distribution. Indeed $I_{n+1}(X_0)$ converges in $\mathbb{L}^2$ to Z_0 that we identify as $I_\infty(X_0)$. As the best linear prediction of a WN is 0, the WN is considered as *linearly unpredictable* and $\sigma^2 = R_\infty^L$ is the smallest poxssible risk of prediction in our context.

However, it is not reasonable to try to estimate n coefficients from n observations $(X_1, \ldots, X_n)$ as $\theta = (\theta_1, \ldots, \theta_n)$ requires the knowledge of $(\rho_X(h))_{0\leq h\leq n}$ through the Yule-Walker equation, and these correlations are unknown. Usually, one estimate the autocorrelations empirically:

Definition 20. *The empirical autocorrelation is defined as*

$$\widehat{\rho}_X(h) = \frac{\sum_{t=1}^{n-h} X_t X_{t+h}}{\sum_{t=1}^{n} X_t^2}, \qquad 0 \leq h \leq n-1.$$

Notice that by definition, we have the following properties

- $|\widehat{\rho}_X(h)| \leq 1$ for the same reason than $|\rho_X(h)| \leq 1$: Cauchy-Schwartz inequality,
- $\widehat{\rho}_X(h)$ is likely biased, i.e. $\mathbb{E}[\widehat{\rho}_X(h)] \neq \rho_X(h)$.

In practice, one would like to test wether $\rho_X(h) = 0$ from the estimator $\widehat{\rho}_X(h)$. It is possible under the strong assumption, uncheckable, that (X_t) is a SWN.

Theorem. *If (X_t) is a SWN then $\widehat{\rho}_X(h)$ converges (a.s) to $\rho_X(h)$ if h is fixed and $n \to \infty$ and in this case, for any $h \geq 1$, we have*

$$\sqrt{n}\widehat{\rho}_X(h) \xrightarrow{d.} \mathcal{N}(0,1).$$

Proof. We want to apply the CLT on (X_tX_{t+h}). It has finite variance $\gamma_X(0)^2$ because of the independence assumption. Also (X_tX_{t+h}) is independent of (X_sX_{s+h}), $s > t$, except for $s = t + h$ but then

$$\mathrm{Cov}(X_tX_{t+h}, X_{t+h}X_{t+2h}) = \mathbb{E}[X_tX_{t+h}^2X_{t+2h}] = 0.$$

Thus one can prove that

$$\sqrt{n}\widehat{\gamma}_X(h) \xrightarrow{d.} \mathcal{N}(0, \gamma_X(0)^2)$$

where $\widehat{\gamma}_X(h) = (n-h)^{-1}\sum_{t=1}^{n-h} X_tX_{t+h}$ is the unbiased empirical estimator of $\gamma_X(h)$. The result also holds for $h = 0$:

$$\sqrt{n}(\widehat{\gamma}_X(0) - \gamma_X(0)) \xrightarrow{d.} \mathcal{N}(0, \gamma_X(0)^2)$$

which implies that $\widehat{\gamma}_X(0) \xrightarrow{\mathbb{P}.} \gamma_X(0)$. We conclude the proof applying Slutsky's theorem. □

The blue dotted band observed in Figure 1.2.1 corresponds to the interval $\pm 1.96/\sqrt{n}$. If the coefficient $\widehat{\gamma}_X(h)$ is outside the band, one can reject with asymptotic confidence rate 95% the hypothesis that (X_t) is a strong white noise. The asymptotic is reasonable when $n - h$ is large because the correct normalisation should be $\sqrt{n-h}$ in the result above and not $\sqrt{n}$ (asymptotically equivalent when h is fixed). On the contrary, there does not exists any converging estimator of $\rho_X(n-h)$ for any h fixed, even when n tends to infinity. (a fortiori $\rho_X(n)$ as we never observed data delayed by n).

As it is unrealistic to estimate n parameters from n observations, we will use a sparse representation of the linear process (X_t):

Definition 21. *An ARMA(p,q) time series is a solution (if it exists) of the model*

$$X_t = \phi_1X_{t-1} + \cdots + \phi_pX_{t-p} + Z_t + \gamma_1Z_{t-1} + \cdots + \gamma_qZ_{t-q}, \qquad t \in \mathbb{Z},$$

with $\theta = (\phi_1, \ldots, \phi_p, \gamma_1, \ldots, \gamma_q)' \in \mathbb{R}^{p+q}$ the parameters of the model and (Z_t) WN(σ^2).

2.1 Moving Averages (MA time series)

The moving average is the simplest sparse representation of the infinite series in the causal representation $X_t = \sum_{j\geq 0}\psi_jZ_{t-j}$ consisting in assuming $\psi_j = 0$ for $j > q$.

Definition 22. *A MA(q), $q \in \mathbb{N} \cup \{\infty\}$ process is a solution to the equation:*

$$X_t = Z_t + \gamma_1Z_{t-1} + \cdots + \gamma_qZ_{t-q}, \qquad t \in \mathbb{Z}.$$

Notice that we extend the notion to the cases where $q = \infty$ so that any causal linear time series satisfies a MA(∞) model.

Example 9. *Let (Z_t) be a WN(σ^2) and let $\gamma \in \mathbb{R}$. Then $X_t = Z_t + \gamma Z_{t-1}$ is a first order moving average, denoted by MA(1). (X_t) is second order stationary because $\mathbb{E}[X_t] = 0$ and*

$$\gamma_X(h) = Cov(Z_t + \gamma Z_{t-1}, Z_{t+h} + \gamma Z_{t+h-1}) = \begin{cases} (1+\gamma^2)\sigma^2 & \text{if } h = 0, \\ \gamma\sigma^2 & \text{if } h = \pm 1, \\ 0 & \text{else.} \end{cases}$$

In general, we have the following very useful property

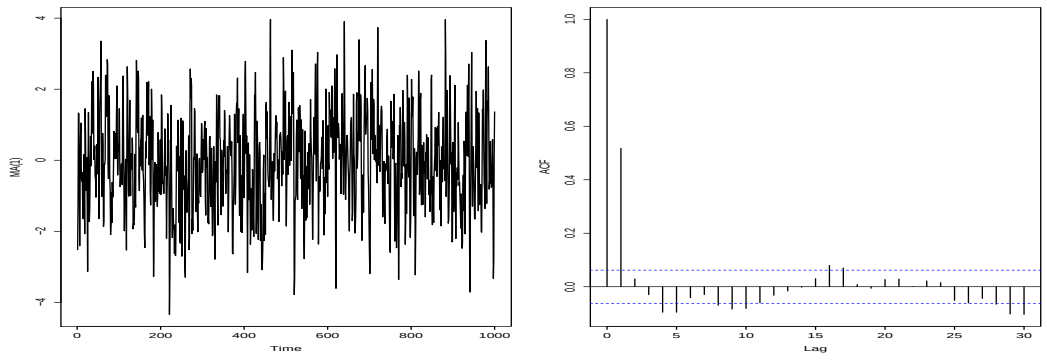

Figure 2.1: A trajectory and the corresponding ACF of the solution of an MA(1) model

Proposition. *If (X_t) satisfies a MA(q) model, we have $\gamma_X(h) = 0$ for all $h > q$.*

Remark.

- *X_t and X_s are uncorrelated as soon as $|t - s| \geq q + 1$.*
- *If Z_t is a SWN(σ^2), then (X_t) is stationary.*
- *More precisely, a MA(q) model is a q-dependent stationary time series when (Z_t) is SNW: X_t and X_s are independent as soon as $|t - s| \geq q + 1$.*

As shown in Figure 2.1, the uncorrelated property is used in practice to estimate the order q of an MA(q); corresponding to the last component which is significantly non-null, i.e. outside the blue confident band (only valid if (Z_t) is a SWN).

2.2 Auto-Regressive models (AR time series)

The second sparse representation is the AR(p) model.

Definition 23. *The time series (X_t) satisfies an AR(p) model, $p \in \mathbb{N} \cup \{\infty\}$, iff it is solution of the equation*

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t, \qquad t \in \mathbb{Z}.$$

It is not sure that it represents a causal linear time series.

Example 10. *Let (Z_t) be a WN(σ^2) and $X_t = \phi X_{t-1} + Z_t$, for $t \in \mathbb{Z}$ (AR(1) process). As we have no initial condition the recurrence equation does not ensure the existence of (X_t). If $|\phi| < 1$, then by iterating the equation we get:*

$$X_t = \phi^k X_{t-k} + \phi^{k-1} Z_{t-k-1} + \cdots + \phi Z_{t-1} + Z_t$$

If a second order stationary solution (X_t) exists, then:

$$\mathbb{E}\left[\left(\phi^k X_{t-k}\right)^2\right] = \phi^{2k}\mathbb{E}\left[X_0^2\right] \underset{k\to+\infty}{\longrightarrow} 0\,.$$

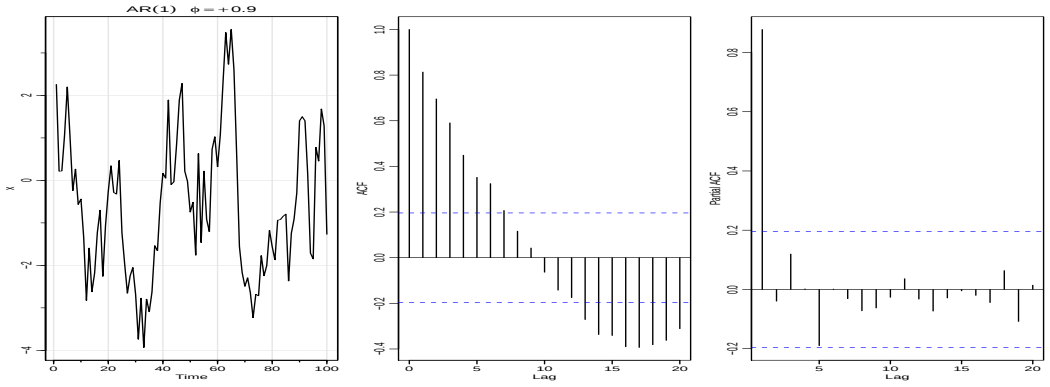

Figure 2.2: A trajectory and the corresponding ACF and PACF of the solution of an AR(1) model

A solution admits a MA(∞) representation:

$$X_t = \sum_{j=0}^{+\infty} \phi^j Z_{t-j}$$

which exists as $\sum_{j=0}^{+\infty} |\phi^j|^2 < \infty$. *We easily check that this representation satisfies* $X_t = \phi X_{t-1} + Z_t$. *Remark that* $\gamma_X(0) = \sigma^2/(1-\phi^2)$ *and that* $\rho_X(h) = \phi^h$, $h \geq 0$.

Remark. *If* $\phi = 1$, *by iterating we get the random walk* $X_t = X_0 + Z_1 + \cdots + Z_t$ *and* $Var(X_t - X_0) = t\sigma^2 \underset{t\to+\infty}{\longrightarrow} +\infty$ *so the random walk is not second order stationary. This course is restricted to the stationary case as the random walk can be preprocessed thanks to differencing.*

We saw that for a MA(1) process, $\gamma_X(h) = 0$ for $h \geq 2$. Here we always have $\gamma_X(h) \neq 0$, $h \geq 0$ see Figure 2.2. Thus, it is not possible to use the ACF to infer the order p of an AR(p) model.

The notion for inferring the order of auto-regression is the partial autocorrelation

Definition 24. *The partial autocorrelation of order* h *is defined as (under the convention* $\Pi_0(X_1) = 0$*)*

$$\widetilde{\rho}_X(h) = \rho_X(X_0 - \Pi_{h-1}(X_0), X_h - \Pi_{h-1}(X_h)), \qquad h \geq 1$$

where $\Pi_{h-1}(X_0)$ *is the projection of* X_0 *on the linear span of* $(X_1, \ldots, X_{h-1})$.

By definition $\widetilde{\rho}_X(1) = \rho_X(1)$. The partial autocorrelations are used to determine graphically the order of an AR(p) model. Indeed, we have

Proposition. *The PACF of a causal AR(p) model (a solution with the expression* $\sum_{j\geq 0} \psi_j Z_{t-j}$ *exists) satisfies* $\widetilde{\rho}_X(h) = 0$ *for all* $h > p$.

Proof. Indeed, for an AR(p) time series we have $\Pi_{h-1}(X_h) = \phi_1 X_{h-1} + \cdots + \phi_p X_{h-p}$ so that

$$\widetilde{\rho}_X(h) = \rho_X(X_0 - \Pi_{h-1}(X_0), Z_h) = 0, \qquad h > p.$$

□

Fortunately, the PACF can be estimated from the Yule-Walker equation using only the h first empirical estimators of the correlations $\widehat{\rho}_X(i)$ for $1 \leq i \leq h$.

2.3 Existence of a causal second order stationary solution of an ARMA model

As for the AR(1) model, some conditions have to be done on the coefficients of the autoregressive part such that the solution can be written as a linear filter

$$X_t = \sum_j \psi_j Z_{t-j}, \qquad t \in \mathbb{Z}.$$

Recall that L defines the lag operator such that $LX_t = X_{t-1}$ and $L^k X_t = X_{t-k}$. One can now rewrite the ARMA model in a compact form

$$\phi(L)X_t = \gamma(L)Z_t\,,$$

where (Z_t) is a WN(σ^2) and the lag polynomials

$$\begin{aligned}\phi(z) &= 1 - \phi_1 z - \cdots - \phi_p z^p,\\ \gamma(z) &= 1 + \gamma_1 z + \cdots + \gamma_q z^q, \qquad z \in \mathbb{C}.\end{aligned}$$

We need to use complex analysis to solve the equation $\phi(L)X_t = \gamma(L)Z_t$ as $X_t = \gamma(L)/\phi(L)Z_t = \psi(L)Z_t$.

Definition 25. *A Laurent series is a function $\mathbb{C} \mapsto \mathbb{C}$ that can be written as $\psi(z) = \sum \psi_j z^j$ where the range of the summation is $j \in \mathbb{Z}$.*

If $\sum |\psi_j| < \infty$, as $\sum \psi_j^2 < \infty$ then $\psi(L)Z_t$ is a linear time series. The behavior of the Laurent series on $S = \{z \in \mathbb{C}, |z| = 1\}$ is crucial for the analysis of the existence of a filter.

Proposition. *1. Assume that $\sum |\psi_{1,j}| < \infty$ and $\sum |\psi_{2,j}| < \infty$. Then the series $\psi_i(z) = \sum_{j\in\mathbb{Z}} \psi_{i,j} z^j$ are well defined on S and*

$$\psi_1(z)\psi_2(z) = \psi_2(z)\psi_1(z) = \sum_{k\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} \psi_{1,j}\psi_{2,k-j} z^k$$

is also well defined on S,

2. On the converse, if $\psi(z)$ is defined on any enlargement of S then $\sum |\psi_{1,j}| < \infty$.

We are now ready to state

Theorem. *If ϕ do not have roots on S then the ARMA model admits a solution*

$$X_t = \frac{\gamma(L)}{\phi(L)} Z_t = \psi(L)Z_t, \qquad t \in \mathbb{Z}$$

and (X_t) is a causal linear time series.

Recall the notion of causality, meaning here that the process (X_t) is a linear transformation of the past WN $(Z_t, Z_{t-1}, \ldots)$. Here, it is equivalent to assert that $\psi_j = 0$ for $j < 0$ and so that the Laurent series is defined on $D = \{z \in \mathbb{C}, |z| \leq 1\}$ since any factor $|z|^j$, $j < 0$, diverges when $|z| \to 0$. As $\psi(z) = \gamma(z)/\phi(z)$ should be defined on D, it means that ϕ does not have roots inside D. So we have the following result

Proposition. *The solution of an ARMA model is*

1. *causal iff ϕ does not have roots inside D, then X_t admits a Wold representation $\sum_{j\geq 0}\psi_j Z_{t-j}$ and the WN Z_t are the limit innovations,*
2. *(linearly) invertible, i.e. $X_t = \sum_{j=1}^{\infty}\varphi_j X_{t-j} + Z_t$ iff γ does not have roots inside D. In particular $\Pi_\infty(X_t) = \sum_{j=1}^{\infty}\varphi_j X_{t-j}$, $t\in\mathbb{Z}$ as soon as (X_t) is second order stationary.*

Proof. We only prove the sufficiency assertion.

Let $z_1,\dots,z_p$ the roots of ϕ (in $\mathbb{C}$). They are non null since $\phi(0)=1$. We can write

$$\phi(z) = \prod_{i=1}^{p}(1 - z_i^{-1}z)$$

Thus, assuming the roots are simple for simplicity, we get the partial fraction decomposition

$$\begin{aligned}\frac{1}{\phi(z)} &= \frac{1}{\prod_{i=1}^{p}(1-z_i^{-1}z)}\\ &= \sum_{i=1}^{p}\frac{a_i}{(1-z_i^{-1}z)}\\ &= \sum_{i=1}^{p}a_i\sum_{j=0}^{\infty}(z_i^{-1}z)^j\\ &= \sum_{j=0}^{\infty}\Big(\sum_{i=1}^{p}a_i(z_i^{-j})\Big)z^j\end{aligned}$$

It is a Laurent series with positive coefficients only and defined over an enlargement of D for any $|z|\leq|z_i|$, $1\leq i\leq p$. Since $\gamma(z)$ has only positive coefficients and defined on $\mathbb{C}$, the product $\psi(z)=\gamma(z)/\phi(z)$ has also positive coefficients and is defined on an enlargement of D. In particular $\psi(L)Z_t$ is a causal linear filter.

The second assertion holds for the same reason than the first one as $\varphi(z) = 1 - \gamma(z)^{-1}\phi(z)$ satisfying $\varphi(0)=\varphi_0=0$. □

Moreover, since ψ and φ are necessarily defined on an enlargement of D, we also have

Proposition. *If the ARMA model is causal or invertible then there exists $C>0$ and $0,\rho<1$ so that $|\psi_j|\leq C\rho^j$ or $|\varphi_j|\leq C\rho^j$, respectively.*

In particular, an ARMA process is a sparse representation of a linear model that models only exponential decaying auto-covariance processes because $\gamma_X(h) = \sigma^2\sum_{j=0}^{\infty}\psi_j\psi_{j+h} = O(\rho^j)$.

Despite potentially infinitely many non-null correlations, (X_t) is said to be (short memory) weakly dependent because $|\gamma_X(h)| \underset{k\to+\infty}{\longrightarrow} 0$ and the decrease is exponential:

$$\exists c>0,\ \rho\in\,]0,1[\,,\ \forall h\in\mathbb{N},\ |\gamma_X(h)| < c\rho^h$$

There are long memory processes (or strongly dependent): $|\gamma_X(h)|\sim h^{-a}$, $a>1/2$.

