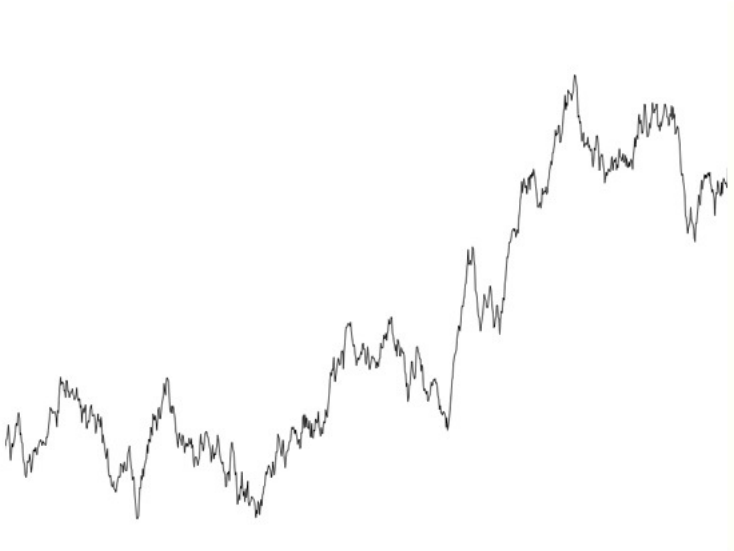
Statistics for Time Series Models

Olivier Wintenberger



2017-2018

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 (generally real-valued). (X_t) is a sequence of random variables on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Observing $(X_1, \dots, X_n)$ at times $t = 1, \dots, n$, the classical issues in statistics is to forecast the future at time n: $X_{n+1}, X_{n+2}, \dots$. In order to do so, one infers the dependence structure of the observed process and one uses it in order to construct a predictor.

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

Definition. (X_t) is strictly (or strongly) stationary if for all $k \in \mathbb{N}$, the joint distribution of $(X_t, \dots, X_{t+k})$ does not depend on t.

Hence, in order to forecast the future at time n, one can subsample $(X_1, \dots, X_n)$ in samples of length k and use the fact that $(X_{n-k}, \dots, X_{n+1})$ and $(X_1, \dots, X_{k+2})$, $(X_2, \dots, X_{k+3})$ are identically distributed... On these subsamples, the last value is observed so that one can assert the predictive power of the predictor from the k first values.

If the process is not reasonably likely to be stationary, one cannot rely on the observations to predict the future. In practice, one has to stationarize our observations first.

1.1 Data pre-processing

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 reasonably seen as stationary. We will not discuss potential pre-processing 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:

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

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 should reasonably be stationary. Of course the decomposition is not unique and it is a hard work to identify each components.

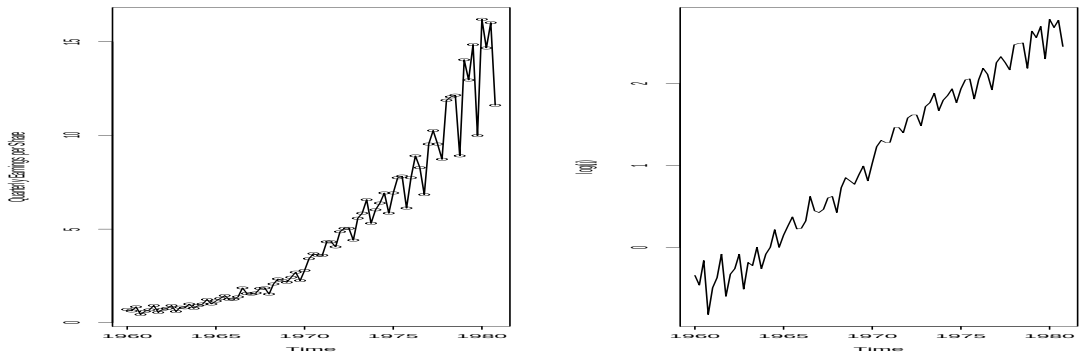


Figure 1.1: Econometrics data exhibiting an exponential (multiplicative) trend that turns into a linear (additive) trend after log-transform

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

Example. For economics data, it is reasonable to take into account an exponential trend from the inflation with; 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 and the inflation:

$$D_t = P_t e^{rt}, \qquad t \geq 0.$$

Due to the presence of an exponential trend, this data cannot be seen as stationary. By applying the log transform, on obtain

$$\log(D_t) = \log(P_t) + rt, \qquad t \geq 0.$$

The exponential trend is transformed in a linear one that we will treat hereafter. Figure shows quarterly earnings per share for the U.S. company Johnson & Johnson from 1960 to 1980 in Figure 1.1.

1.1.1 Differencing

Let us treat the trend part $f(t)$ in the decomposition (1.1), assuming that the seasonality part is null $S_t = 0$. 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 on an example.

Example. Let us regress the price of chicken in cents on the unit of time from 2001 to 2016 (notice on some 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 1.1.1.

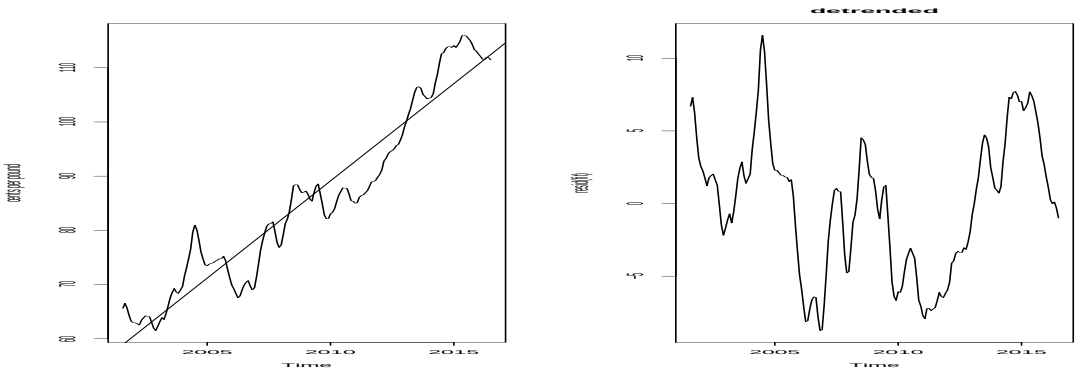


Figure 1.2: Estimation of the stationary component in presence of a linear trend thanks to linear regression on the time.

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

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

A SWN generates the natural filtration $\mathcal{F}_t = \sigma(Z_t, Z_{t-1}, \dots)$. The prediction at time n cannot use any information from the future Z_{n+1}. 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; more interesting problem than prediction are usually treated (estimation and tests).

Definition. Let $(\mathcal{F}_t)$ be a filtration. The process (X_t) is *non-anticipative* relatively to the SWN (Z_t) if $X_t \in \mathcal{F}_t = \sigma(Z_t, Z_{t-1}, \dots)$, $t \geq 1$. The process (X_t) is *invertible* if $X_t \in \mathcal{F}_t = \sigma(Z_t, X_{t-1}, \dots)$, $t \geq 1$.

Notice that an invertible process is non-anticipative. The invertibility is the most important notion in statistics. 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 pas observations.

Example (1.1.1, 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 $(\hat{a}_0(D_1, \dots, D_n), \hat{b}_0(D_1, \dots, D_n))$. Thus the residuals

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

is anticipative because they depend on the future D_s and thus Z_s for $s > t$. Such stationarization transformation does not respect the arrow of time. It is likely that $(\hat{a}_0(D_1, \dots, D_n) - \hat{b}_0(D_1, \dots, D_n)t)$ 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.

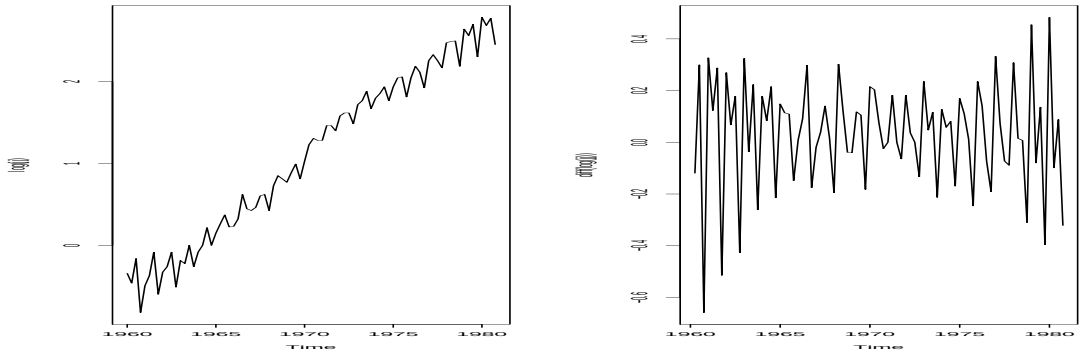


Figure 1.3: A linear (additive) trend is removed thanks to differenciation. The obtained time series has a mean behavior constant in time but exhibits some heteroscedastic behavior, i.e. a non constant variance.

Stationarization transformation that respects the arrow of time are based on the difference operator:

Definition. The lag (or backshift) operator L is defined as $LD_t = D_{t-1}$ for any data D_t, $t \geq 1$. The difference operator $\nabla = Id - L$ is defined so that $\nabla D_t = D_t - D_{t-1}$, $t \geq 1$.

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

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

If (X_t) is stationary, then $b + \nabla X_t$ is also and applying the difference operator stationarizes the data. Notice that the arrow of time is preserved; if (D_t) is non-anticipative with respect to SWN (Z_t) so is (∇D_t).

Example (1.1, continued)**.** On econometric data, the log transformed data $\log(D_t) = \log(P_t) + rt$ exhibit a linear trend. Applying the difference operator, we obtain $\nabla \log(D_t) = \log(P_t/P_{t-1}) + r$ which is reasonably stationary. Neglecting the influence of the interest rate, one calls the obtained process $X_t = \nabla \log(D_t)(*100)$ the log-ratios.

Example (1.1.1, continued)**.** Let us perform the difference operator on the chicken prices and compare it with the residuals of the linear regression. As residuals from a possibly overfitted linear regression, the residuals have a very smooth trajectory that seems simpler to predict than the difference ∇D_t, see Figure 1.1.1. However, it is not the case because one cannot rely on the residuals to directly predict D_{n+1} as

$$\hat{D}_{n+1} = \hat{a}_0(D_1, \ldots, D_n) + \hat{b}_0(D_1, \ldots, D_n)(n+1) + \hat{X}_{n+1},$$

as one should also take into account the errors of approximation

$$\hat{a}_0(D_1, \ldots, D_n) - \hat{a}_0(D_1, \ldots, D_{n+1}) \qquad \text{and} \qquad \hat{b}_0(D_1, \ldots, D_n) - \hat{b}_0(D_1, \ldots, D_{n+1}).$$

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

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

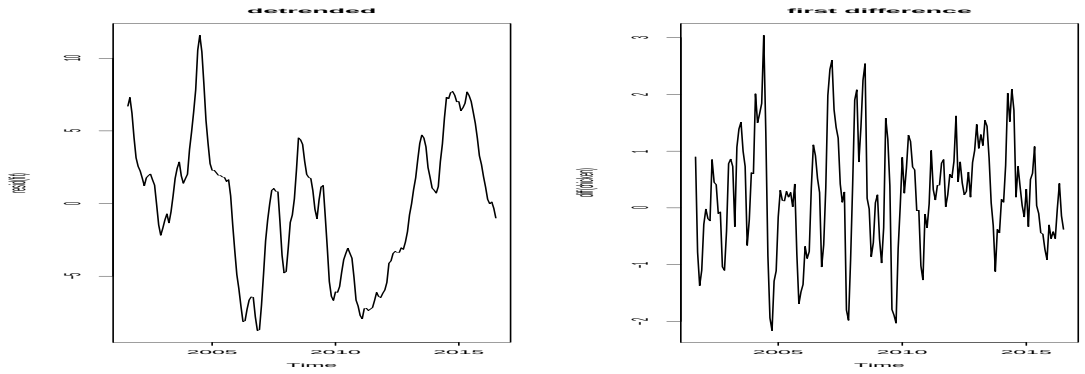


Figure 1.4: The difference operator are more variable then the residuals of the linear regression.

then, differentiating once, we obtain a linear trend

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

As if (X_t) is stationary, so it is (∇X_t), then we are back to the previous case and we stationarize ∇D_t by differentiating:

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

In this case $(2c_0 + \nabla^2 X_t)$ 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 arrow of the time. Moreover, it is simple to come back to the orignal data D_t by the inverse operator, called integration. For instance, denoting $X_t = \nabla^2 D_t$, assuming that it is stationary so that we can construct a predictor $\hat{X}_{n+1}$ then

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

It seems that there is no limit in the differencing process: the more you differentiate 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 $\nabla D_t = D_t - D_{t-1}$ is also stationary, so it is tempting to erroneously differentiate the observations. However, $\mathrm{Var}\,(\nabla D_t) = 2\sigma^2 > \mathrm{Var}\,(D_t)$ and the variance of the differentiate process is larger than the original one. More generally, the stationarization by successive differencing should stop when it increases the variance, i.e. when $\mathrm{Var}\,(\nabla X_t) > \mathrm{Var}\,(X_t)$. Then (X_t) is considered as the stationary version of the data.

1.1.2 Seasonal coefficients

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 $(S_j)_{1 \leq t \leq T}$ are easily estimated by the empirical mean

$$\hat{S}_j = \frac{T}{n} \sum_{1 \leq t = kT + j \leq n} D_t, \qquad 1 \leq j \leq T.$$

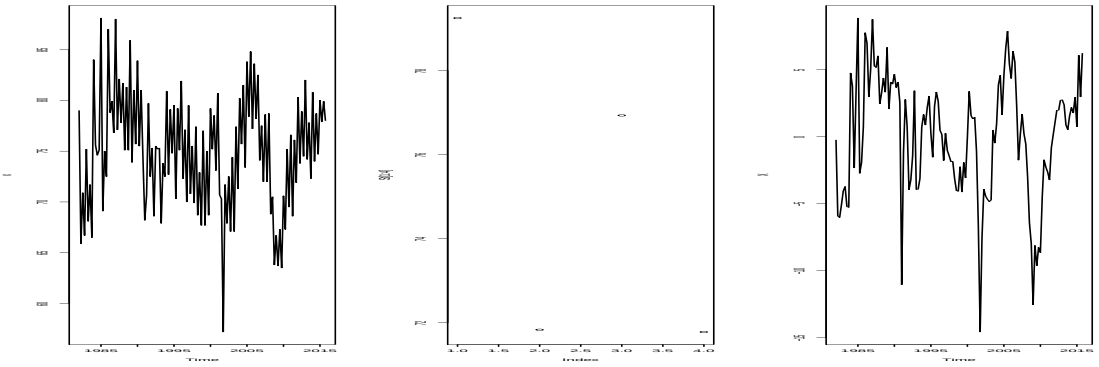


Figure 1.5: The orignal data (D_t), the 4 seasonal coefficients $(\hat{S}_j)_{1\leq j\leq 4}$ and the seasonally adjusted time series $(X_t = D_t - S_t)$. It may be a polynomial trend in (X_t). Differencing could be applied, upstream the seasonal adjustment that should be recalculate.

This transformation breaks the arrow of time. Thus, there is a risk of overfitting and it should be applied only if there is a strong suspicion of seasonality. For instance, for monthly data the period is very likely to be a year, i.e. $T = 12$. It could also be 2 or 3 years but choosing a too long period is dangerous, as it increases the risk of overfitting thanks to the seasonal coefficients.

Example. Consider the quarterly occupancy rate of Hawaiian hotels from 2002 to 2016. There is a strong suspicion of a seasonality of period $T = 4$. Thus, one can compute the 4 seasonal coefficients. 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.

Note that by considering seasonality with possibly period $T = 1$, the seasonally adjusted time series is centered $\mathbb{E}[X_t] = 0$.

1.2 Second order stationarity

We consider now that pre-processing has been applied and that (X_t) is reasonably stationary and centered. Let us first consider that it is reasonably stationary of the second order which implies some homoscedasticity (constant variance in time, not a lot of extreme values).