Example 11. *Any ARMA(p,q) model has auto-correlations that will ultimately decrease to 0 exponentially fast.*

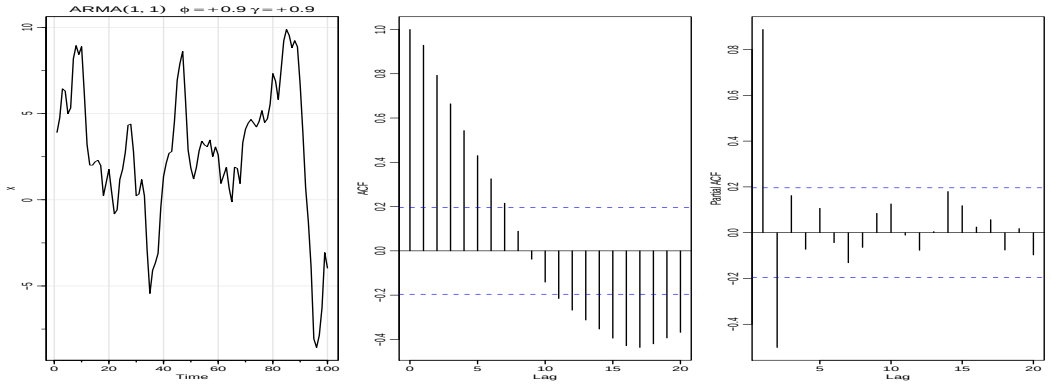

Figure 2.3: A trajectory and the corresponding ACF and PACF of the solution of an ARMA(1,1) model with $\phi 1 = \gamma_1 = 0.9$.

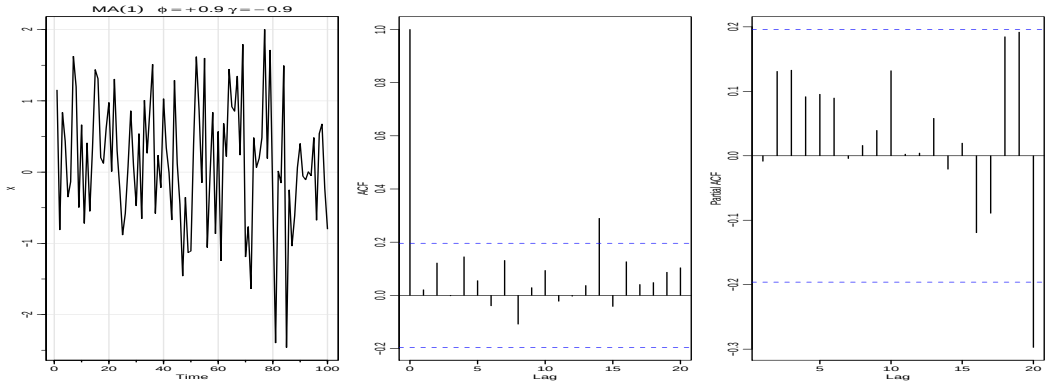

Figure 2.4: A trajectory and the corresponding ACF of the solution of an ARMA(1,1) model with $\phi_1 = -\gamma_1 = 0.9$

Notice that the ARMA(p, q) representation faces the following problem of sparsity, when $pq \neq 0$; if there is a common root for the two polynomials ϕ and γ, let us say z_0 with $|z_0| \neq 1$, then $(z_0 - L)$ is inveritble and the ARMA$(p-1, q-1)$ model

$$(1 - z_0^{-1}L)^{-1}\phi(L)X_t = (1 - z_0^{-1}L)^{-1}\gamma(L)Z_t$$

defines the same linear time series than the original ARMA(p, q) model. This problem is crucial in statistics as the ACF or the PACF do not yield any information on how to choose the orders p and q.

Example 12. *An ARMA$(1, 1)$ with $\phi_1 = -\gamma_1$ is equivalent to a WN. To see this, one checks that the root of the polynomial $\phi(z) = 1 - \phi_1 z$ is the same than the root of $\gamma(z) = 1 + \gamma_1 z$.*

To solve this issue, one uses penalized Quasi Maximum Likelihood approach.

Chapter 3

Quasi Maximum Likelihood for ARMA models

The estimation of the parameter $\theta = (\phi_1, \dots, \phi_p, \gamma_1, \dots, \gamma_q) \in \mathbb{R}^{p+q}$ will be done following the Maximum Likelihood principle. The important concept is the likelihood, i.e. the density of the f_θ of the sample $(X_1(\theta), \dots, X_n(\theta))$ that follows the ARMA(p, q) model with the corresponding $\theta \in \mathbb{R}^{p+q}$.

Definition 26. *The log-likelihood $L_n(\theta)$ is defined as*

$$L_n(\theta) = -2 \log(f_\theta(X_1, \dots, X_n)).$$

The Quasi-Likelihood criterion (QLik) is the log-likelihood when the WN (Z_t) of the model $(X_1(\theta), \dots, X_n(\theta))$, is gaussian WN($\sigma^2$). The Quasi Maximum Likelihood Estimator (QMLE) satisfies

$$\widehat{\theta}_n = \arg\min_{\theta \in \Theta} L_n(\theta),$$

for some admissible parameter region $\Theta \subset \mathbb{R}^{p+q}$.

The concept of QLik is fundamental in these notes. As we will see, the gaussian assumption on (Z_t) is made for the ease of calculating the criterion. It is not the Likelihood, i.e. we do not believe that the observations (X_t) satisfies the ARMA(p, q) model with gaussian noise. It is a convenient way to define a contrast L_n to minimize on Θ. Notice that we do not consider the variance of the noise of the model σ^2 as an unknown parameter. The procedure will automatically provide an estimator of this variance, as an explicit function of the QMLE $\widehat{\theta}_n$.

3.1 The QML Estimator

3.1.1 Gaussian distribution

Definition 27. *A random variable N is gaussian standard if its density is equal to $(2\pi)^{-1/2} e^{x^2/2}$, $x \in \mathbb{R}$. We will denote the distribution $\mathcal{N}(0, 1)$.*

Then X is symmetric, $\mathbb{E}[X] = 0$ and $\mathrm{Var}\,(X) = 1$.

Definition 28. *A random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ and $\sigma > 0$, if there exists $N \sim \mathcal{N}(0, 1)$ such that $X = \mu + \sigma N$ in distribution.*

Then $\mathbb{E}[X] = \mu$ and $\mathrm{Var}\,(X) = \sigma^2$ and

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad x \in \mathbb{R}.$$

Definition 29. *Let $X_k \sim \mathcal{N}(0,1)$, $1 \le k \le n$, be iid. Then, for any $U \in \mathbb{R}^n$, any Σ a $n \times n$ symmetric matrix definite positive, the vector*

$$Y = U + \Sigma^{1/2}(X_1, \dots, X_n)' \qquad \textit{in distribution}$$

is distributed as $\mathcal{N}_d(U, \Sigma)$, the gaussian distribution of dimension d with mean U and variance Σ.

Notice that $\Sigma^{1/2}$ is the square root of Σ, i.e. the only symmetric definite positive matrix A such that $A^2 = \Sigma$. The fundamental result about gaussian random vector is the following

Proposition. *Let Y be a d-dimensional Gaussian random vector that is centered then $\mathbb{E}[Y_i Y_j] = 0$, i.e. $Y_i \perp Y_j$ for $i \neq j$ is equivalent to Y_i independent of Y_j and*

$$\mathbb{E}[Y_i | Y_{i-1}, \dots, Y_1] \in \textit{span}(Y_{i-1}, \dots, Y_1), \qquad 1 \le i \le d\,.$$

The proof is based on a characteristic functions argument. That Y_i and Y_j are gaussian centered r.v. is not enough, consider the case $Y_i = \varepsilon Y_j$ with $\mathbb{P}(\varepsilon = \pm 1) = 2^{-1}$ and ε independent of Y_j.

The proposition has several consequences. In particular, one deduces that for centered observations $X_1, \dots, X_n$ constituting a gaussian vector then $P_j = \Pi_j$, i.e. the conditional expectation is equal to the orthogonal projection.

3.1.2 The QLik loss

The Quasi-Likelihood loss for an ARMA model is computed in the following way. Consider the parameter θ of the ARMA(p,q) model as fixed. One wants to compute the density of the model $(X_1(\theta), \dots, X_n(\theta))$ that are not independent random variables. Thus, the density is not a priori a product. However, we always have

$$f_\theta(x_1, \dots, x_n) = \prod_{t=1}^{n} f_\theta(x_t \mid x_{t-1}, \dots, x_1)$$

where $f_\theta(x_t \mid x_{t-1}, \dots, x_1)$ is the density of the distribution of $X_t(\theta)$ given $X_{t-1}(\theta), \dots, X_1(\theta)$. Under the gaussian assumption, we have

Proposition. *If θ corresponds to a causal ARMA(p,q) model then the distribution of $X_t(\theta)$ given $X_{t-1}(\theta), \dots, X_1(\theta)$ is*

$$\mathcal{N}(\Pi_{t-1}(X_t(\theta)), R_t^L(\theta)), \qquad t \ge 1$$

with the conventions $\Pi_0(X_t(\theta)) = 0$ and $R_t^L(\theta) = \mathbb{E}[(X_t(\theta) - \Pi_{t-1}(X_t(\theta)))^2]$.

Proof. From the causal assumption, $(X_t(\theta))$ is a linear function of (Z_t). Thus all the distributions of $(X_{t+1}(\theta), \dots, X_{t+h}(\theta))$ for any $t \in \mathbb{Z}$ and $h \ge 1$ are gaussian. one says that $(X_t(\theta))$ is a guassian process. In particular the conditional distribution of $X_t(\theta)$ given $X_{t-1}(\theta), \dots, X_1(\theta)$ is gaussian. One has to compute the conditional expectation and the conditional variance. We already know that the conditional expectation coincides with

the best linear predictor $\Pi_{t-1}(X_t(\theta))$. Moreover, we have that the corresponding error of prediction $X_t(\theta) - \Pi_{t-1}(X_t(\theta))$ is orthognal to the past $(X_{t-1}(\theta), \dots, X_1(\theta))$. Thus, it is independent and the conditional variance

$$\begin{aligned}\mathrm{Var}\left(X_t(\theta) \mid X_{t-1}(\theta), \dots, X_1(\theta)\right) &= \mathbb{E}[(X_t(\theta) - \Pi_{t-1}(X_t(\theta)))^2 \mid X_{t-1}(\theta), \dots, X_1(\theta)] \\ &= \mathbb{E}[(X_t(\theta) - \Pi_{t-1}(X_t(\theta)))^2] \\ &= R_t^L(\theta).\end{aligned}$$

□

By definition, $\Pi_{t-1}(X_t(\theta))$ is a linear function of the past $X_{t-1}(\theta), \dots, X_1(\theta)$ depending only on θ. Let us denote by $\Pi_{t-1}(\theta)(x_t)$ the same function expresses on $x_{t-1}, \dots, x_1$. Then the density of the model expresses as

$$f_\theta(x_1, \dots, x_n) = \prod_{t=1}^{n} \frac{1}{\sqrt{2\pi R_t^L(\theta)}} e^{-(x_t - \Pi_{t-1}(\theta)(x_t))^2 / R_t^L(\theta)}.$$

The QLik criterion has the nice additive form, up to a constant

$$L_n(\theta) = \sum_{t=1}^{n} \log(R_t^L(\theta)) + \frac{(X_t - \Pi_{t-1}(\theta)(X_t))^2}{R_t^L(\theta)} + cst.$$

Note that $\Pi_{t-1}(\theta)(X_t) \neq \Pi_{t-1}(X_t(\theta))$ as the first expression uses the observations, the second one the solutions of an ARMA model with parameter θ. In the sequel, we will denote for short the innovation of the ARMA model on the observations as

$$I_t(\theta) = X_t - \Pi_{t-1}(\theta)(X_t).$$

Minimizing this criterion over the set of any possible causal models Θ, we obtain the QMLE.

3.1.3 The QMLE as an M-estimator

The QMLE is an estimator defined as the minimizer of the QLik. There is a vast literature on such class of estimators, called M-estimator.

Definition 30. *An M-estimator is a parameter $\widehat{\theta}_n$ satisfying*

$$\widetilde{\theta}_n \in \arg\min_{\Theta} L_n(\theta) = \arg\min_{\Theta} \sum_{t=1}^{n} \frac{1}{n}\, \ell_t(\theta).$$

The (random) functions ℓ_t are called the loss functions. and L_n is the contrast or cumulative loss function. The set of parameters Θ has to be chosen carefully. One convenient (and safe) way is to choose it as a compact set so that continuity of the loss functions yields the existence of the M-estimator.

Avoiding for a moment the difficult problem of calculating efficiently $(\Pi_t(\theta))$ and $(R_t^L(\theta))$ (i.e. assuming they are known), the QMLE is an M-estimator with

$$\ell_t(\theta) = -2\log(f_\theta(X_t \mid X_{t-1}, \dots, X_1)) + cst = \log(R_t^L(\theta)) + \frac{(X_t - \Pi_{t-1}(\theta)(X_t))^2}{R_t^L(\theta)}.$$

In order to deal with the convergence of M-estimator in non iid settings such as time series we need a generalization of the SLLN called the ergodic theorem.

3.1.4 Stationary ergodic time series

Stability in a stochastic setting refers to many notions. We remind here the main stability notion: the ergodicity. Recall that L denotes the lag operator $LX_t = X_{t-1}$.

Definition 31. *A set C of $\mathbb{R}^{\mathbb{Z}}$ is invariant iff $L^{-1}C = C$ and the stationary time series (X_t) is ergodic iff for all invariant sets $\mathbb{P}((X_t) \in C) = 0$ or $\mathbb{P}((X_t) \in C) = 1$.*

Ergodicity is a notion of stability because of the following theorem

Theorem (Birkhoff)**.** *If (X_t) is an ergodic time series and f is a measurable function such that $\mathbb{E}[|f((X_t))|] < \infty$ then:*

$$\frac{1}{n}\sum_{i=1}^{n} f((X_{i+t})) \to \mathbb{E}[f((X_t))] \qquad a.s.$$

Note that we average a function of the complete sequence (X_t) as required in the applications: In particular, it implies a generalization of the Strong Law of Large Numbers under integrability

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mathbb{E}[X_0] \qquad a.s.$$

Here the stability corresponds to the fact that averaging through time of a constant function of the observations converges to a constant.

In order to apply this powerful result, one needs to exhibit stationary and ergodic time series.

Proposition. *Let (Z_t) be an iid sequence, then (Z_t) is stationary and ergodic*

It is a consequence of the zero-one law of Kolmogorov. From this basic example, it is possible to construct other examples that are useful.

Proposition. *If h is a measurable function and if (Z_t) is a stationary and ergodic sequence then $X_i = h((Z_{i+t}))$ constitutes a stationary ergodic sequence.*

Thus, any linear filter of a SWN is a stationary ergodic time series when it exists. Thus it is the case of any solution of an ARMA model.

Combining the ergodic theorem above and the definition of the M-estimator, one can actually prove that the estimator is converging to θ_0 the minimizer of the risk function $\mathbb{E}[\ell_0]$. Denote $x \vee 0 = x^-$:

Theorem (Pfanzagl (1973))**.** *Assume that (ℓ_t) is a stationary ergodic sequence of losses, that θ_0 is the unique minimizer of $\mathbb{E}[\ell_0]$ and that it exists $\varepsilon > 0$ small enough such that*

$$\mathbb{E}\Big[\inf_{\theta \in B(\theta_0,\varepsilon)} \ell_0^-(\theta)\Big] > -\infty$$

then $\widehat{\theta}_n \to \theta_0$ a.s., i.e. the M-estimator of θ_0 is strongly consistent.

3.2 Consistency of the QMLE

3.2.1 Strong consistency of the QMLE

In the case of the QLik approach, when θ corresponds to a causal and invertible ARMA model, one identifies the loss functions with

$$\ell_t(\theta) = \log(R_t^L(\theta)) + \frac{(X_t - \Pi_{t-1}(\theta)(X_t))^2}{R_t^L(\theta)}\,.$$

This function is not stationary and ergodic since it depends on an increasing number of past data $(X_{t-1}, \ldots, X_1)$. We will use the following approximation result from Straumann

Proposition. *If (f_n) is a sequence of measurable functions: $f_n : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}$ such that $(f_n(Z_t, Z_{t-1}, \ldots))$ converge a.s. for some $t \in \mathbb{Z}$, then it exists a measurable function f such that*

$$\widetilde{f}_t = \lim_{n\to\infty} f_n(Z_t, Z_{t-1}, \ldots) = f(Z_t, Z_{t-1}, \ldots), \quad t \in \mathbb{Z}$$

and $(\widetilde{f}_t)$ is stationary ergodic.

The first term in ℓ_t will converge by continuity as $(R_t^L(\theta))$ is converging to $R_\infty^L(\theta) = \sigma^2$.

For the second term, it depends on the convergence of $(\Pi_{t-1}(\theta)(X_t))$. We already know that $(\Pi_{t-1}(X_t(\theta)))$ is converging. If θ corresponds to an invertible ARMA model, we obtain that

$$X_t(\theta) = \sum_{j=1}^{\infty} \varphi_j(\theta) X_{t-j}(\theta) + Z_t, \qquad t \in \mathbb{Z},$$

where $\varphi(\theta)(z) = 1 - \gamma(\theta)(z)^{-1}\phi(\theta)(z)$ with

$$(\phi(\theta)(z), \gamma(\theta)(z)) = (1 - \theta_1 z - \cdots - \theta_p z^p, 1 + \theta_{p+1} z + \cdots + \theta_{p+q} z^q).$$

Thus, one can identify the best linear prediction (with infinite coefficients)

$$\Pi_\infty(\theta)(X_t(\theta)) = \sum_{j=1}^{\infty} \varphi_j(\theta) X_{t-j}(\theta),$$

and then $R_\infty^L(\theta) = \sigma^2$. We also know that there exist $C > 0$ and $0 < \rho < 1$ so that $|\varphi_j| \leq C\rho^j$ and then

$$\mathbb{E}[(\Pi_{t-1}(\theta)(X_t) - \Pi_\infty(\theta)(X_t))^2] = O(\rho^j)$$

for any second order stationary time series (X_t).

We can apply the Proposition from Straumann and we get the stationary ergodic sequence

$$\widetilde{\ell}_t(\theta) = \log(\sigma^2) + \frac{(X_t - \Pi_\infty(\theta)(X_t))^2}{\sigma^2}$$

with the corresponding risk function with

$$\mathbb{E}[\widetilde{\ell}_0](\theta) = \log(\sigma^2) + \frac{\mathbb{E}[(X_0 - \Pi_\infty(\theta)(X_0))^2]}{\sigma^2}.$$

The sequence of loss functions $(\widetilde{\ell}_t)$ is stationary, ergodic and admits second order moments if (X_t) does and we can apply Pfanzagl Theorem to it. That we can apply Pfanzagl theorem also on the original (ℓ_t) which is approximating $(\widetilde{\ell}_t)$ comes from a Cesaro argument. We obtain

Proposition. *If (X_t) is a stationary ergodic time series such that $\mathbb{E}[X_0^2] < \infty$, if Θ corresponds to a causal ARMA models, if there exists a unique minimizer $\theta_0 \in \Theta$ of*

$$\theta \mapsto \mathbb{E}\left[(X_0 - \Pi_\infty(\theta)(X_0))^2\right] \tag{3.1}$$

then the QMLE is strongly consistent $\widehat{\theta}_n \to \theta_0$ a.s. as $n \to \infty$.