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

- the *autocovariance function*:

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

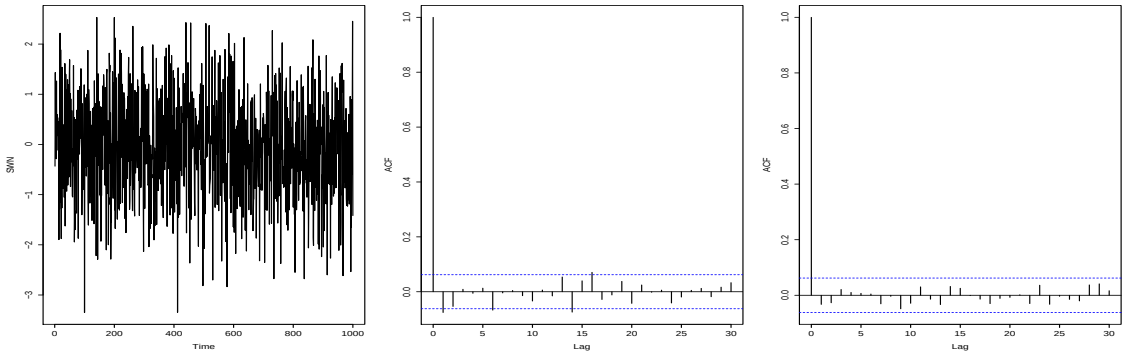


Figure 1.6: A trajectory and the corresponding ACF of a SWN and its squares

- the *autocorrelation function*:

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

- The *cross-covariance function*:

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

  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).

Remark.

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

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

Definition. 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 $(X_t) \in \mathrm{WN}(\sigma^2)$.

Exercise. Find an example of a white noise which is not i.i.d. stationary.

1.2.2 Linear time series

We have the following definition

Definition. 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 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:

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

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 $\mathbb{L}^2$ and a.s. to $X_t = \sum_{j\in\mathbb{Z}} \psi_j Z_{t-j}$ that exists. □

Reminder :
Let $x_t \geq 0$ for all $t \in \mathbb{Z}$ then the series $\sum_{t\in\mathbb{Z}} x_t$ is always well defined in $[0, +\infty]$.
For $x_t \in \mathbb{R}$, if $\sum_{t\in\mathbb{Z}} |x_t| < +\infty$, then $\sum_{t\in\mathbb{Z}} x_t$ is well defined and the order of summation is arbitrary. The series is said to converge absolutely. Thus, when $x_t \geq 0$ or $\sum_{t\in\mathbb{Z}} |x_t| < +\infty$, $\sum_{t\in\mathbb{Z}} x_t$ is well defined as the limit of the sequence $\left(\sum_{|t|\leq n} x_t\right)_{n\in\mathbb{N}}$.

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. The set of all measurable functions $f : \Omega \longrightarrow \mathbb{R}$ such that $\int |f|^2 d\mathbb{P} < +\infty$ is denoted by $\mathcal{L}^2(\mathbb{P})$.

Definition. The inner product associated to

$$||f|| = \sqrt{\int |f|^2 d\mathbb{P}}$$

is

$$\langle f_1, f_2 \rangle = \int f_1 f_2 d\mathbb{P}$$

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

- *Bilinearity:* $\langle \alpha f_1, \beta f_2 \rangle = \alpha\beta \langle f_1, f_2 \rangle$
- *Symmetric:* $\langle f_2, f_1 \rangle = \langle f_1, f_2 \rangle$
- *Non-negative:* $\langle f, f \rangle \geq 0$ *and* $\langle f, f \rangle = 0$ *iff* $f = 0$ *a.s.*
- $|| \cdot ||$ *is a seminorm:* $||f_1 + f_2|| \leq ||f_1|| + ||f_2||$ *and* $||\alpha f|| = |\alpha| \, ||f||$.

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

Proposition. $\big(\mathbb{L}^2(\mathbb{P}), || \cdot ||\big)$ *is a Hilbert space.*

Exercise.

- Cauchy-Schwarz: for any $f, g \in \mathbb{L}^2(\mathbb{P})$, $| < f, g > | \leq \|f\| \|g\|$,
- Show the triangular inequality $\|f + g\| \leq \|f\| + \|g\|$.

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

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

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

$$\|f - g\|^2 = \|f - P_L(f)\|^2 + \|P_L(f) - g\|^2.$$

The Projection theorem has nice probabilistic interpretations. For $\mathbb{P}$ being the distribution of the WN (Z_t), we identify the measurable functions $f \in \mathbb{L}^2(\mathbb{P})$ with the r.v. X such that $\mathbb{P}(X \in A) = \mathbb{P}(f^{-1}(A))$ for any $A \in \mathcal{A}$. Moreover $< X, Y >= \mathbb{E}[XY]$ and so $< X, Y >= \mathrm{Cov}(X, Y)$ 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 closed and

Definition. 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 $\mathbb{E}[Xh(Y)] = 0$ for any measurable and square integrable function h.

Definition. The projection $P_L(X)$ is called the conditional expectation of X on $\mathcal{A}_0$ and is denoted $P_L(X) = \mathbb{E}[X \mid A_0]$.

1.2.4 Best linear prediction

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

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

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

Definition. 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) $R^L_{n+1} = \mathbb{E}[(X_{n+1} - f(X_n, \dots, X_1))^2]$.

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

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

Thus (R^L_n) is a converging sequence with non-negative limit denoted R^L_∞. Moreover $\Pi_n(X_{n+1}) = \theta_1 X_n + \dots + \theta_n X_1$ 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. The system of n equations on the covariances constitute defining the coefficients of the best linear prediction is called the Yule-Walker system and it is equal to

$$(\rho_X(h))_{1 \leq h \leq n} = (\rho_X(i-j))_{1 \leq i,j \leq n} \theta.$$

The equations of prediction are equivalent to the Yule-Walker system

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

Thus one can write in a compact way with $\mathbb{X} = (X_1, \dots, X_n)'$

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

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 with diagonal 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 $R^L_\infty = 0$ one says that the second order stationary time series (X_t) is *deterministic*. For instance $r_t = X$ 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. Let A and B two random variables such that $\mathrm{Var}\,(A) = \mathrm{Var}\,(B) = \sigma^2$, $\mathbb{E}(A) = \mathbb{E}(B) = 0$ and $\mathrm{Cov}(A, B) = 0$. 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

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

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. 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_n^L. In general, the innovations are not stationary as R_n^L 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 $(X_1, \ldots, X_{n+1})$ as the linear span L_n of $(X_1, \ldots, X_n)$ 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 $(X_{-1}, \ldots, X_{-n})$ 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 $(X_{-1}, \ldots, X_{-n})$. 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

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

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),*
- $Cov(Z_t, r_s) = 0$ *for all* t, $s \in \mathbb{Z}$.
- (r_t) *is the deterministic component in the sense that*

$$\mathbb{E}[r_t \mid X_{t-1}, X_{t-2}, \ldots] = r_t, \qquad t \in Z,$$

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. 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 pre-processing 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 $I_{n+1} = X_{n+1} - \Pi_n(X_{n+1})$ converges in distribution to Z_0 that we identify with $I_\infty(X_0)$. 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. The empirical autocorrelation is defined as

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

Notice that by definition, we have the following properties

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

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

Theorem. *If (X_t) is a SWN then $\hat{\rho}_X(h)$ 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*

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

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

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

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

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

which implies that $\hat{\gamma}_X(0) \xrightarrow{\mathbb{P}.} \gamma_X(0)$. 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 $\hat{\gamma}_X(h)$ 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. 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 $j \geq q$.

Definition. 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 MA(∞) model

Example. 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 as MA(1). (X_t) is second order stationary because $\mathbb{E}(X_t) = 0$ and

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

In general, we have the following very useful property

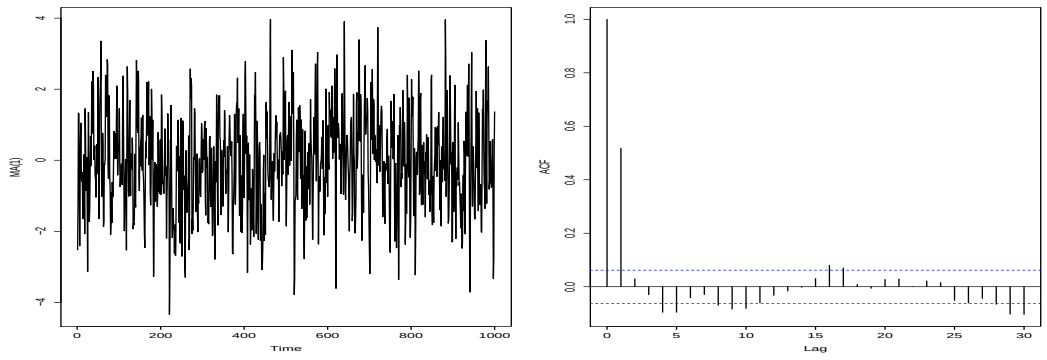


Figure 2.1: A trajectory and the corresponding ACF of the solution of an MA(1) model

Proposition. *If (X_t) is a MA(q) time series, we have $\gamma_X(h) = 0$ for all $h \geq q$.*

Remark.

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

Exercise. Show that (X_t) is stationary as soon as (Z_t) is.

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).

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

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

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

Example. Let (Z_t) be a SWN(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:

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

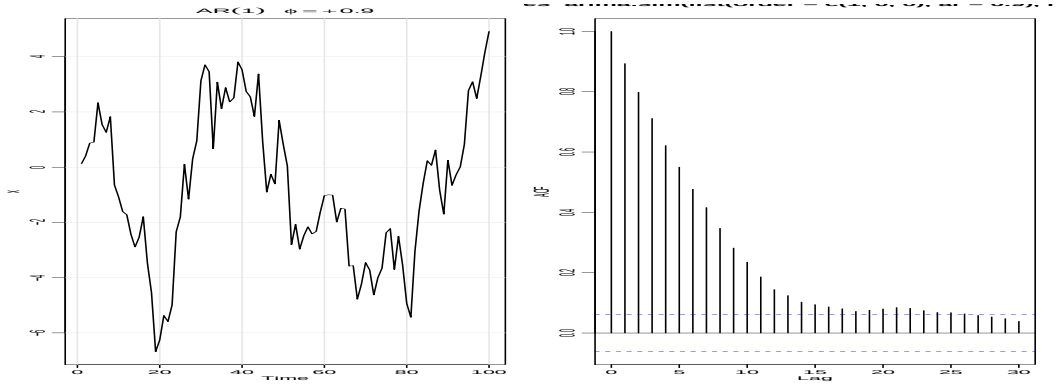


Figure 2.2: A trajectory and the corresponding ACF of the solution of an AR(1) model

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 shall prove that this representation is well defined for non strong WN (Z_t). Remark that $\gamma_X(0) = \sigma^2/(1-\phi^2)$ and that $\rho_X(h) = \phi^h$, $h \geq 0$.

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

The notion used for asserting the order of auto-regression is the partial autocorrelation that uses the innovations

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

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

where $\Pi_{h-1}(X_0)$ is the projection of X_0 on the linear span of $(X_1, \ldots, X_{h-1})$ (null for any causal time series).

By definition $\tilde{\rho}_X(1) = \rho_X(1)$. Notice that for any causal time series we $\Pi_{h-1}(X_0) = 0$ so that

$$\tilde{\rho}_X(h) = \rho_X(X_0, I_h), \qquad h \geq 1.$$

The partial autocorrelations are used to determine graphically the order of an AR(p) model. Indeed, we have

Proposition. *The PACF of an AR(p) time series satisfies $\tilde{\rho}_X(h) = 0$ for all $h > p$*

Proof. Indeed, for an AR(p) time series we have $\Pi_{h-1}(X_0) = 0$ and $\Pi_{h-1}(X_h) = \phi_1 X_{h-1} + \cdots + \phi_p X_{h-p}$ so that

$$\tilde{\rho}_X(h) = \rho_X(X_0, Z_h) = 0, \qquad h > p.$$

□

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

Remark. If $\phi = 1$, by iterating we get the random walk $X_t = X_0 + Z_1 + \cdots + Z_t$ and $\mathrm{Var}\,(X_t - X_0) = t\sigma^2 \underset{t\to+\infty}{\longrightarrow} +\infty$. Let's assume that (X_t) is weak-sense stationary. Then by Minkowsky inequality:

$$\sqrt{\mathrm{Var}\,(X_t - X_0)} \leq \sqrt{\mathrm{Var}\,(X_t)} + \sqrt{\mathrm{Var}\,(X_0)} = 2\sqrt{\mathrm{Var}\,(X_0)}$$

This is a contradiction. So the random walk is not weak-sense stationary. This course is restricted to the stationary case.

Exercise.

- Show that if $|\phi| > 1$, the equation $X_t = \phi X_{t-1} + Z_t$ admits a unique second order stationary solution of the form $X_t = -\sum_{j=1}^{+\infty} \phi^{-j} Z_{t+j}$. It is a linear time series that is not causal.
- Show that if $\phi = -1$, there is no stationary solution.

However, (X_t) is said to be (short memory) weakly dependent in both cases 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): $|\gamma_X(h)| \sim h^{-a}$, $a > 1/2$.

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 backward shift operator such that $L((X_t)) = (X_{t-1})$ and $L^k(X_t) = X_{t-k}$. One can now rewrite the ARMA model in a compact form

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

where (Z_t) is a $\mathrm{WN}(\sigma^2)$ and

$$\begin{aligned}\phi(x) &= 1 - \phi_1 x - \cdots - \phi_p x^p,\\ \gamma(x) &= 1 + \gamma_1 x + \cdots + \gamma_p x^p, \qquad x \in \mathbb{R}.\end{aligned}$$

We need to use complex analysis to solve the equation $\phi(L)X_t = \gamma(L)Z_t$ as $X_t = \phi^{-1}(L)\gamma(L)Z_t = \psi(L)Z_t$.

Definition. 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 $\sup_t \mathbb{E}|Z_t| \leq \sigma < \infty$ then $\psi(L)Z_t$ exists a.s. and in $\mathbb{L}^1$. Thus, 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.

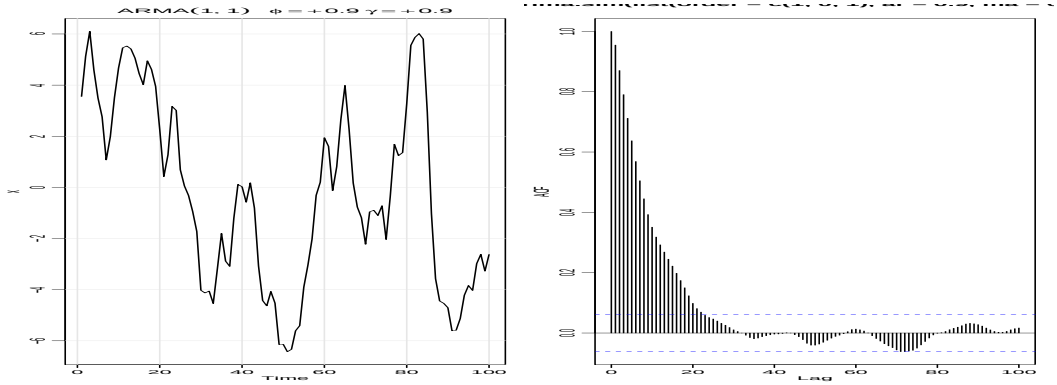


Figure 2.3: A trajectory and the corresponding ACF of the solution of an ARMA(1,1) model with $\phi 1 = \gamma_1 = 0.9$.

Proposition. *Assume that* $\sum |\psi_{1,j}| < \infty$ *and* $\sum |\psi_{2,j}| < \infty$.

1. *the series* $\psi_i(z) = \sum_{j\in\mathbb{Z}} \psi_{i,j} z^j$ *are well defined on* S,
2. $\psi_1(z)\psi_2(z) = \psi_2(z)\psi_1(z) = \sum_{k\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} \psi_{1,j}\psi_{2,k-j} z^k$ *is well defined on* S.