The last assumption of uniqueness depends on the parametrization of the model and on the assumptions on (X_t). If one assumes that the observations (X_t) follows themselves an ARMA model with $\theta_0 \in \Theta$, then θ_0 is unique if the polynomials ϕ and γ for $\theta \in \Theta$ do not have common roots. Let us denote $\mathcal{C} \subset \mathbb{R}^{p+q}$ the set of parameters corresponding to a causal and invertible ARMA(p, q) models with no common roots. We have the following *strong consistency* result

Theorem (Hannan (1970)). *If (X_t) satisfies an ARMA(p, q) model with $\theta_0 \in \mathcal{C}$ and (Z_t) SWN(σ^2), $\sigma^2 > 0$, then the QMLE is strongly consistent $\widehat{\theta}_n \to \theta_0$ a.s.*

Proof. The main difficulty is that $\mathcal{C}$ is an open set by definition. One should work on its closure $\bar{\mathcal{C}}$ that is compact after excluding the points on the boundary $\partial\mathcal{C}$ as potential minimizers, see Proposition 10.8.3. of Brockwell and Davis (2013). We will not detail this technical step here.

The rest of the proof is an application of Pfanzagl theorem as above. The ergodicity and stationarity is ensured because of the causal representation $X_t = \sum_{j\geq 0} \psi_j Z_{t-j}$ where (Z_t) is a SWN, thus iid and thus ergodic and stationery. The unicity of θ_0 is derived from the identity

$$\mathbb{E}\left[(X_0 - \Pi_\infty(\theta)(X_0))^2\right] = \mathbb{E}\left[(\Pi_\infty(\theta_0)(X_0) - \Pi_\infty(\theta)(X_0))^2\right] + \sigma^2,$$

obtained using $X_0 = \Pi_\infty(\theta_0)(X_0) + Z_0$ and orthogonality. The expectation term is null iff $\Pi_\infty(\theta_0)(X_0) = \Pi_\infty(\theta)(X_0) = \sum_{j\geq 0} \varphi(\theta_0) X_{-j-1}$ a.s. As the function

$$(\phi(\theta)(z), \gamma(\theta)(z)) = (1 - \theta_1 z - \cdots - \theta_p z^p, 1 + \theta_{p+1} z + \cdots + \theta_{p+q} z^q)$$

giving $\varphi(\theta)(z) = 1 - \gamma(\theta)(z)^{-1}\phi(\theta)(z)$ is injective we deduce that the expectation term is null iff $\theta = \theta_0$. Then θ_0 is the unique minimizer of the risk. □

3.2.2 Estimation of the variance of the noise

In practice, the variance of the WN $\sigma^2 > 0$ is unknown and one has to estimate this parameter. However, since σ^2 is unknown then $R_t^L(\theta)$ is not accessible. Fortunately, we have the identity $R_t^L(\theta) = \sigma^2 r_t^L(\theta)$ where now $r_t^L(\theta)$ corresponds to the risk of linear prediction assuming that (Z_t) is a gaussian (standardized) WN(1). Moreover the innovations $I_t(\theta) = X_t(\theta) - \Pi_{t-1}(X_t(\theta))$ of the ARMA model with parameter θ is independent of σ^2. Thus considering the QLik functions in (θ, σ^2) we have

$$L_n(\theta, \sigma^2) = \sum_{t=1}^{n} \log(\sigma^2 r_t^L(\theta)) + \frac{(X_t - \Pi_{t-1}(\theta)(X_t))^2}{\sigma^2 r_t^L(\theta)} + cst.$$

Differentiating with respect to σ^2 provides

$$\sigma^2(\theta) := \frac{1}{n} \sum_{t=1}^{n} \frac{(X_t - \Pi_{t-1}(\theta)(X_t))^2}{r_t^L(\theta)}.$$

Plugin in the likelihood $L_n(\theta, \sigma^2(\theta))$ we obtain that $\widehat{\theta}_n$ minimizes the *reduced* likelihood

$$L_n^0(\theta) = n \log\Big(\frac{1}{n}\sum_{t=1}^{n} \frac{(X_t - \Pi_{t-1}(\theta)(X_t))^2}{r_t^L(\theta)}\Big) + \sum_{t=1}^{n} \log(r_t^L(\theta)) + cst\,.$$

Noticing that the asymptotic behavior of the contrast $n^{-1}L_n^0(\theta)$ is similar than above, an application to Pfanzagl theorem on $L_n^0(\theta)$ provides

Proposition. *If (X_t) satisfies an ARMA(p, q) model with $\theta_0 \in \mathcal{C}$ and (Z_t) SWN(σ^2), $\sigma^2 > 0$, then the QMLE provides also a strongly consistent estimator of the variance σ^2*

$$\widehat{\theta}_n \xrightarrow{a.s.} \theta \qquad \text{and} \qquad \widehat{\sigma}_n^2 \xrightarrow{a.s.} \sigma^2, \qquad n \to \infty.$$

Remark. *One considered $\Pi_t(\theta)$ and $r_t^L(\theta)$ as known. It is actually one main crucial issue to compute $\Pi_t(\theta)$ and $r_t^L(\theta)$ efficiently, issue that will be solved later.*

By definition, we have $r_t^L(\theta) \to r_\infty^L(\theta) = 1$ for any $\theta \in \mathcal{C}$ by definition of $r_t^L(\theta)$. Using a Cesaro argument, we obtain $n^{-1}\sum_{t=1}^{n} \log(r_t^L(\theta)) \to 0$ so that

$$\frac{1}{n}L_n^0(\theta) \approx \log\Big(\frac{1}{n}\sum_{t=1}^{n} \frac{(X_t - \Pi_{t-1}(\theta)(X_t))^2}{r_t^L(\theta)}\Big) + cst\,.$$

Thus the QMLE approximatively minimizes the squares

$$\sum_{t=1}^{n} \frac{(X_t - \Pi_{t-1}(\theta)(X_t))^2}{r_t^L(\theta)}.$$

The *least squares* estimator $\widetilde{\theta}_n$ is defined as the minimizer of this contrast. One can also define a least squares estimator of the variance

$$\widetilde{\sigma}_n^2 = \frac{1}{n-p-q}\sum_{t=1}^{n} \frac{(X_t - \Pi_{t-1}(\widetilde{\theta}_n)(X_t))^2}{r_t^L(\widetilde{\theta}_n)}$$

where $n - p - q$ stands for the $n - p - q$ degrees of freedom. The lease squares estimators and the QMLE are asymptotically equivalent.

3.2.3 Misspecification

Definition 32. *The QMLE is well specified when the observations are stationary solutions of an ARMA(p, q) model for the same (p, q) and with gaussian WN(σ^2), $\sigma^2 > 0$. Otherwise it is misspecified.*

Consider now that (X_t) is a centered stationary ergodic time series such that $\mathbb{E}[X_0^2] < \infty$. We do not assume anymore that (X_t) follows an ARMA(p, q) model. One studies the asymptotic behaviour of the QMLE of the ARMA model with $\mathcal{C} \in \mathbb{R}^{p+q}$. Such cases are called *misspecification* as the density used to calculate the contrast is not the correct one. Such setting is very important to obtain results that are satisfied even if the normal and the model assumptions used to derive the QLik loss does not hold. In this context, Pfanzagl theorem still holds and a careful look at the proof of the strong consistency show that it is still valid under certain conditions. Actually, one can always decompose the second order stationary process (X_t) as

$$X_t = \Pi_\infty(X_t) + I_\infty(X_0)$$

where $I_\infty(X_0)$ is orthogonal to the span of the past values $\{X_{-1}, X_{-2}, \ldots\}$. Thus we obtain

$$\mathbb{E}\left[(X_0 - \Pi_\infty(\theta)(X_0))^2\right] = \mathbb{E}\left[(\Pi_\infty(X_0) - \Pi_\infty(\theta)(X_0))^2\right] + R_\infty^L,$$

We are left to discuss the unicity of the minimizer of the function

$$\theta \mapsto \mathbb{E}[(\Pi_\infty(X_0) - \Pi_\infty(\theta)(X_0))^2] =: \mathbb{E}\left[\left(\sum_{j\geq 1}(\varphi_j - \varphi_j(\theta))X_{-j}\right)^2\right].$$

Developing this quantity, we find a function of $u_j = (\varphi_j - \varphi_j(\theta))$:

$$\sum_{i\geq 0}\sum_{j\geq 0} u_i \gamma_X(|j-i|)u_j \in [0, \infty].$$

As $R_\infty^L > 0$, there is no co-linearity in $(X_j)_{j\leq 0}$ and the kernel of this function is restricted to $\{0\}$. It is not hard to show that it is a (possibly infinite) norm on the space of square integrable series. One can define a projection on any closed convex subset of this space, in particular

$$\varphi(\bar{\mathcal{C}}) := \{(\varphi_j(\theta)); \theta \in \bar{\mathcal{C}}\}.$$

However, one has to check that the norm is not infinite over $\varphi(\bar{\mathcal{C}})$. We know that for each elements of $(u_j) \in \varphi(\bar{C})$ there exist $C > 0$ and $0 < \rho < 1$ so that $|u_j| \leq C\rho^j$. Thus, if $\sum_{h\geq 0}|\gamma_X(h)| < \infty$ we have

$$\sum_{i\geq 0}\sum_{j\geq 0} u_i \gamma_X(|j-i|)u_j \leq 2C^2 \sum_{i\geq 0}\rho^i \sum_{h\geq 0}|\gamma_X(h)|\rho^{i+h} \leq 2C^2 \sum_{i\geq 0}\rho^{2i}\frac{\sum_{h\geq 0}|\gamma_X(h)|}{1-\rho} < \infty.$$

Thus $\mathbb{E}[(\Pi_\infty(X_0) - \Pi_\infty(\theta)(X_0))^2]$ is minimized by the projection of the coefficients of $\Pi_\infty(X_0)$ over $\varphi(\bar{C})$. The coefficients of the projection $(\varphi_j(\theta))$ are unique but not necessarily the parameters $\theta \in \partial\mathcal{C}$. For instance, if (X_t) is not an ARMA model with minimal lag polynomials γ and ϕ, one cannot avoid the possibility of parameters $\theta \in \partial\mathcal{C}$ that correspond to polynomial with common roots. We say that a point y converges to a set $\mathcal{X}$ when $d(y, \mathcal{X}) = \inf_{x\in\mathcal{X}}\|y - x\| \to 0$. We obtain

Proposition. *Consider a centered stationary ergodic time series (X_t) such that $\mathbb{E}[X_0^2] < \infty$, $R_\infty^L > 0$ and $\sum_{h\geq 0}|\gamma_X(h)| < \infty$. Then the QMLE converges to the set Θ_0 corresponding to the coefficients $(\varphi_j(\theta_0))$ that uniquely determine the best linear prediction over $\varphi(\bar{C})$.*

Example 13. *Fitting an ARMA(1,1) on a SWN it is not possible to avoid the case of common roots $\phi_1 = -\gamma_1$ as shown by the following code from tsaEZ*

```
> set.seed(8675309)
> x = rnorm(150, mean=5) # generate iid N(5,1)s
> arima(x, order=c(1,0,1)) # estimation

Call:
arima(x = x, order = c(1, 0, 1))

Coefficients:
          ar1     ma1  intercept
      -0.9595  0.9527     5.0462
s.e.   0.1688  0.1750     0.0727

sigma^2 estimated as 0.7986:  log likelihood = -195.98,  aic = 399.96
```

As emphasised in the exemple above, the variance σ^2 is no longer consistently estimated by the square of the innovations of the fitted ARMA(1,1) model due to the extra aditive term $\mathbb{E}[(\Pi_\infty(X_0) - \Pi_\infty(\theta)(X_0))^2]$ called the bias.

3.3 Asymptotic normality and model selection

We have seen that it is crucial in practice to choose a good ARMA model in order to fit the data. The choice of the orders of the ARMA models is a difficult task requiring properties on the QMLE as the asymptotic normality.

3.3.1 Kullback-Leibler divergence

One can identify the risk $\mathbb{E}[\widetilde{\ell}_0]$ associated with the QLik loss with an important notion from information theory that is a pseudo-distance between probability measures.

Definition 33. *The Kullback-Leibler divergence (KL, relative entropy) between two probability measures P_1 and P_2 is defined as*

$$\mathcal{K}(P_1, P_2) = \mathbb{E}_{P_1}[\log(dP_1/dP_2)].$$

The KL divergence has nice properties

Proposition. *We have $\mathcal{K}(P_1, P_2) \geq 0$ and $\mathcal{K}(P_1, P_2) = 0$ iff $P_1 = P_2$ a.s.*

One can identify, up to additive constants, the standardized risk

$$\mathbb{E}[\widetilde{\ell}_0(\theta)] = \log(\sigma^2) + \frac{\mathbb{E}\left[(X_0 - \Pi_\infty(\theta)(X_0))^2\right]}{\sigma^2}$$

as twice the expectation of the KL divergence of

$$2\mathbb{E}[\mathcal{K}(P_{X_0|X_{-1},X_{-2},\ldots}, \mathcal{N}(\Pi_\infty(\theta)(X_0), \sigma^2))]$$

where the expectation is taken over the distribution of the past $(X_{-1}, X_{-2}, \ldots)$ and the KL divergence is understood conditional to this past.

Thus, if (X_t) follows an ARMA model with parameter θ_0, we have that θ_0 is the unique minimizer of the risk $\mathbb{E}[\widetilde{\ell}_0]$ but also of the conditional risk

$$\mathbb{E}[\widetilde{\ell}_0(\theta) \mid X_{-1}, X_{-2}, \ldots] = 2\mathcal{K}(P_{X_0|X_{-1},X_{-2},\ldots}, \mathcal{N}(\Pi_\infty(\theta)(X_0), R^L_\infty(\theta))) + cst.$$

3.3.2 Asymptotic normality of the MLE

Let us turn to the general case of any ML Estimator

$$\widetilde{\theta}_n \in \arg\min_\Theta \widetilde{L}_n(\theta) = \arg\min_\Theta \sum_{t=1}^n \widetilde{\ell}_t(\theta)$$

where $\widetilde{\ell}_t = -2\log(f_\theta(X_t \mid X_{t-1}, X_{t-2}, \ldots))$ constitutes a stationary sequence of contrast such that θ_0 is the unique minimizer of $\mathbb{E}[\widetilde{\ell}_0]$ on a compact set $\Theta \subset \mathbb{R}^d$, $d \geq 0$ being the dimension of the parametric estimation. We assume that the conditions of integrability in Pfanzagl theorem are satisfied so that $\widetilde{\theta}_n$ is strongly consistent. The asymptotic normality

of the MLE follows in most of the cases under extra assumptions. If $\widetilde{L}_n$ is sufficiently regular (2-times continuously differentiable) then a Taylor expansion gives

$$\nabla \widetilde{L}_n(\widetilde{\theta}_n) \approx \nabla \widetilde{L}_n(\theta_0) + \nabla^2 \widetilde{L}_n(\widetilde{\theta}_n)(\widetilde{\theta}_n - \theta_0)\,. \tag{3.2}$$

Notice that if $\widetilde{\theta}_n \in \overset{\circ}{\Theta}$ the interior of the compact set then $\nabla \widetilde{L}_n(\widetilde{\theta}_n) = 0$ as the MLE is the minimizer of the likelihood contrast by assumption. So we have to study the properties of the two first derivative of the contrast $\widetilde{L}_n$. Let us first show that the two first derivatives of $\widetilde{L}_n$ have nice properties at $\widetilde{\theta}_0$:

Definition 34. *The score vector is defined as the gradient of the QLik loss (up to constant)*

$$S_t(\theta) = \partial_\theta \log(f_\theta(X_t \mid X_{t-1}, X_{t-2}, \ldots)).$$

The Fisher's information is $\mathcal{I}(\theta_0) = -\mathbb{E}[\partial^2_\theta \log(f_\theta(X_t \mid X_{t-1}, X_{t-2}, \ldots))]$.

We have the following property, deriving from the definition of θ_0 as the unique minimizer of $\mathbb{E}[-2\log(f_\theta(X_t \mid X_{t-1}, X_{t-2}, \ldots)) \mid X_{t-1}, X_{t-2}, \ldots]$ from the discussion on the KL divergence, we obtain

Proposition. *If* $\theta_0 \in \overset{\circ}{\Theta}$ *is the unique minimizer of the conditional risk*

$$-\mathbb{E}[\log(f_\theta(X_t \mid X_{t-1}, X_{t-2}, \ldots)) \mid X_{t-1}, X_{t-2}, \ldots]$$

then the score vector is centered $\mathbb{E}[S_0(\theta_0) \mid \mathcal{X}_{-1}, X_{-2}, \ldots] = 0$ *and* $\mathcal{I}(\theta_0)$ *is a symmetric definite positive matrix. If moreover the model is* well-specified *so that* $f(X_t \mid X_{t-1}, X_{t-2}, \ldots) = f_{\theta_0}(X_t \mid X_{t-1}, X_{t-2}, \ldots)$*, then* $\mathcal{I}(\theta_0) = Var(S_0(\theta_0))$ *and its inverse(*$\times n$*) is the smallest possible variance of any unbiased estimator, called the Cramer-Rao bound.*

Proof. As θ_0 is the minimizer of the conditional KL in the interior of a compact set, the derivative is null at this point. Thus the score is centered by differentiating under the integral. Moreover, the Fisher information is definite otherwise the minimizer is not unique.