We are now ready to state

Theorem. *If* ϕ *do not have roots on* S *then the ARMA model admits a solution* $X_t = \phi^{-1}(L)\gamma(T)Z_t = \psi(T)Z_t$, $t \in \mathbb{Z}$ *and* (X_t) *is a causal linear time sereies.*

Recall the notion of causality, meaning here that the process (X_t) is a linear transformation of the past $(Z_t, Z_{t-1}, \ldots)$. Here, it is equivalent to assert that $\psi_j = 0$ for $j < 0$ and so that the Laurent series is holomorphic on $D = \{z \in \mathbb{C}, |z| \leq 1\}$ as its Taylor representation exists on S and so also for smaller $|z|$. As $\psi(z) = \phi^{-1}(z)\gamma(z)$, 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 if* ϕ *does not have roots inside* D,
2. *(linearly) invertible, i.e.* $\varphi(L)X_t = \sum_{j=0}^{\infty} \varphi_j X_{t-j} = Z_t$ *if* γ *does not have roots inside* D.

The second assertion holds for the same reason than the first one as $\varphi(z) = \gamma^{-1}(z)\phi(z)$. Moreover, from Cauchy's integral theorem on holomorphic functions defined on some extension of the unit disc, 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 $\gamma_X(h) = \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h} = O(\rho^j)$.

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

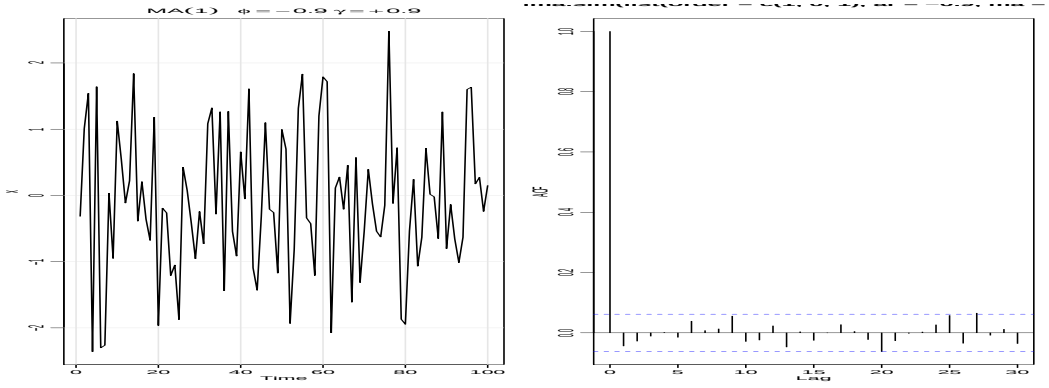


Figure 2.4: A trajectory and the corresponding ACF of the solution of an ARMA(1,1) model with $-\phi 1 = \gamma_1 = 0.9$

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 even more crucial than the ACF or the PACF do not yield any information on how to choose the orders p and q.

Example. 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. 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 $(Z_t(\theta))$, the noise of the model $(X_1(\theta), \dots, X_n(\theta))$, is gaussian WN$(\sigma^2)$. The Quasi Maximum Likelihood Estimator (QMLE) satisfies

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

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 only for calculating the criterion. It is not the Likelihood, i.e. we do not want to believe that the observations (X_t) satisfies the ARMA(p, q) model. 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 a deterministic function of the QMLE $\hat{\theta}_n$.

3.1 The QML Estimator

3.1.1 Gaussian distribution

Definition. 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 $\text{Var}\,(X) = 1$.

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

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

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.*

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 can deduce that for centered observations $X_1, \ldots, X_n$ 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 an ARMA(p,q) as fixed. One wants to compute the density of the model $(X_1(\theta), \ldots, X_n(\theta))$ that are not independent random variables. Thus, the density is not a priori a product. However, one always has

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

where $f_\theta(x_t \mid x_{t-1}, \ldots, x_1)$ is the density of the distribution of $X_t(\theta)$ given $X_{t-1}(\theta), \ldots, X_1(\theta)$. 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 $X_{t-1}(\theta), \ldots, X_1(\theta)$ is*

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

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 $(X_{t+1}(\theta), \ldots, X_{t+h}(\theta))$ for any $t \in \mathbb{Z}$ and $h \geq 1$ are gaussian. one says that $(X_t(\theta))$ is a guassian process. Thus, the conditional distribution of $X_t(\theta)$ given $X_{t-1}(\theta), \ldots, X_1(\theta)$ 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 $(X_{t+1}(\theta), \dots, X_{t+h}(\theta))$. Thus, it is independent and the conditional variance

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

□

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

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

In the sequel, we will denote for short the innovation of the ARMA model on the observatiosn 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 litterature on such class of estimators, called M-estimator.

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

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

The (random) functions ℓ_t are called the contrast. The set of parameters $'\Theta$ has to be chosen carefully. One convenient (and safe) way is to choose it as a compact set so that continuity of the contrast yields the existence of the M-estimator.

Avoiding the difficult problem of calculating efficiently $(\Pi_t(\theta))$ and $(R_t^L(\theta))$ (i.e. assuming they are known), we obtain the asymptotic behavior

$$\begin{aligned}\frac{1}{n} L_n(\theta) &= \frac{1}{n} \sum_{t=1}^{n} \log(f_\theta(X_t \mid X_{t-1}, \dots, X_1)) \\ &\approx \frac{1}{n} \sum_{t=1}^{n} \log(f_\theta(X_t \mid X_{t-1}, X_{t-2}, \dots)) \\ &\xrightarrow{a.s.} \mathbb{E}[\log(f_\theta(X_t \mid X_{t-1}, X_{t-2}, \dots))].\end{aligned}$$

Here some comments should be provided about these approximations.

The first approximation is valid as a Cesaro mean if

$$\log(f_\theta(X_t \mid X_{t-1}, \dots, X_1)) - \log(f_\theta(X_t \mid X_{t-1}, X_{t-2}, \dots)) \to 0, \qquad t \to \infty.$$

As

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

the first term will converge by continuity as $(R_t^L(\theta))$ is converging. The limit is

$$\log(R_t^L(\theta)) \to_{t\to\infty} \log(R_\infty^L(\theta)).$$

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

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

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

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

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)

The second approximation is made thanks to 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 T denotes the backward shift operator $T((X_t)) = (X_{t-1})$.

Definition. A set C of $\mathbb{R}^{\mathbb{Z}}$ is invariant iff $T^{-1}(C) = C$ 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 a strictly stationary and ergodic time series and f is a measurable function such that $\mathbb{E}[|f((X_t))|] < \infty$ then:*

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

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 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 that basic result, it is possible to construct other examples

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. Sometimes, it is difficult to have an explicit representation so we will use the following 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 there exists measurable function f such that*

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

and (X_t) is stationary ergodic

The ergodicity implies a more general notion of stability than the stability for averaging provided by the SLLN. Here we will average a function of the complete past (X_t) as required in the applications:

Theorem (Kingman)**.** *Assume that there exists measurable functions g_n, $n \geq 1$ that are subadditive, i.e.*

$$g_{n+m}((X_t)) \leq g_n((X_t)) + g_m((X_{t+n})), \qquad n, m \geq 1$$

then if (X_t) is stationary and ergodic and g_1^+ is integrable we have

$$\frac{g_n((X_t))}{n} \to \inf_{k\geq 1} \frac{\mathbb{E}[g_k((X_t))]}{k} \geq -\infty, \qquad a.s.$$

Notice that $g_n((X_t)) = \sum_{t=1}^n f(X_t)$ are subbadditive functions such that $k^{-1}\mathbb{E}[g_k((X_t))] = \mathbb{E}[f((X_t))]$ for all k by linearity. In general, by subadditivity, we always have $k^{-1}\mathbb{E}[g_k((X_t))] \leq \mathbb{E}[g_1((X_t))]$. An M-estimator is minimizing the cumulative losses also called the contrast. Combining the Kingman 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)**.** *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 $\hat{\theta}_n \to \theta_0$ a.s., i.e. the M-estimator of θ_0 is strongly consistent.

3.2 Strong 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 risk with the function

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

Note that the corresponding contrast is

$$\ell_t(\theta) = \log(R_\infty^L(\theta)) + \frac{(X_t - \Pi_\infty(\theta)(X_t))^2}{R_\infty^L(\theta)} + cst$$

which is an approximation of the log-likelihood of the distribution $\mathcal{N}(\Pi_t(\theta)(X_t), R_t^L(\theta))$ when θ corresponds to a causal and invertible ARMA(p,q) model. It is stationary, ergodic and admits second order moments if (X_t) does. Applying Pfanzagl Theorem, we obtain

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

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

then the QMLE $\hat{\theta}_n \to \theta_0$ a.s.

The last assumption 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 non anticipative and invertible ARMA(p,q) models with no common roots. We have the following *strong consistency* result

Theorem. *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 $\hat{\theta}_n \to \theta_0$ 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.

The rest of the proof is an application of Pfazagl theorem as above. The ergodicity and stationarity is ensured because fo 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

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

obtained using $X_0 = \Pi_\infty(\theta_0)(X_0) + Z_0$ and orthogonality. From strong convexity of the quadratic risk, $R_\infty^L(\theta)$ is minimized in θ_0 only. Because $x \mapsto \log(x) + x^{-1}$ is minimized in a unique point $1 = R_\infty^L(\theta_0)/\sigma^2$, 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. Thus $R_t^L(\theta)$ is not accessible. However, one can check that it is equal to $\sigma^2 r_t^L(\theta)$ where now $r_t^L(\theta)$ corresponds to the risk of linear prediction assuming that (Z_t) is a gaussian (standardized) WN(1). As the minimizer is derived from the nullity of the derivatives of the contrast, the M-estimator defined assuming that (Z_t) is a gaussian (standardized) WN(1) is still the QMLE $\hat{\theta}_n$. From Pfanzagl theorem we have also the convergence of the minimum of the QLik function, we obtain

$$\sum_{t=1}^{n} \frac{(X_0 - \Pi_t(\hat{\theta}_n)(X_0))^2}{\sigma^2 r_t^L(\hat{\theta}_n)} \xrightarrow{a.s.} 1.$$

We obtain

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 a strongly consistent estimator of the variance σ^2*

$$\hat{\sigma}_n^2 := \frac{1}{n} \sum_{t=1}^{n} \frac{(X_0 - \Pi_t(\hat{\theta}_n)(X_0))^2}{r_t^L(\hat{\theta}_n)} \xrightarrow{a.s.} \sigma^2, \qquad n \to \infty,$$

where $r_t^L(\hat{\theta}_n)$ is the risk of linear prediction under standardization of the WN.

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.

One can also show that $r_t^L(\hat{\theta}_n) \to r_\infty^L(\theta_0) = 1$ a.s. for any $\theta_0 \in \mathcal{C}$ because of the standardization. Substituting σ^2 by its estimator in the QLik loss, we obtain

$$\frac{1}{n} \sum_{t=1}^{n} \left(\log(\hat{\sigma}_n^2 r_t^L(\hat{\theta}_n)) + \frac{(X_t - \Pi(\hat{\theta}_n)(X_t))^2}{\hat{\sigma}_n^2 r_t^L(\hat{\theta}_n)} \right) \approx \log(\hat{\sigma}_n^2) + 1,$$

from the definition of $\hat{\sigma}_n^2$ and as

$$\frac{1}{n} \sum_{t=1}^{n} \log(r_t^L(\hat{\theta}_n)) \xrightarrow{a.s.} 0.$$

Notice that we cannot optimize the likelihood simultaneously on θ and σ^2 as

$$X_t = \sum_{j \geq 0} \psi_j Z_{t-j} = \sum_{j \geq 0} (c\psi_j)(c^{-1} Z_{t-j})$$

thus θ_0 and σ^2 are not uniquely determined.

3.2.3 Misspecification

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 build on the contrast is not the correct one. They are very important as the aim is to obtain results that are satisfied even if the normal assumption 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 if one can decompose the second order term. Actually, one can always decompose the second order stationary process (X_t) as

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

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

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

By the previous discussion, one can always consider first $R_\infty^L > 0$ as known and then estimating it by standardizing the WN. Thus, we are left in the unicity of the minimizer of the function

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

Developping 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 now 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 coefficents $(\varphi_j(\theta))$ are unique but not the corresponding value of the parameters $\theta_0 \in \Theta_0$. Indeed, one cannot avoid the parameters $\theta_0 \in \partial\mathcal{C}$, in partiular the ones that correpond 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 $\varphi(\theta_0)$ that uniquely determine the best linear prediction over $\varphi(\bar{C})$.*

Example. Fitting an ARMA(1,1) one 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 is also no longer consistently estimated as the minimum of the risk $\mathbb{E}[\ell_0]$ is no longer one, 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

3.3.1 Kullback-Leibler divergence

One can identify the risk $\mathbb{E}[\ell_0]$ associated with the QLik loss with an important notion from information theory that is a kind of distance between probability measures.

Definition. 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.*

Assume that $\sigma^2 = R^L_\infty > 0$ is unknown so that in the following $r^L_\infty(\theta)$ is the standardized linear prediction risk. One can identify, up to additive constants, the risk

$$\mathbb{E}[\ell_0(\theta)] = \log(\sigma^2 r^L_\infty(\theta)) + \frac{\mathbb{E}\left[(X_0 - \Pi_\infty(\theta)(X_0))^2\right]}{\sigma^2 r^L_\infty(\theta)}$$

with twice the expectation of the KL divergence of

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

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 $\mathbb{E}[\ell_0]$ but also of the conditional risk

$$\mathbb{E}[\ell_0(\theta) \mid X_{-1}, X_{-2}, \ldots] = 2\mathcal{K}(P_{X_0|X_{-1},X_{-2}}, \mathcal{N}(\Pi_\infty(\theta)(X_0), R^L_\infty(\theta))).$$

3.3.2 Asymptotic normality of the MLE

Let us turn to the ML estimator

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

where $\ell_t = -2\log(f_\theta(X_t \mid X_{t-1}, X_{t-2}, \ldots))$ consitutes a stationary sequence of contrast such that θ_0 is the unique minimizer of $\mathbb{E}[\ell_0]$ 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 also satisfied so that $\hat{\theta}_n$ is strongly consistent. The asymptotic normality of the QMLE follows in most of the cases under extra assumptions. If L_n is sufficiently regular (2-times continuously differentiable) then a Taylor expansion gives

$$\partial_\theta L_n(\hat{\theta}_n) = \partial_\theta L_n(\theta_0) + \partial_\theta^2 L_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0) \tag{3.2}$$

with $\tilde{\theta}_n \in [\theta_0, \hat{\theta}_n]$. Notice that as $\hat{\theta}_n$ is strongly convergent, then $[\theta_0, \hat{\theta}_n] \to \{\theta_0\}$ a.s. Moreover, if $\hat{\theta}_n \in \overset{\circ}{\Theta}$ the interior of the compact set then $\partial_\theta L_n(\hat{\theta}_n) = 0$ as the QMLE is the minimizer of the QLik contrast by assumption. So we have to study the properties of the two first derivative of the contrast L_n. Let us first show that the two first derivatives of L_n have nice properties at θ_0:

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

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

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

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 predictive power $\mathbb{E}[-\log(f_\theta(X_t \mid X_{t-1}, X_{t-2}, \ldots)) \mid X_{t-1}, X_{t-2}, \ldots]$ then the score vector is centered $\mathbb{E}[S_0 \mid \mathcal{X}_{-1}, X_{-2}, \ldots] = 0$ 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 $\mathcal{I}(\theta_0) = Var(S_0)$ and its inverse is the smallest possible variance of unbiased estimator, called the Cramer-Rao bound.*

Proof. As θ_0 is the minimizer of the predictive power 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 $f_{\theta_0} = f_{\theta_0}(X_0 \mid X_{-1}, X_{-2}, \ldots)$. Then we have

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

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

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

The proof of the Cramer-Rao bound is classical. □

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 predictive power* $\mathbb{E}[-\log(f_\theta(X_t \mid X_{t-1}, X_{t-2}, \ldots)) \mid X_{t-1}, X_{t-2}, \ldots]$*, if the contrast* $\ell_t = -2\log(f_\theta(X_t \mid X_{t-1}, X_{t-2}, \ldots))$ *is twice continuously differentiable and integrable, then the MLE is* asymptotically normal

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

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 difference of martingale. The CLT extends to such square integrable difference of martingales and we obtain