Assume now that $f(X_t \mid X_{t-1}, X_{t-2}, \ldots)$ coincides with $f_{\theta_0}(X_t \mid X_{t-1}, X_{t-2}, \ldots)$ which is identically distributed as $f_{\theta_0} := f_{\theta_0}(X_0 \mid X_{-1}, X_{-2}, \ldots)$ by stationarity. Then we have

$$0 = \mathbb{E}[\partial_\theta \log(f_{\theta_0}) \mid X_{-1}, X_{-2}, \ldots] = \mathbb{E}\left[\frac{\partial_\theta f_{\theta_0}}{f_{\theta_0}} \mid X_{-1}, X_{-2}, \ldots\right] = \int \partial_\theta f_{\theta_0}.$$

Assuming that one can differentiate under the integral, we then also have $\int \partial^2_\theta f_{\theta_0} = 0$. Simple calculation yields

$$I(\theta_0) = \mathbb{E}\left[\frac{\partial_\theta f_{\theta_0} \partial_\theta f_{\theta_0}^\top - f_{\theta_0}\partial^2_\theta f_{\theta_0}}{f^2_{\theta_0}}\right] = \mathbb{E}[S_0(\theta_0)S_0(\theta_0)^\top] = \mathrm{Var}\,(S_0(\theta_0)).$$

The proof of the Cramer-Rao bound is classical for $d = 1$. Any unbiased estimator θ_n satisfies $\mathbb{E}[\theta_n] = \theta_0$ which can be written as

$$\mathbb{E}\Big[\int \theta_n \prod_{t=1}^{n} f_{\theta_0}(X_t \mid X_{t-1}, X_{t-2}, \ldots)\Big] = \theta_0.$$

By differentiation over θ_0, we obtain $\sum_{t=1}^n \mathbb{E}[\int \theta_n \partial_\theta f_{\theta_0}(X_t \mid X_{t-1}, X_{t-2}, \ldots)] = 1$. Thus as the scores are centered and by Cauchy-Schwarz inequality we obtain

$$\begin{aligned}
1 &= \sum_{t=1}^n \mathbb{E}\Big[\int (\widehat{\theta}_n - \theta_0) \partial_\theta f_{\theta_0}(X_t \mid X_{t-1}, X_{t-2}, \ldots) \Big] \\
&= \sum_{t=1}^n \mathbb{E}\left[\int (\widehat{\theta}_n - \theta_0) \sqrt{f_{\theta_0}(X_t \mid X_{t-1}, X_{t-2}, \ldots)} \frac{\partial_\theta f_{\theta_0}(X_t \mid X_{t-1}, X_{t-2}, \ldots)}{\sqrt{f_{\theta_0}(X_t \mid X_{t-1}, X_{t-2}, \ldots)}} \right] \\
&\leq \sum_{t=1}^n \mathbb{E}\left[\sqrt{\int (\widehat{\theta}_n - \theta_0)^2 f_{\theta_0}(X_t \mid X_{t-1}, X_{t-2}, \ldots) \mathbb{E}[S_t(\theta_0)^2 | X_{t-1}, X_{t-2}, \ldots]} \right] \\
&\leq \sum_{t=1}^n \mathrm{Var}\,(\widehat{\theta}_n) \mathrm{Var}\,(S_t(\theta_0)) = n \mathrm{Var}\,(\widehat{\theta}_n) \mathrm{Var}\,(S_0(\theta_0))
\end{aligned}$$

and the desired result follows. □

The inverse of the Fisher information is interpreted as the best possible asymptotic variance. We obtain

Theorem. *If there exists $\theta \in \overset{\circ}{\Theta}$ which is the unique minimizer of the conditional risk, if the contrast $\widetilde{\ell}_t = -2\log(f_\theta(X_t \mid X_{t-1}, X_{t-2}, \ldots))$ is twice continuously differentiable and integrable, then the MLE is* asymptotically normal

$$\sqrt{n}(\widetilde{\theta}_n - \theta_0) \xrightarrow{d.} \mathcal{N}(0, I(\theta_0)^{-1} \mathit{Var}(S_0) I(\theta_0)^{-1}).$$

Moreover, it is asymptotically efficient*, i.e. the asymptotic variance coincides with the Cramer-Rao bound, when the model is well-specified.*

Proof. The sequence of score vectors (S_t) constitutes a differences of martingale process. The CLT extends to such square integrable differences of martingale and we obtain

$$-\frac{1}{\sqrt{n}} \nabla \widetilde{L}_n(\theta_0) = 2\frac{1}{\sqrt{n}} \sum_{t=1}^n S_t \xrightarrow{d.} \mathcal{N}(0, 4\mathrm{Var}\,(S_0)).$$

One can also use the ergodic theorem and the strong consistency of $\widehat{\theta}_n$ to obtain

$$\frac{1}{n} \nabla^2 \widetilde{L}_n(\widetilde{\theta}_n) = -2\frac{1}{n} \sum_{t=1}^n \partial_\theta^2 \log(f_{\widetilde{\theta}_n}(X_t \mid X_{t-1}, X_{t-2}, \ldots)) \xrightarrow{a.s.} 2\mathcal{I}(\theta_0).$$

Thus, starting from the identity (3.2), we obtain

$$\begin{aligned}
& 0 \approx \nabla \widetilde{L}_n(\theta_0) + \nabla^2 \widetilde{L}_n(\widetilde{\theta}_n)(\widetilde{\theta}_n - \theta_0) \\
\Leftrightarrow \quad & -\nabla \widetilde{L}_n(\theta_0) \approx \nabla^2 \widetilde{L}_n(\widetilde{\theta}_n)(\widetilde{\theta}_n - \theta_0) \\
\Leftrightarrow \quad & -\frac{1}{\sqrt{n}} \nabla L_n(\theta_0) = \frac{1}{n} \nabla^2 \widetilde{L}_n(\widetilde{\theta}_n) \sqrt{n}(\widehat{\theta}_n - \theta_0).
\end{aligned}$$

The LHS of the last identity converges in distribution to $\mathcal{N}(0, 4\mathrm{Var}\,(S_0))$, the RHS is a.s. equivalent to $2\mathcal{I}(\theta_0)\sqrt{n}(\widehat{\theta}_n - \theta_0)$ so that the desired result is obtained. □

3.3.3 Asymptotic normality of the QMLE

We assume here that σ^2 is known. As θ_0 was uniquely determined in Theorem 3.2.1, as $\mathcal{C}$ is an open set so that $\theta_0 \in \overset{\circ}{\mathcal{C}}$, we immediately obtain the asymptotic normality of the QMLE:

Theorem (Hannan (1970))**.** *If (X_t) satisfies an ARMA(p, q) model with $\theta_0 \in \mathcal{C}$ and (Z_t) SWN(σ^2), $\sigma^2 > 0$, then the QMLE is asymptotically normal*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d.} \mathcal{N}_{p+q}\left(0,\, Var\,(AR_p, \ldots, AR_1, MA_q, \ldots, MA_1)^{-1}\right)$$

where (AR_t) and (MA_t) are AR(p) and AR(q) models driven by the coefficient θ_0, the same SWN(1) (η_t) and satisfying

$$\phi(\theta_0)(L)AR_t = \gamma(\theta_0)(L)MA_t = \eta_t, \qquad t \in \mathbb{Z}.$$

In misspecified cases, the uniqueness of θ_0 is not ensured and the asymptotic normality result is not possible in full generality.

Proof. We deal with the stationary ergodic losses rather than their approximations used for calculating $\widehat{\theta}_n$. Indeed the approximation is converging exponentially fast in $\mathbb{L}^2$ and will not have any consequence on the asymptotic properties of $\widehat{\theta}_n$. One first checks the differentiability and integrability conditions on the QLik contrast

$$\widetilde{\ell}_t(\theta) = \log(\sigma^2) + \frac{(X_t - \Pi_\infty(\theta)(X_t))^2}{\sigma^2}.$$

The score is defined as

$$S_t(\theta) = \frac{(X_t - \Pi_\infty(\theta)(X_t))}{R^L_\infty(\theta)} \partial_\theta \Pi_\infty(\theta)(X_t).$$

From the identity $X_t - \Pi_\infty(\theta_0)(X_t) = Z_t$ we obtain the expression

$$S_t = \frac{1}{\sigma^2} Z_t \partial_\theta \Pi_\infty(\theta_0)(X_t).$$

One checks easily that $\mathbb{E}[S_0 \mid X_{t-1}, X_{t-2}, \ldots] = 0$ even when (Z_t) is not gaussian. Its variance is

$$\mathrm{Var}\,(S_0) = \frac{1}{\sigma^2}\mathbb{E}[\partial_\theta \Pi_\infty(\theta_0)(X_t) \partial_\theta \Pi_\infty(\theta_0)(X_t)^\top]$$

Similarly, one computes the Fisher information

$$\begin{aligned}\mathcal{I}(\theta_0) &= \frac{1}{\sigma^2}\mathbb{E}[\partial_\theta \Pi_\infty(\theta_0)(X_t) \partial_\theta \Pi_\infty(\theta_0)(X_t)^\top - Z_t \partial^2_\theta \Pi_\infty(\theta_0)(X_t)] \\ &= \frac{1}{\sigma^2}\mathbb{E}[\partial_\theta \Pi_\infty(\theta_0)(X_t) \partial_\theta \Pi_\infty(\theta_0)(X_t)^\top].\end{aligned}$$

As $\Pi_\infty(\theta_0)(X_t) = \gamma(L)^{-1}\phi(L)X_t$ we have that

$$\partial_{\phi_k} \Pi_\infty(\theta_0)(X_t) = \gamma(L)^{-1} L^k X_t = \phi(L)^{-1} Z_{t-k} = \sigma AR_{t-k}.$$

Similarly, we have

$$\partial_{\gamma_k} \gamma(L)^{-1} = \partial_{\gamma_k} \gamma(L) \gamma(L)^{-2} = L^k \gamma(L)^{-2}$$

so that

$$\partial_{\gamma_k} \Pi_\infty(\theta_0)(X_t) = L^k \gamma(L)^{-2} \phi(L) X_t = \gamma(L)^{-1} Z_{t-k} = \sigma MA_{t-k}.$$

The desired result follows. □

Note that the asymptotic variance of $\widehat{\theta}_n$ does not depend on σ^2. It complements the fact that θ and σ^2 can be estimated separately in ARMA models.

From the proof, we have an alternative expression for the asymptotic variance as

$$I(\theta_0)^{-1} = \sigma^2 \mathbb{E}[\partial_\theta \Pi_\infty(\theta_0)(X_t) \partial_\theta \Pi_\infty(\theta_0)(X_t)^\top]^{-1}.$$

As soon as (Z_t) is SWN($\sigma^2 = R_\infty^L > 0$), the identity $I(\theta_0) = \mathrm{Var}\,(S_0)$ holds and the QMLE is efficient.

Note that the asymptotic variance can be estimated by computing the covariances of (AR_t) and (MA_t) driven by the QMLE $\widehat{\theta}_n$ (actually one can compute it explicitly in term of the coefficients θ of the polynomial of (AR_t) and (MA_t) or one can use numerical approximations, see exercices class).

3.3.4 Asymptotic properties of the maximum of the reduced likelihood

We give an heuristic of the order second behavior of the bias of $L_n^0(\widehat{\theta}_n)$ for p and q fixed. We use a Taylor expansion

$$L_n^0(\theta_0) \approx L_n^0(\widehat{\theta}_n) + \nabla L_n^0(\widehat{\theta}_n)(\theta_0 - \widehat{\theta}_n) + \frac{1}{2}(\theta_0 - \widehat{\theta}_n)^\top \nabla^2 L_n^0(\widehat{\theta}_n)(\theta_0 - \widehat{\theta}_n).$$

From the definition of $\widehat{\theta}_n$ as the minimizer of L_n^0, the first derivative is null and the first order term in the expansion vanishes. We assume the optimal asymptotic normality result achieving the Cramer-Rao bound $\mathcal{I}(\theta_0) = \nabla^2 \widetilde{\ell}_0(\theta_0)]^{-1}$ as asymptotic variance

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d.} \mathcal{N}_{p+q}\left(0, 2\mathbb{E}[\nabla^2 \widetilde{\ell}_0(\theta_0)]^{-1}\right).$$

Then Slutsky Lemma applies on the second order term and we get

$$\frac{1}{2}\sqrt{n}(\theta_0 - \widehat{\theta}_n)^\top \frac{1}{n} \nabla^2 L_n^0(\widehat{\theta}_n) \sqrt{n}(\theta_0 - \widehat{\theta}_n) \approx N^\top N$$

where N is a standard $p+q$ gaussian vector. We have $\mathbb{E}[N^\top N] = p+q$. Summarizing our findings, we obtain that in expectation we have

$$\mathbb{E}[L_n^0(\widehat{\theta}_n)] \approx \mathbb{E}[L_n^0(\theta_0)] - (p+q).$$

This result, depending on $\widehat{\theta}_n$ throught the predictions $\Pi_t(\widehat{\theta}_n)(X_t)$ and not on the parametric estimation, might be extended in misspecified cases. The quality of the prediction at $\widehat{\theta}_n$ estimated on the sample $(X_1, \dots X_n)$ is strictly better than the best possible prediction (using θ_0). Moreover it depends on the number $p+q$ of parameter one has to estimate. This statement is not contradictory because using X_t twice in $(X_t - \Pi_t(\widehat{\theta}_n(p,q))(X_t))^2$, once for calculating $\widehat{\theta}_n$ and another time for calculating the error, one under estimates the risk of prediction. It is because $\widehat{\theta}_n$ uses the future in $\Pi(\widehat{\theta}_n(p,q))(X_t)$ to predict X_t.

3.3.5 Akaike and other information criteria

One faces a crucial issues when fitting an ARMA model to observations that are not issued from an ARMA model themselves (the model is misspecified, which is always the case in practice). Thus, in order to find the sparsest ARMA representation for our observation (X_t) it is fundamental to have some criteria in order to choose the smallest order (p,q) of the model.

A good measure between distributions is the KL-divergence, see Section 3.3.1. From an ARMA(p,q) model, the QML approach will predict the future value thanks to the distribution $\mathcal{N}(\Pi_\infty(\widehat{\theta}_n)(X_0), \widehat{\sigma}_n^2)$. Let us define

Definition 35. *The* predictive power *of the model ARMA(p, q) fitted by the QMLE $\widehat{\theta}_n$ is*

$$K(P_{X_0|X_{-1},X_{-2},\dots}, \mathcal{N}(\Pi_\infty(\theta)(X_0), \sigma^2))\Big|_{\theta=\widehat{\theta}_n}.$$

It is the KL divergence between the distribution of the future of the observation given the past and the distribution of the prediction given the ARMA(p, q) model fitted by the QMLE.

By comparing the predictive power for different orders (p, q) and choosing the smallest number of parameters $p+q$ that achieves the maximal predictive power, one should choose the sparsest ARMA representation with the best prediction. Let us denote $\widehat{\theta}_n(p, q)$ the QMLE for the ARMA(p, q) on the reduced likelihood. Akaike idea is to approximate (-2 times) the predictive power by penalizing the quantity

$$\frac{1}{n}L_n^0(\widehat{\theta}_n) = \log(\sigma^2(\widehat{\theta}_n(p,q))) + \frac{1}{n}\sum_{t=1}^{n}\log(r_t^L(\widehat{\theta}_n(p,q))) + 1.$$

However, the above expression is a biased estimator of (-2 times) the predictive power because the sample $(X_1, \dots X_n)$ is used twice in $(X_t - \Pi(\widehat{\theta}_n(p, q))(X_t))^2$, once for calculating $\widehat{\theta}_n$ and another time for estimating the function $\mathbb{E}[\ell_0]$. More precisely, we have

Definition 36. *We define three information criteria as penalized log-likelihood*

1. *Akaike Information Criterion: AIC* $= \frac{1}{n}L_n^0(\widehat{\theta}_n(p, q)) + \frac{2(p+q)}{n}$,
2. *Bayesian Information Criterion: BIC* $= \frac{1}{n}L_n^0(\widehat{\theta}_n(p, q)) + \frac{\log n(p+q)}{n}$,
3. *Akaike Information Criterion corrected: AICc* $= \frac{1}{n}L_n^0(\widehat{\theta}_n(p, q)) + \frac{2(p+q)}{n-p-q-1}$.

We have $\frac{1}{n}L_n^0(\widehat{\theta}_n) \approx \log(\widehat{\sigma}_n^2(p, q)) + 1$ when $r_t(\widehat{\theta}_n(p, q)) \to 1$ (i.e. the well-specified case) and some authors considered instead AIC $= \log(\widehat{\sigma}_n(p, q)) + \frac{2(p+q)}{n}$, BIC $= \log(\widehat{\sigma}_n(p, q)) + \frac{\log(n)(p+q)}{n}$ and AICc $= \log(\widehat{\sigma}_n^2(p, q)) + \frac{2(p+q)}{n-p-q-1}$.

The procedure is then to select the order $(\widehat{p}_n, \widehat{q}_n)$ that minimizes one of the information criterion. Notice that one can compare the penalties and as $AIC < AICc < BIC$ for a fixed model, the order chosen by the procedure will be reversed; BIC will choose the sparsest model whereas AIC will choose the model with the largest number of parameters.

If the observations (X_t) satisfies an ARMA(p,q) model then, asymptotically,

- BIC procedure chooses the correct order,
- AIC and, a fortiori, AICc, select the best predictive model.

Notice that the best predictive model is not necessarily the true model. AICc is preferred to AIC that can over-fit when n is small. The last item follows from the heuristic

Proposition (Akaike (1974))**.** *The AIC defined above are asymptotically unbiased estimators of the predictive power of the ARMA(p, q) model.*

Proof. We give the heuristic for the AIC only. From the discussion Section 3.3.4 we have

$$L_n^0(\widehat{\theta}_n) \approx L_n^0(\theta_0) - (p+q).$$

On the opposite, we have the Taylor expansion of the predictive power term is

$$\mathbb{E}[\widetilde{\ell}_0(\widehat{\theta}_n)] \approx \mathbb{E}[\widehat{\ell}_0(\theta_0)] + \mathbb{E}[\nabla_\theta \widetilde{\ell}_0(\theta_0)^\top(\widehat{\theta}_n - \theta_0)] + \frac{1}{2}(\widehat{\theta}_n - \theta_0)^\top \mathbb{E}[\nabla^2 \widetilde{\ell}_0(\theta_0)](\widehat{\theta}_n - \theta_0).$$

The first order term is null as θ_0 is the unique minimizer of $\mathbb{E}[\widetilde{\ell}_0]$. Moreover, thanks to the asymptotic normality of the QMLE as in Section 3.3.4 we obtain

$$\mathbb{E}[\sqrt{n}(\widehat{\theta}_n - \theta_0)^\top \mathbb{E}[\nabla^2 \widetilde{\ell}_0(\theta_0)]\sqrt{n}(\widehat{\theta}_n - \theta_0)] \approx \mathbb{E}[N^\top N] \approx p + q.$$

We obtain the desired result

$$\begin{aligned}\frac{1}{n}\mathbb{E}[L_n^0(\widehat{\theta}_n(p,q))] + \frac{2(p+q)}{n} &\approx \frac{1}{n}\mathbb{E}[L_n^0(\theta_0(p,q))] + \frac{p+q}{n} \\ &\approx \mathbb{E}[\widetilde{\ell}_0(\theta_0)] + \frac{p+q}{n} \\ &\approx \mathbb{E}[\widetilde{\ell}_0(\widehat{\theta}_n)].\end{aligned}$$

□

3.3.6 Interval of prediction.

The aim of time series model is to produce forecasting under the condition that (X_t) is stationary. We will assert the point and interval predictions produced by the ARMA model and we will discuss its ability.

Let us first consider the one step prediction. The prediction of X_{n+1} is given by $\widehat{X}_{n+1} = \Pi_n(\widehat{\theta}_n)(X_{n+1})$. An interval of prediction is often more useful than a point prediction. Denoting $\hat{\sigma}^2(1) = R_n^L(\widehat{\theta}_n)$ the QMLE produces a natural interval of confidence α such as

$$\widehat{I}_\alpha(X_{n+1}) = [\widehat{X}_{n+1} - q^N_{1-\alpha/2}\sigma_n(1); \widehat{X}_{n+1} + q^N_{1-\alpha/2}\widehat{\sigma}_n(1)]$$

where $q^N_{1-\alpha/2}$ is the quantile of order $1-\alpha/2$ of the standard gaussian r.v. N. It is an estimator of the best interval for X_{n+1} given the past which is defined as

$$I_\alpha(X_{n+1}) = [q_\beta(X_{n+1} \mid X_n, \ldots, X_1), q_{\alpha-\beta}(X_{n+1} \mid X_n, \ldots, X_1)]$$

where $q_\beta(X_{n+1} \mid X_n, X_{n-1}, \ldots)$ is the quantile of order $0 \geq \beta \geq 1$ of the conditional distribution of X_{n+1} given the observations $X_1, \ldots, X_n$ and β is chosen such that the length of the interval is the smallest possible. Often, we assume that the conditional distribution is symmetric and then $\beta = \alpha/2$.

For an ARMA model, it is also possible to produce h step prediction intervals for any $h \geq 1$ as

$$\widehat{I}_\alpha(X_{n+h}) = [\Pi_n(\widehat{\theta}_n)(X_{n+h}) - q^N_{1-\alpha/2}\widehat{\sigma}_n(h); \Pi_n(\widehat{\theta}_n)(X_{n+h}) + q^N_{1-\alpha/2}\widehat{\sigma}_n(h)]$$

where $\Pi_n(\theta)(X_{n+h})$ is the best linear projection of X_{n+h} on the span of the observation given the ARMA model θ such that

$$\Pi_n(\theta)(X_{n+h}) \approx \sum_{i=1}^{p} \phi_i \Pi_n(\theta)(X_{n+h-i}) + \sum_{j=h}^{q} \theta_j (X_{n+h-j} - \Pi_{n+h-j-1}(\theta)(X_{n+h-j}))$$

and $\widehat{\sigma}_n(h)$ is the associated variance

$$\widehat{\sigma}_n^2(h) \approx \widehat{\sigma}_n^2 \sum_{j=0}^{h-1} \psi_j(\widehat{\theta}_n)^2.$$

Notice that the issue of the explicit and efficient computations of those quantities will be treated later.

The usefulness of the interval of prediction is that it provides *indicators of risk*;

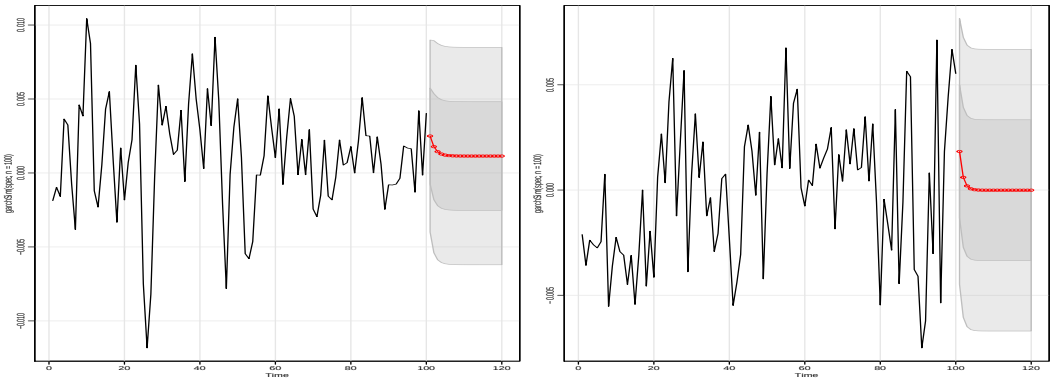

Figure 3.1: The interval of prediction does not take into account the present variability of the time series. When the present variability is low (on the left), the interval is too conservative (too large). On the opposite, when the present variability is high (on the right), the interval is too optimistic (too low).

Definition 37. *The length of the interval* $|\widehat{I}_\alpha(X_{n+1}))|$ *is an indicator of the risk of prediction with confidence level* $1-\alpha$*; in the symmetric case, the lower and upper points are indicators of the risk of lower and higher values with level* $\alpha/2$ *called* Values at Risks *(VaR, quantiles of the conditional distribution).*

The prediction forecast provides good indicators of any level if it has a large predictive power

$$-2\mathbb{E}[\mathcal{K}(P_{X_0|X_{-1},X_{-2}}), \mathbb{P}_\theta(X_0 \mid X_{-1}, X_{-2}, \ldots))]\Big|_{\theta=\widehat{\theta}_n}.$$

The QMLE for ARMA models the conditional distribution

$$\mathbb{P}_{\widehat{\theta}_n}(X_0 \mid X_{-1}, X_{-2}, \ldots) = \mathcal{N}(\Pi_n(\widehat{\theta}_n)(X_{n+1}), \sigma^2(\widehat{\theta}_n)).$$

There $\widehat{\sigma}^2 \approx \sigma^2$ is approximatively a constant and the model on the conditional probability is dependent on the present observations only for the mean $\Pi_n(\widehat{\theta}_n)(X_{n+1})$. Thus, ARMA models produce good point prediction but may fail for interval of predictions. The center of the interval of prediction is accurate in view of the past values but not the length of the interval that adapts not well to the present behavior of the time series.

Example 14. *Let us consider* $X_t = \phi X_t + Z_t$ *where* (Z_t) *is a WN*(σ^2)*. Then the interval of prediction of confidence level* $1-\alpha$ *of horizon* h *is given by*

$$\widehat{I}_\alpha(X_{n+h}) = [\widehat{\phi}_n^h X_n - q^N_{1-\alpha/2}\widehat{\sigma}_n(h), \widehat{\phi}_n^h X_n + q^N_{1-\alpha/2}\widehat{\sigma}_n(h)]$$

where $\widehat{\phi}_n = \sum_{t=2}^n X_t X_{t-1} / \sum_{t=1}^n X_t^2$ *is the QMLE and*

$$\widehat{\sigma}_n^2(h) = \frac{1-\widehat{\phi}_n^{h+1}}{1-\widehat{\phi}_n}\frac{1}{n}\Big(\sum_{t=2}^n (X_t - \widehat{\phi}_n X_{t-1})^2 + X_1^2(1-\widehat{\phi}_n)\Big)$$

is the estimation of the variance. Then the variance and the length of the interval of prediction does not depend on the present variability of the time series as shown in Figure 14

Chapter 4

GARCH models

In order to estimate risk indicators more adaptive to the actual variability of the observed time series, the concept of volatility has been introduced:

Definition 38. *Consider a second order stationary time series. Its volatility at time t is its conditional variance given the past*

$$\sigma_t^2 = Var(X_t \mid X_{t-1}, X_{t-2}, \ldots).$$

Notice that the volatility is a *predictable* process in the sense that at time t it depends on the past. Assuming the gaussian assumption on the conditional distribution, a better 1-step prediction interval from an ARMA model is given by

$$[\Pi_n(\widehat{\theta}_n)(X_{n+1}) - q^N_{1-\alpha/2}\sigma^2_{n+1}, \Pi_n(\widehat{\theta}_n)(X_{n+1}) + q^N_{1-\alpha/2}\sigma^2_{n+1}],$$

where σ^2_{n+1} is the volatility at time $n+1$. It produces nice risk indicators and the length of the interval of prediction adapts to the present volatility of the time series. As the volatility is predictable, one can estimate it thanks to some model different than ARMA models.

We consider (Z_t) an observed WN. This WN is actually most of the time the residuals (innovations) of an ARMA model fitted by the QMLE in a first step of the analysis.

Definition 39. *The GARCH(p, q) model (Generalized Autoregressive Conditional Heteroscedastic) is solution, if it exists, of the system:*

$$\begin{cases} Z_t = \sigma_t W_t, \qquad t \in \mathbb{Z}, \\ \sigma_t^2 = \omega + \beta_1\sigma^2_{t-1} + \cdots + \beta_p\sigma^2_{t-p} + \alpha_1 X^2_{t-1} + \cdots + \alpha_q Z^2_{t-q}, \end{cases}$$

with $\omega > 0$, $\alpha_i, \beta_i \geq 0$ and $(W_t) \in SWN(1)$.

Remark. *If $\beta_i = 0$, $1 \leq i \leq p$, GARCH(0,q)=ARCH(q). If $\alpha_i = 0$, $1 \leq i \leq q$, $\sigma_t^2 = \omega/(1 - \beta_1 + \ldots + \beta_p)$ is degenerate.*

In the sequel, we focus for simplicity on $p = q = 1$.

4.1 Existence and moments of a GARCH(1,1)

We say that (Z_t) is a non-anticipative solution of a GARCH(1,1) model if $Z_t \in \mathcal{F}_t = \sigma(W_s, s \leq t)$.

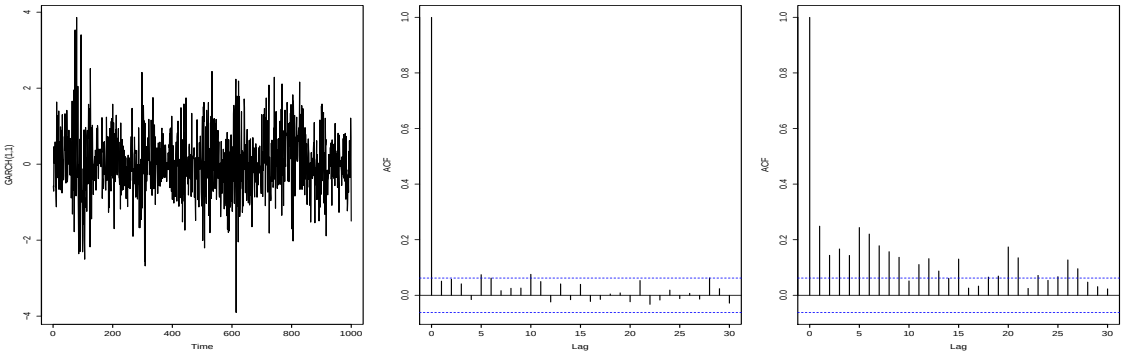

Figure 4.1: A trajectory and the corresponding ACF of the solution of a GARCH(1,1) model and its squares (to be compared with the SWN case)

Proposition. *A GARCH(1,1) model such that $\alpha + \beta < 1$ has a non-anticipative solution (Z_t) which is a stationary WN($\sigma^2 := \omega/(1-\alpha+\beta)$. Then $\sigma_t^2 = Var(Z_t \mid Z_{t-1}, Z_{t-2}, \ldots)$ is the predictable ($\sigma_t^2 \in \mathcal{F}_{t-1}$) volatility of (Z_t).*

Proof. Write $\sigma_t^2 = \omega + (\beta + \alpha W_{t-1}^2)\sigma_{t-1}^2$ as an AR(1) model with random coefficients. We have an explicit solution, which is non-anticipative and stationary (if the series converges)

$$\sigma_t^2 = \omega + (\beta + \alpha W_{t-1}^2)\left(\omega + (\beta + \alpha W_{t-2}^2)\sigma_{t-2}^2\right) = \omega\left(\sum_{j=1}^{+\infty}\prod_{k=1}^{j}(\beta + \alpha W_{t-k}^2) + 1\right)$$

Let $Y_j = \prod_{k=1}^{j}(\beta + \alpha W_{t-k}^2)$. As soon as $\sum_{j=1}^{+\infty}\mathbb{E}\left[|Y_j|\right] < +\infty$, the series $\sum_{j=1}^{+\infty} Y_j$ converges a.s. absolutely. We have:

$$\mathbb{E}\left[|Y_j|\right] = \mathbb{E}\left[\prod_{k=1}^{j}(\beta + \alpha W_{t-k}^2)\right] = \prod_{k=1}^{j}\mathbb{E}\left[\beta + \alpha W_{t-k}^2\right] = (\beta+\alpha)^j$$

If $\alpha + \beta < 1$, then $\sum_{j=1}^{+\infty}(\beta+\alpha)^j < +\infty$ and σ_t^2 a.s. exists, is predictable and $\mathbb{E}[\sigma_t^2] = \sigma^2$. So $Z_t = \sigma_t W_t$ exists and $\mathbb{E}[Z_t^2] = \mathbb{E}[\sigma_t^2 W_t^2] = \mathbb{E}[\sigma_t^2]$ because $\mathbb{E}[W_t^2] = 1$ and σ_t^2 is predictable. Moreover, $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = \sigma_t \mathbb{E}[W_t \mid \mathcal{F}_{t-1}] = 0$ and, for $s < t$, $\mathbb{E}[Z_s Z_t] = \mathbb{E}[Z_s \sigma_t \mathbb{E}[Z_t \mathcal{F}_{t-1}]] = 0$. □

Remark.

- *The volatility $\sigma_t^2 = \omega + \beta\sigma_{t-1}^2 + \alpha Z_{t-1}^2$ is also invertible if $\beta < 1$, i.e. $\sigma_t^2 = \sigma(W_{t-1}, \sigma_{t-1}^2, \sigma_{t-2}^2, \ldots)$.*
- *The WN is* unpredictable*, i.e. $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = 0$ so that the best prediction is* 0*. One also say that (Z_t) is a martingale differences sequence.*

If $|X_{t-1}|$ is large, then $\sigma_t^2 \geq \alpha X_{t-1}^2$ is too and thus X_t has a large conditional variance. We talk about periods of high volatility. Thanks to non-linearity, the model captures a conditionally heteroscedastic behavior, which we observe in finance for example.

The stationary solution of a GARCH(1,1) exists under much weaker solution. Stationary solutions that are not second order stationary satisfies $\mathbb{E}[Z_t^2] = \infty$, one says they are *heavy tailed*.

Theorem. *If $\mathbb{E}\left[\log(\beta + \alpha W_0^2)\right] < 0$ and $\mathbb{E}\left[|\log(\beta + \alpha W_0^2)|\right] < \infty$, then the GARCH(1,1) model has a (strictly) stationary solution.*

Proof. Let (Y_t') iid, $Y_t' = \log(\beta + \alpha W_t^2)$. By the strong law of large numbers:

$$\frac{1}{n}\sum_{t=1}^{n} Y_t' \xrightarrow{\text{a.s.}} \mathbb{E}\left[Y_0\right] = \mathbb{E}\left[\log(\beta + \alpha W_0^2)\right] < +\infty$$

Besides, $\sum_{j=1}^{n} Y_j = \sum_{j=1}^{n} \prod_{k=1}^{j} (\alpha W_{t-k}^2 + \beta)$ converges a.s. absolutely if it satisfies the Cauchy criteria. Let us show that $Y_j^{\frac{1}{j}} \xrightarrow{\text{a.s.}} \rho$ with $\rho < 1$.

$$\begin{aligned}
\mathbb{P}\left(Y_j^{\frac{1}{j}} \longrightarrow \rho\right) = 1 &\Leftrightarrow \mathbb{P}\left[\left(\prod_{k=1}^{j} \alpha W_{t-k}^2 + \beta\right)^{\frac{1}{j}} \longrightarrow \rho\right] = 1 \\
&\Leftrightarrow \mathbb{P}\left[\exp\left(\frac{1}{j}\sum_{t=1}^{j} Y_t'\right) \longrightarrow \rho\right] = 1 \\
&\Leftrightarrow \mathbb{P}\left(\frac{1}{j}\sum_{t=1}^{j} Y_t' \longrightarrow \log\,\rho\right) = 1
\end{aligned}$$

This equality is true with $\log\,\rho = \mathbb{E}\left[\log(\beta + \alpha W_0^2)\right] < 0$. □

Remark. *If $\alpha+\beta < 1$, then by Jensen's inequality $\mathbb{E}\left[\log(\beta + \alpha W_0^2)\right] \leq \log\left(\mathbb{E}\left[\beta + \alpha W_0^2\right]\right) = \log(\alpha + \beta) < 0$.*

Example 15. *Consider the ARCH(1) model with $\beta = 0$ et $W_0 \sim \mathcal{N}(0,1)$, then $\mathbb{E}\left[\log(\alpha W_0^2)\right] < 0 \Leftrightarrow \alpha < 2e^{\gamma} \simeq 3,56$. The stationary condition is much weaker than the second order stationary condition $\alpha < 1$ (as $\beta = 0$).*

Remark. *The GARCH(1,1) model under the condition $\mathbb{E}\left[\log(\beta + \alpha W_0^2)\right] < 0$ ($\Rightarrow \beta < 1$) is invertible:*

$$\sigma_t^2 = \sum_{j=0}^{+\infty} \beta^j(\omega + \alpha Z_{t-j-1}^2), \qquad t \in \mathbb{Z}.$$

The GARCH model is a special case of a *stochastic volatility model*. We call stochastic volatility model (X_t) a solution of

$$\begin{cases} Z_t = \sigma_t W_t, \qquad t \in \mathbb{Z}, \\ \sigma_t > 0 \text{ is a predictable non anticipative sequence.} \end{cases}$$

4.2 The Quasi Maximum Likelihood for GARCH models

Let us consider the QML approach for constructing an M-estimator for a GARCH(1,1) model $(Z_t(\theta))_{t\in\mathbb{Z}}$ with $\theta = (\omega, \alpha, \beta) \in \mathbb{R}^3$. Assume that (W_t) is gaussian $\mathcal{N}(0,1)$ and that $\mathbb{E}[\log(\beta + \alpha W_0^2)] < 0$ such that the conditional log-likelihood of the stationary model is

$$-2\log(f(Z_t(\theta) \mid Z_{t-1}(\theta), Z_{t-2}(\theta), \ldots)) = \log(\sigma_t^2(\theta)) + \frac{Z_t(\theta)^2}{\sigma_t^2(\theta)}$$

as

$$\sigma_t^2(\theta) = \sum_{j=0}^{+\infty} \beta^j(\omega + \alpha Z_{t-j-1}(\theta)^2)$$

is invertible because $\beta < 1$. We also have

$$\sigma_t^2(\theta) = \omega + \beta\sigma_{t-1}^2(\theta) + \alpha Z_{t-1}(\theta)^2, \qquad t \in \mathbb{Z},$$

which is observable for $t \geq 2$. We observe only $Z_1, \ldots, Z_n$ and we approximate $\sigma_t^2(\theta)$ with $\widehat{\sigma}_t^2(\theta)$ such that

$$\widehat{\sigma}_t^2(\theta) = \omega + \beta\widehat{\sigma}_{t-1}^2(\theta) + \alpha Z_{t-1}^2, \qquad \text{from} \qquad \widehat{\sigma}_0^2(\theta) \text{ arbitrary}, \tag{4.1}$$

The approximation error is a.s. bounded as $O(\beta^t)$.

Definition 40. *The QMLE is the M-estimator defined as*

$$\widehat{\theta}_n \in \arg\min_{\Theta} \sum_{t=1}^{n} \log(\widehat{\sigma}_t^2(\theta)) + \frac{Z_t^2}{\widehat{\sigma}_t^2(\theta)}$$

where $\Theta = (0,\infty) \times [0,\infty) \times [0,1)$ *and* $(\widehat{\sigma}_t^2(\theta))$ *is defined recursively thanks to* (4.1).

Notice that the condition $\mathbb{E}[\log(\beta + \alpha W_0^2)] < 0$ is not explicit and cannot be used in the definition of the QMLE. It is enough to ensure that the model is invertible $\beta < 1$ so that the arbitrary initial choice in (4.1) is not important.