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

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

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

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

$$\begin{aligned} & 0 = \partial_\theta L_n(\theta_0) + \partial_\theta^2 L_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0) \\ \Leftrightarrow \quad & -\partial_\theta L_n(\theta_0) = \partial_\theta^2 L_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0) \\ \Leftrightarrow \quad & -\frac{1}{\sqrt{n}}\partial_\theta L_n(\theta_0) = \frac{1}{n}\partial_\theta^2 L_n(\tilde{\theta}_n)\sqrt{n}(\hat{\theta}_n - \theta_0). \end{aligned}$$

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 $2I(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0)$ so that the desired result is obtained. □

3.3.3 Asymptotic normality of the QMLE

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)**.** *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*

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

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

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

Proof. One first check the differentiability and integrability conditions on the QLik contrast

$$\ell_t(\theta) = \log(\sigma^2 r_\infty^L(\theta)) + \frac{(X_t - \Pi_\infty(\theta)(X_t))^2}{\sigma^2 r_\infty^L(\theta)}.$$

The score is defined as

$$S_t = -\frac{\nabla_\theta r_\infty^L(\theta_0)}{2r_\infty^L(\theta_0)} + \frac{(X_t - \Pi_\infty(\theta_0)(X_t))}{\sigma^2 r_\infty^L(\theta_0)}\nabla_\theta \Pi_\infty(\theta_0)(X_t) + \frac{(X_t - \Pi_\infty(\theta_0)(X_t))^2}{2\sigma^2 r_\infty^L(\theta_0)^2}\nabla_\theta r_\infty^L(\theta_0).$$

From the identities $r^L_\infty(\theta_0) = 1$, $\nabla_\theta r^L_\infty(\theta_0) = 0$ (because θ_0 minimizes the function r^L_∞) and $X_t - \Pi_\infty(\theta_0)(X_t) = Z_t$ we obtain the expression

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

One checks easily that $\mathbb{E}[S_0 \mid X_{t-1}, X_{t-2}, \ldots] = 0$ as expected. Its variance is

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

Similarly, one computes the Fisher information

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

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

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

Similarly, we have

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

so that

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

The desired result follows. □

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

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

Notice also that we have the identity $I(\theta_0) = \mathrm{Var}\,(S_0)$ and that the QMLE is efficient as soon as (Z_t) is gaussian WN. Notice that the asymptotic variance of $\hat\theta_n$ does not depend on σ^2. It complements the fact that θ and σ^2 can be estimated separately in ARMA models. Finally notice that the asymptotic variance can be estimated by computing the covariances of (AR_t) and (MA_t) driven by the QMLE $\hat\theta_n$ (actually one can compute explicitly $\mathrm{Var}\,(AR_p, \ldots, AR_1, MA_q, \ldots, MA_1)^{-1}$ in term of the coefficients θ of the polynomial of (AR_t) and (MA_t) or one can use numerical approximations).

3.3.4 Asymptotic properties of the predictions

Notice that when σ^2 is unknown, we have the QLik loss

$$\frac{1}{n}\sum_{t=1}^n\left(\log(\hat\sigma^2_n r^L_t(\hat\theta_n)) + \frac{(X_t - \Pi_{t-1}(\hat\theta_n)(X_t))^2}{\hat\sigma^2_n r^L_t(\hat\theta_n)}\right) \approx \log\left(\frac{1}{n}\sum_{t=1}^n (X_t - \Pi_\infty(\hat\theta_n)(X_t))^2\right) + 1,$$

as, under $\sum_{h\geq 0}|\gamma_X(h)| < \infty$, the error of approximation of $\Pi_\infty(\hat\theta_n)(X_t)$ by $\Pi_\infty(\hat\theta_n)(X_t)$ is exponentially decreasing with t in $\mathbb{L}^2$. Thus, the QMLE is equivalent to an ordinary least square estimator over an infinite number of explanatory variables (the whole past) under a constraint on the shape of the coefficients (the parameters are represented by an

ARMA(p,q) equation). However, in misspecified cases, the uniqueness of θ_0 is not ensured and the asymptotic normality result is not possible in full generality.

However, as $(\Pi(\hat{\theta}_n)(X_t))$ is the unique minimizer, the derivatives in φ at this point are null and

$$2\sum_{t=1}^{n}(X_t - \Pi_\infty(\hat{\theta}_n)(X_t))\Pi_\infty(\theta)(X_t) \approx 0, \qquad \theta \in \mathcal{C}.$$

By the remarkable identity, we then have

$$\sum_{t=1}^{n}(X_t - \Pi_\infty(\hat{\theta}_n)(X_t))^2 \approx \sum_{t=1}^{n}(X_t - \Pi_\infty(\theta_0)(X_t))^2 - \sum_{t=1}^{n}(\Pi_\infty(\theta_0)(X_t) - \Pi_\infty(\hat{\theta}_n)(X_t))^2.$$

If the uniqueness of θ_0 is likely (i.e. the QMLE converges in $\mathcal{C}$), the asymptotic normality follows and we can apply the Delta-method on the last term

$$\begin{aligned}
\sum_{t=1}^{n}(\Pi_\infty(\theta_0)(X_t) - \Pi_\infty(\hat{\theta}_n)(X_t))^2 &\approx \sum_{t=1}^{n}(\nabla_\theta\Pi_\infty(\theta_0)(X_t)^\top(\hat{\theta}_n - \theta_0))^2 \\
&\approx \sum_{t=1}^{n}\nabla_\theta\Pi_\infty(\theta_0)^\top(X_t)(\hat{\theta}_n - \theta_0)(\hat{\theta}_n - \theta_0)^\top\nabla_\theta\Pi_\infty(\theta_0)(X_t) \\
&\approx \sum_{t=1}^{n}\mathrm{Tr}(\nabla_\theta\Pi_\infty(\theta_0)(X_t)\nabla_\theta\Pi_\infty(\theta_0)(X_t)^\top(\hat{\theta}_n - \theta_0)(\hat{\theta}_n - \theta_0)^\top) \\
&\approx \mathrm{Tr}\Big(\frac{\sigma^2}{n}\sum_{t=1}^{n}\nabla_\theta\Pi_\infty(\theta_0)(X_t)\nabla_\theta\Pi_\infty(\theta_0)(X_t)^\top \\
&\qquad \mathbb{E}[\nabla_\theta\Pi_\infty(\theta_0)(X_t)\nabla_\theta\Pi_\infty(\theta_0)(X_t)^\top]^{-1}NN^\top\Big) \\
&\approx \sigma^2\mathrm{Tr}(\mathrm{Var}\,(N)) = \sigma^2\mathrm{Tr}(I_{p+q}) = \sigma^2(p+q)
\end{aligned}$$

where $N \in \mathbb{R}^{p+q}$ is a standardized and centered gaussian vector with $\mathrm{Var}\,(N) = I_{p+q}$ the identity matrix. We obtain that

$$\sum_{t=1}^{n}(X_t - \Pi_\infty(\hat{\theta}_n)(X_t))^2 \approx \sum_{t=1}^{n}(X_t - \Pi_\infty(\theta_0)(X_t))^2 - \sigma^2(p+q).$$

This result, depending only on the predictions, may be extended in misspecified cases. The quality of the prediction at $\hat{\theta}_n$ estimated on the sample $(X_1, \dots X_n)$ is strictly better than the best possible prediction (using θ_0). This statement is not contradictory because using twice in $(X_t - \Pi(\hat{\theta}_n(p,q))(X_t))^2$, once for calculating $\hat{\theta}_n$ and another time for estimating the function $\mathbb{E}[\ell_0]$, one under estimates the risk of prediction. It is because $\hat{\theta}_n$ uses the future at to predict X_t with $\Pi(\hat{\theta}_n(p,q))(X_t)$.