Assume that (Z_t) is WN(σ^2). The QLik risk is, using the tower property,

$$\mathbb{E}\left[\log\left(\sum_{j=0}^{+\infty}\beta^j(\omega+\alpha Z_{t-j-1}^2)\right) + \frac{Z_0^2}{\sum_{j=0}^{+\infty}\beta^j(\omega+\alpha Z_{t-j-1}^2)}\right]$$
$$= \mathbb{E}\left[\log\left(\sum_{j=0}^{+\infty}\beta^j(\omega+\alpha Z_{t-j-1}^2)\right) + \frac{\sigma_0^2}{\sum_{j=0}^{+\infty}\beta^j(\omega+\alpha Z_{t-j-1}^2)}\right],$$

where σ_0^2 is the true volatility. The integrand is larger than 1 and equal to one iff $\sigma_0^2 = \sum_{j=0}^{+\infty}\beta^j(\omega+\alpha Z_{t-j-1}^2)$ a.s.. Thus, the QLik risk is minimized by the volatility satisfying the GARCH(1,1) equation that is the closest to the true volatility. Notice that the risk is not equivalent to the square risk as it was the case for the ARMA model. Actually, it is very robust to heavy tailed (Z_t). Even if then the volatility does not exist when $\mathbb{E}[Z_0^2] = \infty$, the QMLE for GARCH(1,1) is very useful to build risk indicators and prediction intervals. We have

Theorem. *Assume that* (Z_t) *is a stationary and ergodic time series so that* $\mathbb{E}\log^+(Z_0)^2 < \infty$. *Then the QMLE converges to the set of minimizers of the QLik risk*

$$d(\widehat{\theta}_n, \Theta_0) \to 0, \qquad a.s.$$

If moreover $(\widehat{\theta}_n)$ *converges to* $\theta_0 \in \overset{\circ}{\Theta}$ *and* (Z_t) *satisfies a volatility model* $Z_t = \sigma_t W_t$ *with* (W_t) *SWN(1) and* $\mathbb{E}[W_0^4] < \infty$ *then*

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d.} \mathcal{N}_3\left(0, (\mathbb{E}[W_0^4]-1)\mathbb{E}\left[\frac{\nabla_\theta\sigma_0^2(\theta_0)\nabla_\theta\sigma_0^2(\theta_0)^\top}{\sigma_0^4(\theta_0)}\right]^{-1}\right).$$

In particular we have the identities

$$\mathrm{Var}\,(S_0) = \frac{\mathbb{E}[W_0^4]-1}{4}\mathbb{E}\left[\frac{\nabla_\theta\sigma_0^2(\theta_0)\nabla_\theta\sigma_0^2(\theta_0)^\top}{\sigma_0^4(\theta_0)}\right]$$
$$\mathcal{I}(\theta_0) = \frac{1}{2}\mathbb{E}\left[\frac{\nabla_\theta\sigma_0^2(\theta_0)\nabla_\theta\sigma_0^2(\theta_0)^\top}{\sigma_0^4(\theta_0)}\right].$$

The QMLE is efficient only if (W_t) gaussian WN(1). In this case $\mathbb{E}[W_0^4] - 1 = 2$ and the inverse of the Fisher information is the Cramer-Rao bound.

4.3 Simple testing on the coefficients

4.3.1 Tests of nullity

Having computed the QMLE $(\widehat{\theta}_n)$, a natural issue is overfitting. Thus, one will test whether one can reject the null hypothesis

1. ARCH model $\theta_3 = \beta = 0$, and then,
2. SWN model $\theta_2 = \alpha = 0$.

To do so, one will construct a region of reject of the form $\widehat{\theta}_i > c_i$ for some constant c_i well chosen. Assuming the conditions of the asymptotic normality met, one will denote the asymptotic variances

$$\mathrm{se}_i^2 = (\mathbb{E}[W_0^4] - 1)\mathbb{E}\left[\frac{\nabla_\theta \sigma_0^2(\theta_0)\nabla_\theta \sigma_0^2(\theta_0)^\top}{\sigma_0^4(\theta_0)}\right]^{-1}_{ii}.$$

Assume that the asymptotic properties still hold on the boundary of the parameter set Θ_0 so that $\widehat{\theta}_i \approx (\mathrm{se}_i/\sqrt{n})\max\{N, 0\}$ for N a standard gaussian r.v. under the null hypothesis. Denoting Φ the distribution function of N and using the independence of N and $\mathbb{1}_{N>0}$ which is Bernoulli$(1/2)$ distributed, the p-value of the test is

$$\mathbb{P}(\sqrt{n}\widehat{\theta}_i/\mathrm{se}_i < \max\{N, 0\}) = \Phi(-\sqrt{n}\widehat{\theta}_i/\mathrm{se}_i)$$

the smallest level of the test that rejects the null hypothesis, i.e. the probability to reject the null hypothesis abusively. Note that due to the boundary effect we have $\widehat{\theta}_i \geq 0$ and the p-value is necessary smaller than $1/2$.

One issue arises: there is no explicit expression of se_i in term of θ so one has to estimate the asymptotic variance in another way than the usual plug-in method $\theta = \widehat{\theta}_n$. To do so, we differentiate the recursive equation (4.1) followed by $\widehat{\sigma}_t^2(\theta)$

$$\nabla\widehat{\sigma}_t^2(\theta) = \begin{pmatrix} 1 \\ Z_{t-1}^2 \\ \widehat{\sigma}_{t-1}^2(\theta) \end{pmatrix} + \beta\nabla\widehat{\sigma}_{t-1}^2(\theta),$$

staring from an arbitrary initial value that is forgotten exponentially fast when $\beta < 1$. Thus one can approximate

$$\mathbb{E}\left[\frac{\nabla_\theta \sigma_0^2(\theta_0)\nabla_\theta \sigma_0^2(\theta_0)^\top}{\sigma_0^4(\theta_0)}\right] \approx \frac{1}{n}\sum_{t=1}^{n} \frac{\nabla\widehat{\sigma}_t^2(\widehat{\theta}_n)\nabla\widehat{\sigma}_t^2(\widehat{\theta}_n)}{\widehat{\sigma}_t^2(\widehat{\theta}_n)^2},$$

invert the approximation and estimate

$$\mathbb{E}[W_0^4] - 1 \approx \frac{1}{n}\sum_{t=1}^{n} \widehat{W}_t^4 - 1$$

where $\widehat{W}_t = Z_t/\widehat{\sigma}_t(\widehat{\theta}_n)$ are the residuals of the GARCH(1,1) model. Doing so, one obtains a consistent estimator of se_i.

Another issue arises: there is no uniqueness of $\widehat{\theta}_0$ under the null $\alpha = 0$ as then the random volatility is degenerate to the constant $\omega/(1-\beta)$. The asymptotic normality of the QMLE could not hold in this case. The idea is to check first whether $\beta = 0$, if yes then use the QMLE computed for the ARCH(1) model (adapting the previous construction under the constraint $\beta = 0$) and then test $\alpha = 0$ on the obtained $\widehat{\alpha}_n$.

4.3.2 Test of second order stationarity

Another natural test is weather the fitted model satisfied the second order condition $\alpha+\beta < 1$. Under the null hypothesis $\alpha+\beta \geq 1$ and $\mathbb{E}[\log(\beta+\alpha Z_0^2)] < 0$, we have $(\omega_0, \alpha_0, \beta_0) \in \overset{\circ}{\Theta}_0$ when $\alpha_0+\beta_0 = 1$ and $0 < \beta_0 < 1$. θ_0 is uniquely determined as the minimizer of the QLik risk and the asymptotic normality holds. We have

$$\sqrt{n}(\widehat{\alpha}_n + \widehat{\beta}_n - 1) \xrightarrow{d.} \mathcal{N}(0, \mathrm{se}_2^2 + \mathrm{se}_3^2 + 2c_{23})$$

where

$$c_{23} = \mathbb{E}[W_0^4] - 1)\mathbb{E}\left[\frac{\nabla_\theta \sigma_0^2(\theta_0)\nabla_\theta \sigma_0^2(\theta_0)^\top}{\sigma_0^4(\theta_0)}\right]_{23}^{-1}$$

can be consistently estimated in the same way than in the previous subsection.

The p-value of the corresponding test, with reject region of the form $\widehat{\alpha}_n + \widehat{\beta}_n < 1 - c$ for some constant $c > 0$, is

$$\mathbb{P}\left(N \leq \sqrt{n}(\widehat{\alpha}_n + \beta_n - 1)/\sqrt{\mathrm{se}_2^2 + \mathrm{se}_3^2 + 2c_{23}}\right) = \Phi\left(\sqrt{n}(\widehat{\alpha}_n + \beta_n - 1)/\sqrt{\mathrm{se}_2^2 + \mathrm{se}_3^2 + 2c_{23}}\right)$$

because the rejection region is one-sided.

4.3.3 Invertibility test

If $\widehat{\beta}_n \lesssim 1$ under the constraint $\beta < 1$, which is often the case in finance, it is legitimate to ask whether the condition of invertibility is satisfied. If one assumes that under the null $\beta \geq 1$ and $\mathbb{E}[\log(\beta+\alpha Z_0^2)] > 0$ then one can proceed to a test rejecting on β. Under $\mathbb{E}[\log(\beta+\alpha Z_0^2)] > 0$, as $\sigma_t^2 > 0$, it is not difficult to prove that $\sigma_t^2 \to +\infty$ infinitely fast from $\sigma_0^2 = 0$. Thus, we are in an explosive case where the heteroscedasticity yields unstability and the variability will always increase. In that situation, the initial arbitrary value in the recursive formula (4.1) defining the QMLE is not important. What matters is the rate of divergence of the volatility which is driven by the coefficients (α, β). In this context Θ should be chosen equal to $(0, \infty)^3$ and not restricted over $[0, 1)$ for β in order to let $\widehat{\beta}_n$ be larger than 1. One can show that the QMLE is asymptotically normal when the model is well specified

$$\sqrt{n}(\widehat{\beta}_n - \beta_0) \xrightarrow{d.} \mathcal{N}(0, \mathrm{se}^2)$$

where

$$se^2 = \frac{(1+\mu_1)\mu_2}{\beta_0^2(1-\mu_1)(1-\mu_2)}$$

with

$$\mu_i = \mathbb{E}\left[\left(\frac{\beta_0}{\alpha_0 W_0^2 + \beta_0}\right)^i\right]$$

Notice that se can be estimated from the residuals $\widehat{W}_t$ and plugging in $\widehat{\beta}_n$. The p-value of the test with reject region of the form $\widehat{\beta}_n < c$ is of the form

$$\mathbb{P}(N \leq \sqrt{n}(\widehat{\beta}_n - 1)/se) = \Phi(\widehat{\beta}_n - 1)/se).$$

Notice that if one cannot reject the test (the p-value is too large) then we are not confident in being in the invertible domain. In that case, one suspects that the stationary condition is not satisfied on the centered (Z_t) that may have the behavior of a centered random walk. One should try to difference the original process one more time as, for instance, there is no consistent estimator of ω and the volatility is not predictable.

4.4 Intervals of prediction

Once we found the good volatility model for the conditional variance (GARCH(1,1), ARCH(1) or a constant from the previous discussion), the volatility is predicted by

$$\widehat{\sigma}^2_{n+1}(\widehat{\theta}_n) = \widehat{\omega}_n + \widehat{\beta}_n\widehat{\sigma}^2_n(\widehat{\theta}_n) + \widehat{\alpha}_n Z_n^2.$$

Thus we obtain the interval of prediction of confidence level α as

$$\widehat{I}_\alpha(Z_{n+1}) = [-q^N_{1-\alpha/2}\widehat{\sigma}_{n+1}(\widehat{\theta}_n), q^N_{1-\alpha/2}\widehat{\sigma}_{n+1}(\widehat{\theta}_n)].$$

It is centered on 0 and the point prediction is useless. However the length of the interval is very useful for risk assessment. Similarly, one can produce h step prediction intervals using the recursion

$$\widehat{\sigma}^2_{n+h}(\widehat{\theta}_n) = \widehat{\omega}_n + (\widehat{\beta}_n + \widehat{\alpha}_n)\widehat{\sigma}^2_{n+h-1}(\widehat{\theta}_n), \qquad h \geq 1,$$

estimating Z^2_{n+h-1} non observed by $\widehat{\sigma}^2_{n+h-1}(\widehat{\theta}_n)$.

As the volatility of the noise is also the volatility of the original process, one can build from the two-stage estimation (QML approach on (X_t) with the ARMA model and on the residuals $(Z_t) = (I_t(\widehat{\theta}_n))$ with the volatility model) a prediction interval on X_{n+h}. Recall that $\widehat{\sigma}^2_n(h)$ is the estimation of $\mathbb{E}[(X_{n+h} - \Pi_n(X_{n+h}))^2] \geq \sigma^2$. Denote $\widehat{\sigma}^2_n$ the approximation of $\sigma^2 = R^L_\infty$. We can build

$$\widehat{I}_\alpha(X_{n+h}) = [\Pi_n(\widehat{\theta}_n)(X_{n+h}) - q^N_{1-\alpha/2}\sqrt{\widehat{\sigma}^2_n(h) - \widehat{\sigma}^2_n + \widehat{\sigma}^2_{n+h}(\widehat{\theta}_n)};$$
$$\Pi_n(\widehat{\theta}_n)(X_{n+h}) + q^N_{1-\alpha/2}\sqrt{\widehat{\sigma}^2_n(h) - \widehat{\sigma}^2_n + \widehat{\sigma}^2_{n+h}(\widehat{\theta}_n)}],$$

with some abuse of notation as there is two different $\widehat{\theta}_n$, one for the ARMA and another for the GARCH.

Since the GARCH modeling holds on the residual of an ARMA model, this two-step procedure does not respect the flow of information. Actually, one could also consider the likelihood of

$$\begin{aligned} X_t &= \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + \gamma_1 Z_{t-1} + \cdots + \gamma_q Z_{t-q}, \\ Z_t &= \sigma_t W_t, \qquad t \in \mathbb{Z} \\ \sigma^2_t &= \omega + \beta\sigma^2_{t-1} + \alpha\varepsilon^2_{t-1}, \end{aligned}$$

under the assumption that (W_t) is a gaussian WN(1). Then the parameters are

$$\theta = (\phi_1, \ldots, \phi_p, \gamma_1, \ldots, \gamma_q, \omega, \alpha, \beta)^\top \in \mathbb{R}^d, \qquad d = p + q + 3,$$

is estimated by the QMLE minimizing $\widehat{L}_n(\theta) = \sum_{t=1}^n \widehat{\ell}_t(\theta)$ computed recursively as follows: Starting from arbitrary initial values, observing recursively X_t,

1. compute the approximative innovation $\widehat{I}_t(\theta) = X_t - \widehat{X}_t(\theta)$ and the QLIK loss $\widehat{\ell}_t(\theta) = \log(\widehat{\sigma}^2_t(\theta)) + \widehat{I}_t(\theta)^2/\widehat{\sigma}^2_t(\theta)$,
2. update the variance of the WN $\widehat{\sigma}^2_{t+1}(\theta) = \omega + \beta\widehat{\sigma}^2_t(\theta) + \alpha\widehat{I}_t(\theta)^2$,
3. predict the next observation $\widehat{X}_{t+1}(\theta) = \phi_1 X_t + \cdots + \phi_p X_{t-p+1} + \gamma_1\widehat{I}_t(\theta) + \cdots + \gamma_q\widehat{I}_{t-p+1}(\theta)$.

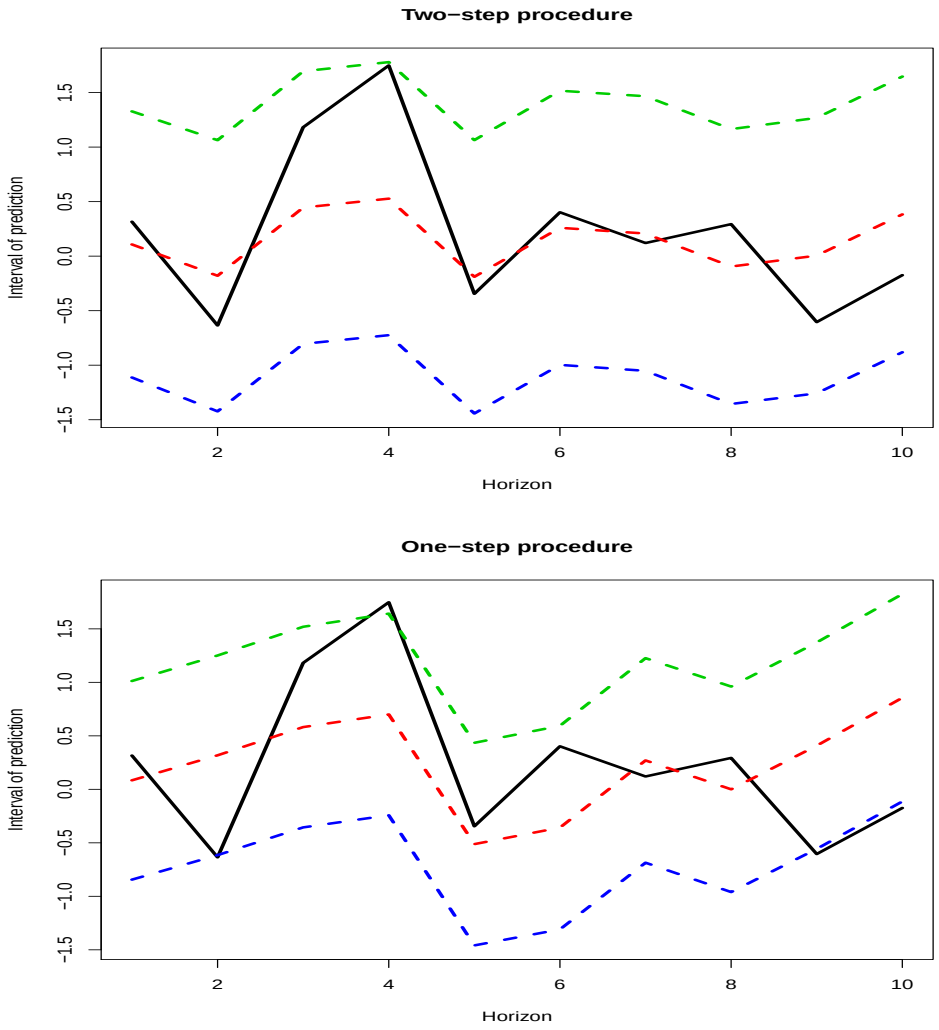

Figure 4.2: Weekly Monetary Supply forecasting in March 2015 training over 40 years, data from `https://research.stlouisfed.org`. Two-step and one-step procedures on ARMA(10,13)-GARCH(1,1).

This one-step QMLE is strongly consistent

Theorem (Francq and Zakoian (2019)). *If the observations satisfy the ARMA(p,q)-GARCH(1,1) model with $\theta_0 \in \Theta$ satisfying the condition of stationarity of the GARCH model the Hannan's condition $\mathcal{C}$ and $\beta < 1$, then the QMLE is strongly consistent. If it satisfies the condition for finite moments of order 4 of the gaussian GARCH model then it is asymptotically normal.*

The advantage of this one-step procedure is that it respects the flow of the information. However the asymptotic normality is achieved under necessary 4th order moment conditions that may be optimistic since risky time series may not have such moment properties.

Part III

Online algorithms

Chapter 5

The Kalman filter

5.1 The state space models

By contrast with the AR models, it is much more difficult to find the best possible (linear) prediction of an ARMA model. Indeed, as soon as the MA part is non degenerate, the filter can have infinitely many non null coefficients. One way to circumvent the problem is to consider the ARMA model as a more general linear model called state space models. Those models have been introduced in signal processing and the best linear prediction can be computed recursively by the Kalman's recursion.

Definition 41. *A state space linear model of dimension r with constant coefficient is given by a system of a space equation and state equations of the form*

$$\begin{cases} X_t = G^\top \mathbf{Y}_t + Z_t, & \text{Space equation,} \\ \mathbf{Y}_t = \mathbf{F}\mathbf{Y}_{t-1} + \mathbf{V}_t, & \text{State equation.} \end{cases}$$

where (Z_t) and $(\mathbf{V}_t)$ are uncorrelated WN with variances σ^2 and $\mathbf{Q}$, $G \in \mathbb{R}^r$, $\mathbf{F} \in \mathcal{M}(r,r)$ and $\mathbf{Y} \in \mathbb{R}^r$ is the random state of the system.

In the cases were both (Z_t) and $(\mathbf{V}_t)$ are SWN the state-space models have a nice interpretation: the state $\mathbf{Y}$ is a Markov chain that governs the distribution of the observations X in the sense that conditionally on $(\mathbf{Y}_t)$ the X_t's are independent. It is a specific case of Hidden Markov model with continuous state. Notice that $(\mathbf{V}_t)$ is actually a WN in $\mathbb{R}^r$, meaning a weak stationary sequence of uncorrelated vectors with mean $0 \in \mathbb{R}^r$ and covariance matrix $\mathbf{Q}$. Notice that the different coordinates of the space $\mathbf{Y_t} = (Y_{1,t}, \dots, Y_{r,t})'$ can be correlated at each time t.

State space representations are not unique. We shall give two representations for an ARMA (p,q) model. The first one directly shows up from the compact equation $\phi(T)X_t = \gamma(T)Z_t$ and it has dimension $r = \max(p, q+1)$. Hereafter we use the convention that the coefficients $\phi_j = 0$ and $\gamma_j = 0$ for any $j > p$ and $j > q$ respectively. We can write

$$\begin{cases} X_t = (1, \gamma_1, \dots, \gamma_{r-1})^\top \mathbf{Y}_t, & \text{Space equation,} \\ \mathbf{Y}_t = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_r \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 1 & 0 \end{pmatrix} \mathbf{Y}_{t-1} + \begin{pmatrix} Z_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}, & \text{State equation.} \end{cases}$$

In the causal case, it is possible to establish a better representation, i.e. a state space representation with the lower dimension $r = \max(p, q)$:

$$\begin{cases} X_t = (1, 0, \dots, 0)^\top \mathbf{Y}_t + Z_t, & \text{Space equation,} \\ \mathbf{Y}_t = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 \\ \phi_r & \cdots & \phi_2 & \phi_1 \end{pmatrix} \mathbf{Y}_{t-1} + \begin{pmatrix} \psi_1 \\ \vdots \\ \vdots \\ \psi_r \end{pmatrix} Z_{t-1}, & \text{State equation.} \end{cases}$$

where $\psi_1, \dots, \psi_r$ are the coefficients of $z, \dots, z^r$ in the Laurent series ψ. For a proof of this result, see p.470-471 of B&D. This representation is called the canonical representation. It is very useful as $\mathbf{Y}_{t,h} = \Pi_{t-1}(X_{t+h-1})$, the h step prediction at time $t-1$. Notice also that in this representation $\mathbf{Y_t}$ is predictable.

Any ARMA model can be represented as a state-space model. Of course the contrary is not true. Consider for instance a time series (X_t) that could be predicted with k explanatory variables $\mathbf{X}_{t-1}$. Here explanatory variables are indexed by $t-1$ as they are supposed to be observed before the variable of interest. Then one can consider the state-space model

$$\begin{cases} X_t = (1, 0, \dots, 0)^\top \mathbf{Y}_t + Z_t, & \text{Space equation,} \\ \mathbf{Y}_t = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 \\ \phi_r & \cdots & \phi_2 & \phi_1 \end{pmatrix} \mathbf{Y}_{t-1} + \begin{pmatrix} \lambda_1^\top \\ \vdots \\ \vdots \\ \lambda_r^\top \end{pmatrix} \mathbb{X}_{t-1} + \begin{pmatrix} \psi_1 \\ \vdots \\ \vdots \\ \psi_r \end{pmatrix} Z_{t-1}, & \text{State equation,} \end{cases}$$

where $\mathbb{X}_{t-1}$ is a $k \times r$ matrix that stacks the vectors $\mathbf{X}_{t-1}, \dots, \mathbf{X}_{t-1}$ and the λ_i are coefficients of dimension k that quantifies the linear influence of the past $\mathbf{X}_{t-1}$ on the h step prediction $\mathbf{Y}_{t-1,h}$ at time $t-1$. Such system is called ARMAX state-space representation (see Hannan and Deistler (2012)) but this parametrization is not the unique one and suffers over-parametrization. One could prefer the parametrization such that $\lambda_i = \lambda$ for all $1 \leq i \leq k$. It is difficult to find the good representation for such models. We will not investigate further this model because of that drawback and we will prefer state-space models with random coefficients, see below.

5.2 The Kalman's recursion

To start the Kalman's recursion, let us take an arbitrary initial values $\widehat{\mathbf{Y}}_\mathbf{0}$ and Ω_0. Assume now that we have a recursive procedure providing at each step $\widehat{X}_n = \Pi_{n-1}(X_n)$, $R_n^L = \mathrm{Var}\,(I_n)$, $\widehat{\mathbf{Y}}_n = \Pi_{n-1}(\mathbf{Y}_n)$ and $\Omega_n = \mathbb{E}[(\mathbf{Y}_n - \widehat{\mathbf{Y}}_n)(\mathbf{Y}_n - \widehat{\mathbf{Y}}_n)^\top]$, the covariance matrix of the prediction error of the state $\mathbf{Y}_n$.

Let us compute $\widehat{X}_{n+1} = \Pi_n(X_n)$ in a recursive way. Applying the linear projection Π_n on the state equation $X_{n+1} = G^\top \mathbf{Y}_{n+1} + Z_{n+1}$ it is clear that

$$\widehat{X}_{n+1} = G^\top \widehat{\mathbf{Y}}_{n+1}.$$

By definition of the innovation I_n and the decomposition of Proposition 1.2.5, we have

$$\widehat{\mathbf{Y}}_{n+1} = \Pi_n(\mathbf{Y}_{n+1}) = \Pi_{n-1}(\mathbf{Y}_{n+1}) + P_{I_n}(\mathbf{Y}_{n+1}).$$

The first term is computed recursively using the space equation

$$\Pi_{n-1}(\mathbf{Y}_{n+1}) = \mathbf{F}\Pi_{n-1}(\mathbf{Y}_n) = \mathbf{F}\widehat{\mathbf{Y}}_n.$$

So it remains to compute recursively the second term $P_{I_n}(\mathbf{Y}_{n+1})$. By definition of the orthogonal projection, there exists $\theta \in \mathbb{R}^r$ such that $P_{I_n}(\mathbf{Y}_{n+1}) = \theta I_n$ and $\mathbf{Y}_{n+1} - \theta I_n \perp I_n$. So

$$\mathbb{E}[(\mathbf{Y}_{n+1} - \theta I_n)I_n] = 0 \Leftrightarrow \theta\mathbb{E}[I_n^2] = \mathbb{E}[\mathbf{Y}_{n+1}I_n].$$

We recognize the risk of linear prediction $\mathbb{E}[I_n^2] = R_n^L$. We can also compute recursively

$$\begin{aligned}
\mathbb{E}[\mathbf{Y}_{n+1}I_n] &= \mathbb{E}[\mathbf{Y}_{n+1}(G^\top(\mathbf{Y}_n - \widehat{\mathbf{Y}}_n) + Z_n)] \\
&= \mathbb{E}[(\mathbf{F}\mathbf{Y}_n + \mathbf{V}_n)(G^\top(\mathbf{Y}_n - \widehat{\mathbf{Y}}_n) + Z_n)] \\
&= \mathbb{E}[(\mathbf{F}(\mathbf{Y}_n - \widehat{\mathbf{Y}}_n)G^\top(\mathbf{Y}_n - \widehat{\mathbf{Y}}_n)] \\
&= \mathbf{F}\Omega_n G
\end{aligned}$$

by orthogonality of $\widehat{\mathbf{Y}}_n$ with $\mathbf{Y}_n - \widehat{\mathbf{Y}}_n$ and Z_n and of Z_n with $\mathbf{V}_n$ and $\mathbf{Y}_n$. So arranging all those terms, we derive the formula

$$\widehat{\mathbf{Y}}_{n+1} = \mathbf{F}\widehat{\mathbf{Y}}_n + \frac{\mathbf{F}\Omega_n G}{R_n^L}(X_n - G^\top\widehat{\mathbf{Y}}_n)$$

Let us denote $\mathbf{K}_n = \mathbf{F}\Omega_n G/R_n^L$ and call it the Kalman's gain. Finally, in order to apply the complete recursion, one has to compute Ω_{n+1} and R_{n+1}^L. Using the identity

$$\Omega_{n+1} = \mathbb{E}[\mathbf{Y}_{n+1}\mathbf{Y}_{n+1}^T] - \mathbb{E}[\widehat{\mathbf{Y}}_{n+1}\widehat{\mathbf{Y}}_{n+1}^\top]$$

together with the state equation and the recursive formula $\widehat{\mathbf{Y}}_{n+1} = \mathbf{F}\widehat{\mathbf{Y}}_n + \mathbf{K}_n I_n$, we obtain

$$\begin{aligned}
\Omega_{n+1} &= \mathbf{F}\mathbb{E}[\mathbf{Y}_n\mathbf{Y}_n^\top]\mathbf{F}^\top + \mathbf{Q} - \mathbf{F}\mathbb{E}[\widehat{\mathbf{Y}}_n\widehat{\mathbf{Y}}_n^\top]\mathbf{F}^\top - \mathbf{K_n}\mathbb{E}[I_n^2]\mathbf{K_n}^\top \\
&= \mathbf{F}\Omega_n\mathbf{F}^\top + \mathbf{Q} - \mathbf{K_n}G^\top\Omega_n\mathbf{F}^\top.
\end{aligned}$$

To compute R_{n+1}^L, we use the identity $I_{n+1} = X_{n+1} - G^\top\widehat{\mathbf{Y}}_{n+1} = G^\top(\mathbf{Y}_{n+1} - \widehat{\mathbf{Y}}_{n+1}) + W_{n+1}$ and by orthogonality between Z_n and the linear span of $\mathbf{Y}_{n+1}$ and $X_1, \ldots, X_n$:

$$R_{n+1}^L = \mathbb{E}[I_{n+1}^2] = \mathbb{E}[(G^\top(\mathbf{Y}_{n+1} - \widehat{\mathbf{Y}}_{n+1}) + W_{n+1})^2] = G^\top\Omega_{n+1}G + \sigma^2.$$

Finally, we have the following theorem

Theorem (Kalman (1960))**.** *In a state-space model with constant coefficients, if $\widehat{Y}_0$ and Ω_0 are well-chosen, one can compute recursively $\widehat{X}_n = \Pi_{n-1}(X_n)$, $R_n^L = \mathbb{E}[(X_n - \widehat{X}_n)^2]$, $\widehat{\mathbf{Y}}_n = \Pi_{n-1}(\mathbf{Y}_n)$ and $\Omega_n = \mathbb{E}[(\mathbf{Y}_n - \widehat{\mathbf{Y}}_n)(\mathbf{Y}_n - \widehat{\mathbf{Y}}_n)^\top]$ by the following recursion*

$$\begin{aligned}
\widehat{\mathbf{Y}}_{n+1} &= \mathbf{F}\widehat{\mathbf{Y}}_n + \frac{1}{R_n^L}\mathbf{F}\Omega_n G(X_n - G^\top\widehat{\mathbf{Y}}_n) \\
\widehat{X}_{n+1} &= G^\top\widehat{\mathbf{Y}}_{n+1} \\
\Omega_{n+1} &= \mathbf{F}\Omega_n\mathbf{F}^\top + \mathbf{Q} - \frac{1}{R_n^L}\mathbf{F}\Omega_n GG^\top\Omega_n\mathbf{F}^\top \\
R_{n+1}^L &= G^\top\Omega_{n+1}G + \sigma^2.
\end{aligned}$$

The Kalman's recursion has several advantages, even in for AR models when compared to the Yule Walker approach:

- It is a recursive procedures, particularly well suited in signal processing or high-frequency data, i.e. when observations are observed consecutively,
- Each step requires the inversion of a scalar R_n^L and not the entire covariance matrix,
- The recursion can handle missing values nicely.

The Kalman's recursion has one major drawback for statistical application: It requires to know the coefficients in the state and space equations. In practice, we want to estimate the parameters $\theta = (\phi_1, \ldots, \phi_p, \gamma_1, \ldots, \gamma_q)$ of an ARMA model. One way to conciliate this contradiction is to use the Bayesian approach. We will not pursue this approach here.

Two issues arise: the first one is about the regularity conditions that are related with optimization problems. This fundamental issue will not be treated in the notes as a diagnostic of convergence is usually provided by any procedure like *nlminb* in R. The second issue is about the condition on the past. As the past is not observed, it will be replaced by some arbitrary past and then it will be fundamental to check the stability of the procedure with respect to this arbitrary choice. This issue will constitute one major topic of these notes.

5.3 Application to state space models

Let us consider a model that fit into the class of the state space models. The gaussian assumption used to derive the QLik loss holds on $\mathbf{V}_t$ and Z_t non degenerate. Notice that to derive the QLik loss one can always restrict to the standard case $\mathrm{Var}\,(Z) = \sigma^2 = 1$. Then the linear risk of prediction is the standardized one $r_t^L = R_t^L/\sigma^2$. The natural filtration of the problem is $\mathcal{F}_t = \sigma(X_t, \ldots, X_1, \widehat{\mathbf{Y}}_0, \Omega_0)$ as under the iid assumption the state equation describes a Markov chain. Here θ correspond to the vector containing the parameters of the model, i.e. the elements of $\mathbf{F}$, G and $\mathbf{Q}$.

Conditionally on $\mathcal{F}_{t-1}$ the distribution of $G^\top \mathbf{Y}_t + Z_t$ in the model is a gaussian r.v. with mean $\Pi_{t-1}(G^\top \mathbf{Y}_t + Z_t) = G^\top \widehat{\mathbf{Y}}_t = \widehat{X}_t(\theta)$ and variance $\mathrm{Var}\,(G^\top(\mathbf{Y}_i - \widehat{\mathbf{Y}}_i)) + 1 = r_t^L(\theta)$. As both the reduced innovations $I_t^0(\theta) = X_t - \widehat{X}_t(\theta)$ and their standardized variances $r_t^L(\theta)$ are computing by the Kalman's recursion, we have the sequential algorithm

- *Initialization:* θ, initial values $\widehat{\mathbf{Y}}_{\mathbf{0}}(\theta)$ and $\Omega_0(\theta)$.
- *New observation* X_n*:*
  1. Compute the innovation $I_n^0(\theta) = X_n - \widehat{X}_n(\theta)$,
  2. Compute the next linear prediction $\widehat{X}_{n+1}(\theta)$ and the associated standardized risk $r_{n+1}^L(\theta)$ thanks to the Kalman's recursion.

From this sequential algorithm, it is then simple to derive the reduced Quasi Maximum Likelihood Estimator for state-space models:

Definition 42. *The QMLE of a stat-space model is defined as a minimizer*

$$\widehat{\theta}_n \in \arg\min_{\Theta} L_n^0(\theta) = \arg\min_{\Theta} \sum_{t=1}^{n} \frac{I_t^{0\,2}(\theta)}{\sigma^2(\theta) r_t^L(\theta)} + \log(\sigma^2(\theta) r_t^L(\theta))$$

where $I_t^0(\theta)$ *and* $r_t^L(\theta)$ *are defined recursively thanks to the standardized procedure described above assuming that* $R = \sigma^2 = 1$. *An estimator of* σ^2 *is provided by*

$$\sigma^2(\widehat{\theta}_n) = \frac{1}{n} \sum_{t=1}^{n} \frac{I_t^{0\,2}(\widehat{\theta}_n)}{r_t^L(\widehat{\theta}_n)}$$

Notice that, neglecting the optimization issues, one should write $\theta_n(\widehat{\mathbf{Y}}_\mathbf{0})$ as the whole procedure depends on the initial state $\widehat{\mathbf{Y}}_\mathbf{0}(\theta)$ chosen arbitrarily in practice, because the initial distribution P_{θ_0} driving the observations is unknown.

Chapter 6

State-space models with random coefficients

6.1 Linear regression with time-varying coefficients

Assume that we observe some variable of interest (X_t) together with some explanatory variables $\mathbf{X}_{t-1} \in \mathbb{R}^k$. Here again we index the explanatory variables with $t-1$ and consider that there are observed before X_t such that one can use them to build prediction intervals. In statistics, the most usual model to fit a prediction is the linear regression one

$$X_t = \theta^T \mathbf{X}_{t-1} + Z_t, \qquad t \in \mathbb{Z}.$$

The unknown parameter $\theta \in \mathbb{R}^k$ is usually estimated thanks to the Ordinary Mean Squares (OMS) which is equivalent to the MLE under the gaussian assumption on (Z_t). The only difference with the time series setting is that $(X_t, \mathbf{X}_{t-1})$ is considered iid. Most of the time, Y'_{t-1} is even considered deterministic. One calls this setting the fixed design setting. It is very close to the time series setting as, in the latter case, we used the principle of conditioning on the past so that, at time t, $\mathbf{X}_{t-1}$ is considered as fixed.

Example 16. *Consider $\mathbf{X}_{t-1} = (X_{t-1}, \dots, X_{t-k})^\top \in \mathbb{R}^k$ then the linear model is equivalent to an AR(k) model. For $k = 1$, the OMS*

$$\frac{\sum_{t=2}^n X_t X_{t-1}}{\sum_{t=2}^n X_t^2} \approx \widehat{\theta}_n$$

approximates the QMLE. The only difference is the denominator $\sum_{t=2}^n X_t^2$ instead of $\sum_{t=1}^n X_t^2$ so that the constraint of stationarity (absolute value of the estimated coefficient less than one) is not satisfied for the OMS.

In this chapter, we investigate the time-varying model

$$X_t = \theta_t^T \mathbf{X}_{t-1} + Z_t, \qquad t \in \mathbb{Z}.$$

We will first see the properties of the simple time-varying model when $\mathbf{X}_{t-1} = X_{t-1}$ and then see how the Kalman's recursion can be used to estimate the (time varying) parameter (θ_t).

6.2 The unit root problem and Stochastic Recurrent Equations (SRE)

One of the most interesting application of the random coefficients setting is to consider the auto regressive case $\mathbf{X}_{t-1}$ equals to the observation X_t (and then $k = 1$):

$$X_t = \theta_t X_{t-1} + Z_t, \qquad t \in \mathbb{Z}.$$

Such model has various nice properties, depending on the behavior of the time-varying coefficients (θ_t).

Consider the case (θ_t) is iid $\mathcal{N}(\phi, \beta)$. Then, denoted $\theta_t = \sqrt{\beta}N_t + \phi$ with (N_t) standard normal, we obtain the identity (in distribution)

$$X_t = \theta_t X_{t-1} + Z_t = \phi X_{t-1} + \sqrt{\beta} N_t X_{t-1} + Z_t, \qquad t \in \mathbb{Z}.$$

It is an SRE, i.e. an auto-regressive model with random coefficients. Notice that the volatility of the GARCH model satisfies such recursion too. The special case $\phi = 1$ is not excluded as the stationary solution condition is

$$\mathbb{E}[\log(|\theta_0|)] = \mathbb{E}[\log(|\phi + \sqrt{\beta}N_0|)] < 0.$$

Actually one can choose of ϕ as big as 1.25 by choosing accordingly the value of β. The stationary solution of such SRE exhibits heavy tails comparable to Pareto distribution

Theorem (Goldie et al. (1991))**.** *Under the stationary condition, there exists a unique $\alpha > 0$ such that*

$$\mathbb{E}[|\theta_0|^\alpha] = 1.$$

Under some other conditions on the distribution of V_0, there exists a coefficient $c > 0$ such that

$$\mathbb{P}(X_0 > x) \sim_{x\to\infty} \frac{c}{2}x^{-\alpha}, \qquad \mathbb{P}(X_0 \leq -x) \sim_{x\to\infty} \frac{c}{2}x^{-\alpha}.$$

The parameter α is the index of heavy tail. The time series (X_t) admits finite moments of order $p < \alpha$ and infinite moments of order $p > \alpha$. Goldie Theorem is very important as the SRE solution appears as natural heavy-tailed time series including risk indication through α.

For $\phi = 1$, one can easily check that necessarily $\alpha < 2$ meaning that the time series (X_t) does not have finite variance. The second order stationarity condition $\phi^2 + \beta < 1$ is not satisfied. An AR model with a random coefficient models naturally an ARCH effect:

Proposition (Klüppelberg et al. (2004))**.** *The SRE with (Z_t) gaussian WN(ω) with $\omega > 0$ is equivalent to the AR(1)-ARCH(1) model*

$$\begin{cases} X_t &= \phi X_{t-1} + Z_t, \\ Z_t &= \sigma_t W_t, \\ \sigma_t^2 &= \omega + \beta X_{t-1}^2, \qquad t \in \mathbb{Z}, \end{cases}$$

where W_t are gaussian WN(1).

Take care that the Z_t of the SRE and AR-ARCH representation do not coincide (even in distribution).

Another very interesting time-varying autoregressive model is when (θ_t) itself is solution of an AR(1) model

$$\theta_t = F\theta_{t-1} + H\eta_t.$$

The model is called doubly-stochastic. It also exhibits heavy tailed phenomenon and its extremal behavior is really sensitive to the values of F and H. Notice that the previous example corresponds to $F = 0$, $H = \sqrt{\beta}$ and $\eta_t = \phi/\sqrt{\beta} + N_t$.

Those models exhibbit heavy tails because $\mathbb{E}[\theta_0] \approx 1$: the random multiplicative coefficient is fluctuating around 1 in the AR(1) representation. In many economics applications, it is relevant to consider such models as, when fitting an AR(1) with constant coefficients ϕ, the estimator of this coefficient is often close to 1. We then say we face the *unit root problem* because the values $|\phi| \geq 1$ are excluded from the classical inference to produce stable estimation. It is a well-known problem that has been treated in many ways; one can for instance consider Integrated ARMA models (ARIMA) that admits an unstable state-space representation associated to a stable Kalman's recursion or one can also use the cointegration analysis. Here we will develop a third approach based on Kalman's recursion.

6.3 State space models with random coefficients

The main idea is to consider the random coefficients (θ_t) as hidden states following a recursive equation. Let us consider the state-space model

$$\begin{cases} X_{t+1} = \mathbf{X}_t^T\theta_t + Z_{t+1} & \text{Space equation,} \\ \theta_{t+1} = \mathbf{F}\theta_t + \mathbf{V}_{t+1} & \text{State equation,} \end{cases}$$

where the coefficients $(\mathbf{X}_t)$ are random and $(\mathbf{V}_t)$ and (Z_t) are SWN($\mathbf{Q}$) and SWN(σ^2), respectively. The main assumption is that $(\mathbf{X}_t)$ is stationary ergodic sequences adapted to the filtration $\mathcal{F}_t = \sigma(\eta_t, Z_t, \eta_{t-1}, Z_{t-1}, \ldots)$. Note that there is a shift in the indices in the space equation: we first observe $\mathbf{X}_t$ in order to predict X_{t+1}.

Under the gaussian assumption, working recursively conditionally on $\mathcal{F}_t$ and using that for normal vectors orthogonality and independence is equivalent, one can extend the Kalman's recursion.

Theorem (Kalman (1960))**.** *In a state-space model with random coefficients, under the normal condition and if $\widehat{Y}_0$ and Ω_0 are well-chosen, one can compute recursively $\widehat{\theta}_n = \Pi_n(\theta_n)$ and the standardized risk $\Omega_n = \mathbb{E}[(\theta_n - \widehat{\theta}_n)(\theta_n - \widehat{\theta}_n)^\top]$ by the following recursion given the observation of $(X_n, \mathbf{X}_{n-1})$*

$$\widehat{\theta}_n = \mathbf{F}\widehat{\theta}_{n-1} + \frac{1}{r_n^L}\mathbf{F}\Omega_{n-1}\mathbf{X}_{n-1}^T(X_n - \mathbf{X}_{n-1}^T\widehat{\theta}_n)$$

$$\Omega_n = \mathbf{F}\Omega_{n-1}\mathbf{F}^\top + \mathbf{Q} - \frac{1}{\mathbf{X}_{n-1}^\top\Omega_{n-1}\mathbf{X}_{n-1} + \sigma^2}\mathbf{F}\Omega_{n-1}\mathbf{X}_{n-1}\mathbf{X}_{n-1}^\top\Omega_{n-1}\mathbf{F}^\top.$$

The main difference with the previous recursion is that as X_n is observed simultaneously with $\mathbf{X}_n$ one can dynamically estimate X_{n+1} using $\widehat{X}_{n+1} = \mathbf{X}_n^T\widehat{\theta}_n$ together with the reduced risk $R_{n+1}^L = \mathbf{X}_n^\top\Omega_n\mathbf{X}_n + 1$. Noticing that under the gaussian conditional assumption we have $\widehat{X}_{n+1} = \mathbb{E}[X_{n+1} \mid \mathcal{F}_n, \widehat{\theta}_0, \Omega_0]$ and $R_{n+1}^L = \mathrm{Var}\,(X_{n+1} \mid \mathcal{F}_{n+1}, \widehat{\theta}_0, \Omega_0)$ when the

arbitrary initial values for $\widehat{\theta}_0$, and Ω_0 are included in the filtration, one can compute the QLik contrast recursively as before.

Assume that the state-space model is parametrized over some hyperparameters $\lambda \in \mathbb{R}^d$. Let $\widehat{\theta}_0$ be some starting coefficient corresponding to the (unique) most likely fit on the observations, i.e. the usual OLS. The reduced QLik L_n^0 is then computed from the recursion:

- *Initialization:* $\widehat{\theta}_0$ and Ω_0
- *New observation* X_n*:*
  1. Compute the innovation $I_n(\lambda) = X_n - \widehat{X}_n(\lambda)$,
  3. Compute the next linear prediction $\widehat{X}_{n+1}(\lambda)$ and the associated risk $R_{n+1}^L(\lambda)$ thanks to the Kalman's recursion.

From this sequential algorithm, it is then simple to derive the reduced QMLE for state-space models with random coefficients

$$\widehat{\lambda}_n \in \arg\min_{\Theta} L_n(\lambda) = \arg\min_{\Theta} \sum_{t=1}^{n} \frac{{I_t}^2(\lambda)}{R_t^L(\lambda)} + \log(R_t^L(\lambda))$$

where $I_t(\lambda)$ and $R_t^L(\lambda)$ are defined recursively as above. Here

$$\sigma^2 = \frac{1}{n}\sum_{t=1}^{n} {I_t}^2(\lambda)\,,$$

is a good estimator of σ^2.

6.4 Dynamical models

The common choice $\mathbf{F}_t = I_k$ is made in this prospect as it does not require any calibration. It corresponds to the dynamical models used in Bayesian forecasting. The main step of the Kalman's recursion

$$\widehat{\theta}_n = \widehat{\theta}_{n-1} + \frac{1}{R_n^L}\Omega_{n-1}\mathbf{X}_{n-1}^T(X_n - \widehat{X}_n)$$

coincides with a stochastic Newton recursive method. More precisely, if one consider the problem of minimization of the quadratic loss

$$\theta \mapsto \ell_t(\theta) = (X_t - \theta^T\mathbf{X}_{t-1})^2$$

then one can use a stochastic gradient approach based where

$$\nabla\ell_t(\theta) = -2\mathbf{X}_{t-1}(X_t - \theta^T\mathbf{X}_{t-1}), \qquad \nabla^2\ell_t(\theta) = 2\mathbf{X}_{t-1}\mathbf{X}_{t-1}^\top.$$

Then the common Stochastic Newton algorithm updates as

$$\widehat{\theta}_{n+1} = \widehat{\theta}_n - 2\eta\Big(\sum_{t=1}^{n}\nabla^2\ell_t(\theta)\Big)^{-1}\nabla\ell_n(\widehat{\theta}_n)\,.$$

This algorithm converges may converge to the unique minimum of $\mathbb{E}[\ell_0]$ under strong convexity assumption at an optimal rate n^{-1} when the learning rate is well chosen. One can identify

$$2\eta\Big(\sum_{t=1}^{n}\nabla^2\ell_t(\theta)\Big)^{-1} = O\Big(\Big(\sum_{t=1}^{n}\mathbf{X}_t\mathbf{X}_t^\top\Big)^{-1}\Big) \qquad \text{and} \qquad \frac{1}{R_n^L}\Omega_{n-1}$$

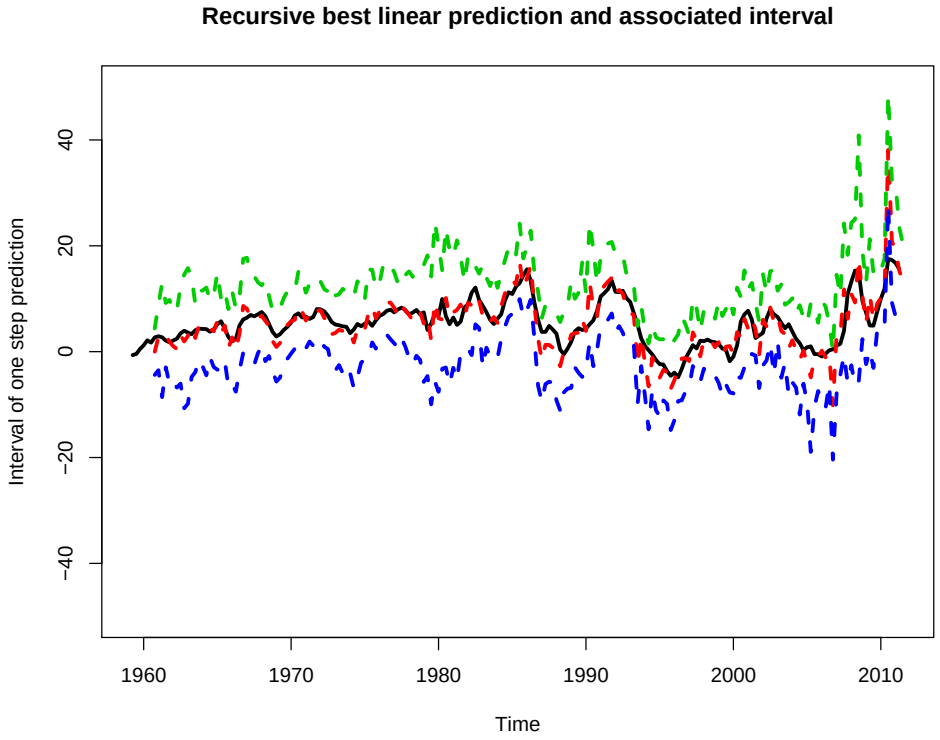

Figure 6.1: Quarterly Monetary Supply one-step forecasting by Kalman recursion on training sample from 1959 to 2014, data from `https://research.stlouisfed.org`

in the Kalman filter when $\mathbf{Q}$ is assumed to be null. The case $\mathbf{Q}$ identically null called the static case is rather restrictive in view of the state equation $\theta_{n+1} = \theta_n$. It corresponds to the special case where the coefficients θ_t are constant. But then the Kalman filter produces a gradient step that approximate the (optimal) Newton step without requiring second order matrices nor inversion of matrices.

Moreover, the Kalman filter offers much more flexibility than the common gradient based algorithms as it allows for non-identically null $\mathbf{Q}$. In such cases the Kalman filter does not converges (as there is no constant coefficient to converge to). It rather "tracks" the hidden random state θ_n and one talks about "tracking" algorithms.

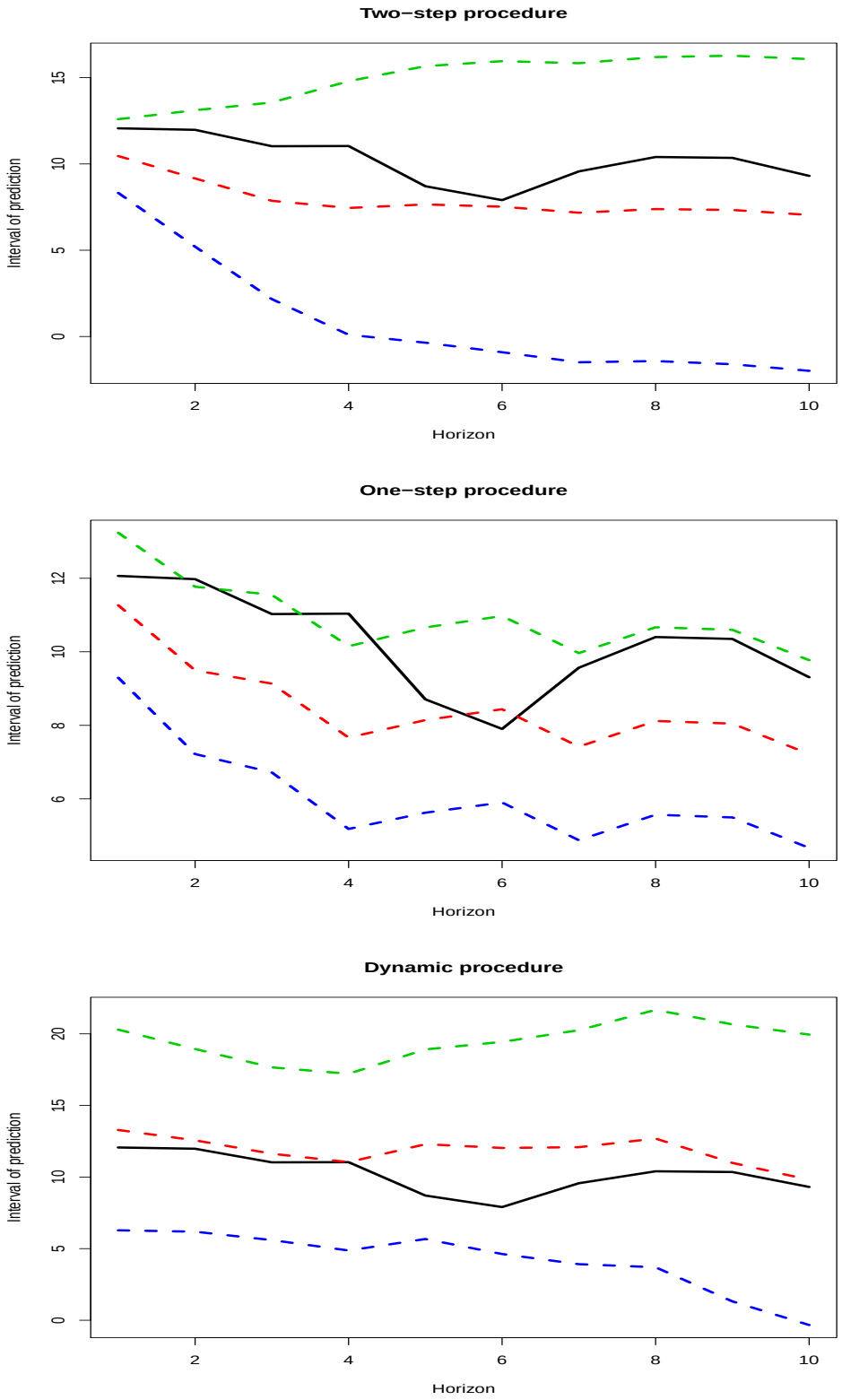

Figure 6.2: Quarterly Monetary Supply forecasting by two-step, onestep ARMA(5,6)-ARCH(1) and dynamic model on ten quarters from 2014, data from `https://research.stlouisfed.org`. The dynamic model uses four lagged data and past consumer consumption, treasury bills and surplus federal government as explanatory variables. The coefficients are static except the one of the consumer consumption.

Bibliography

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

George EP Box and Gwilym M Jenkins. *Time series analysis: forecasting and control*, volume 734. John Wiley & Sons, 2011.

Peter J Brockwell and Richard A Davis. *Time Series: Theory and Methods*. Springer Science & Business Media, 2013.

Christian Francq and Jean-Michel Zakoian. *GARCH models: structure, statistical inference and financial applications*. John Wiley & Sons, 2019.

Charles M Goldie et al. Implicit renewal theory and tails of solutions of random equations. *The Annals of Applied Probability*, 1(1):126–166, 1991.

Edward James Hannan. *Multiple time series*, volume 38. John Wiley & Sons, 1970.

Edward James Hannan and Manfred Deistler. *The statistical theory of linear systems*. SIAM, 2012.

R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960.

Claudia Klüppelberg, Serguei Pergamenchtchikov, et al. The tail of the stationary distribution of a random coefficient ar (q) model. *The Annals of Applied Probability*, 14(2): 971–1005, 2004.

J Pfanzagl. Asymptotic expansions related to minimum contrast estimators. *The Annals of Statistics*, pages 993–1026, 1973.