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 $\mathcal{N}(\Pi_\infty(\hat{\theta}_n)(X_0), \hat{\sigma}_n^2)$. Let us define

Definition. The *predictive power* of the model ARMA(p,q) fitted by the QMLE is

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

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 $\hat{\theta}_n(p,q)$ the QMLE for the ARMA(p,q) model and

$$\hat{\sigma}_n^2(p,q) = \frac{1}{n}\sum_{t=1}^{n} \frac{(X_t - \Pi_t(\hat{\theta}_n(p,q))(X_t))^2}{r_t^L(\hat{\theta}_n(p,q))}.$$

Akaike idea is to approximate (-2 times) the predictive power by penalizing the quantity

$$\frac{1}{n}L_n(\hat{\theta}_n) = \frac{1}{n}\sum_{t=1}^{n}\left(\log(\hat{\sigma}_n^2(p,q)r_t^L(\hat{\theta}_n(p,q))) + \frac{(X_t - \Pi(\hat{\theta}_n)(X_t))^2}{\hat{\sigma}_n^2(p,q)r_t^L(\hat{\theta}_n(p,q))}\right).$$

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 $(X_t - \Pi(\hat{\theta}_n(p,q))(X_t))^2$, once for calculating $\hat{\theta}_n$ and another time for estimating the function $\mathbb{E}[\ell_0]$. More precisely, we have

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

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

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

The procedure is then to select the order $(\hat{p}_n, \hat{q}_n)$ 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

- 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. *The AIC and AICc 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. Consider the approximation of the QLik $\frac{1}{n}L_n(\hat{\theta}_n(p,q))$ as

$$\log(\hat{\sigma}_n^2(p,q)) + \frac{\frac{1}{n}\sum_{t=1}^n (X_t - \Pi_t(\hat{\theta}_n(p,q))(X_t))^2}{\hat{\sigma}_n^2(p,q)}.$$

We will show that, up to an additive constant, it is an unbiased estimator of the mean of the predictive power

$$\log(\hat{\sigma}_n^2(p,q)) + \frac{\mathbb{E}[(X_0 - \Pi_\infty(\hat{\theta}_n(p,q))(X_0))^2 \mid \hat{\theta}_n(p,q)]}{\hat{\sigma}_n^2(p,q)}.$$

From the discussion Section 3.3.4 we have

$$\begin{aligned}\sum_{t=1}^n (X_t - \Pi_t(\hat{\theta}_n(p,q))(X_t))^2 &\approx \sum_{t=1}^n (X_t - \Pi_\infty(\hat{\theta}_n(p,q))(X_t))^2 \\ &\approx \sum_{t=1}^n (X_t - \Pi_\infty(\theta_0(p,q))(X_t))^2 - \sigma^2(p+q).\end{aligned}$$

On the opposite, we have the decomposition of the predictive power term

$$\begin{aligned}&\mathbb{E}[(X_0 - \Pi_\infty(\hat{\theta}_n(p,q))(X_0))^2 \mid \hat{\theta}_n(p,q)] \\ &\quad\approx \mathbb{E}[(\Pi_\infty(\theta_0(p,q))(X_0) - \Pi_\infty(\hat{\theta}_n(p,q))(X_0))^2 \mid \hat{\theta}_n(p,q)] \\ &\qquad + \mathbb{E}[(X_0 - \Pi_\infty(\theta_0(p,q))(X_0))^2]\end{aligned}$$

where the last identity is from the fact that θ_0 is the unique minimizer of

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

and thus the derivative w.r.t. φ of the square risk is approximatively null (similar discussion than in Section 3.3.4 on the square risk and not the square loss). Thanks to the asymptotic normality of the QMLE, we evaluate

$$\begin{aligned}&\mathbb{E}[(\Pi_\infty(\theta_0)(X_0) - \Pi_\infty(\hat{\theta}_n(p,q))(X_0))^2 \mid \hat{\theta}_n(p,q)] \\ &\qquad\approx \frac{1}{n}\sum_{t=1}^n (X_t - \Pi_\infty(\theta_0(p,q))(X_t))^2 \\ &\qquad\approx \frac{\sigma^2(p+q)}{n}.\end{aligned}$$

Assuming that $\hat{\sigma}_n^2 \approx \sigma^2$ (this point is true only asymptotically, a refinement taking into account the expectation of $\mathbb{E}[\hat{\sigma}_n^{-2}]$ yields AICc), we obtain the desired result

$$\begin{aligned}\frac{1}{n}L_n(\hat{\theta}_n(p,q)) + \frac{2(p+q)}{n} &\approx \log(\hat{\sigma}_n^2(p,q)) + \frac{\mathbb{E}[(X_0 - \Pi_\infty(\theta_0(p,q))(X_0))^2]}{\hat{\sigma}_n^2(p,q)} + \frac{p+q}{n} \\ &\approx \log(\hat{\sigma}_n^2(p,q)) + \frac{\mathbb{E}[(X_0 - \Pi_\infty(\hat{\theta}_n(p,q))(X_0))^2 \mid \hat{\theta}_n(p,q)]}{\hat{\sigma}_n^2(p,q)}.\end{aligned}$$

□

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 $\hat{X}_{n+1} = \Pi_n(\hat{\theta}_n)(X_{n+1})$. Notice that by construction it is an estimator of $\Pi_n(\theta_0)(X_{n+1})$ and we have from previous discussion

$$\mathbb{E}[(X_{n+1} - \hat{X}_{n+1})^2] \lesssim R_\infty^L + \mathbb{E}[(\Pi_\infty(X_{n+1}) - \Pi_n(\theta_0)(X_{n+1}))^2] + \frac{\sigma^2(p+q)}{n}.$$

Such inequality is called an *oracle inequality*. The best prediction within the model is $\Pi_n(\theta_0)(X_{n+1})$ and is called the oracle. The risk of prediction of the oracle is approximatively

$$R_\infty^L + \mathbb{E}[(\Pi_\infty(X_{n+1}) - \Pi_n(\theta_0)(X_{n+1}))^2]$$

the sum of the best risk of prediction and the square of the bias of the model ARMA(p, q).

An interval of prediction is often more useful than a point prediction. The QMLE produced a natural interval of confidence α such as

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

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 $\beta = \alpha/2$.

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

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

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 $\hat{\sigma}_n(h)$ is the associated risk

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

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*;

Definition. The length of the interval $|\hat{I}_\alpha(X_{n+1}))|$ 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).

In the conditional gaussian case all these quantities are proportional to the conditional variance $\mathrm{Var}\,(X_{n+1} \mid X_n, \ldots, X_1)$ also called the *volatility*. It is the main indicator to assess risks in finance and insurance. The prediction forecast provides good indicators of any level if it has a large predictive power

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

where $\mathbb{P}_{\hat{\theta}_n}(X_0 \mid X_{-1}, X_{-2}, \ldots)$ is the conditional distribution of the stationary version of the model fitted by the QMLE $\hat{\theta}_n$.

The QMLE for ARMA models estimate those indicators with a quantity proportional to $\hat{\sigma}^2 \approx \sigma^2$ which is approximatively a constant. It is a drawback on the conditional distribution

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

which is dependent on the present observations only for the mean $\Pi_n(\hat{\theta}_n)(X_{n+1})$. 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. Let us consider $X_t = \phi X_t + Z_t$ where (Z_t) is a $\mathrm{WN}(\sigma^2)$. Then the interval of prediction of confidence level $1-\alpha$ is given by

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

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

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

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 3.3.6

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

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

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

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

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

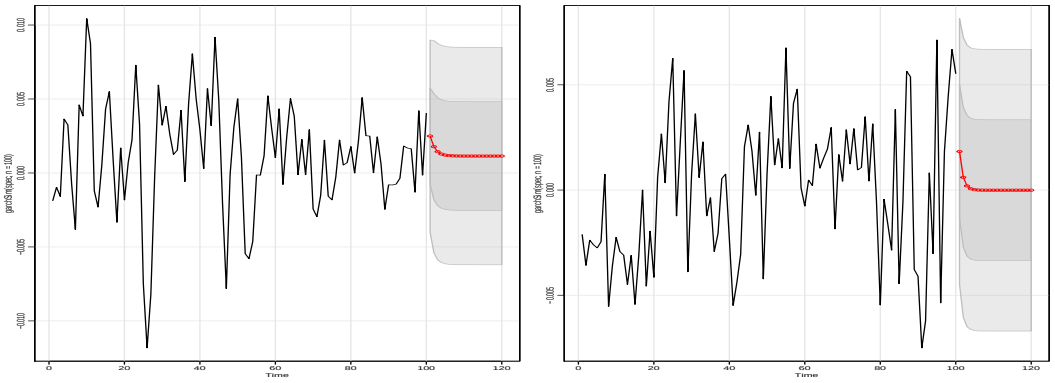


Figure 3.1: The interval of prediction does not take into account the present variability of the time series. When the present variability is low (on the left), the interval is too conservative (too large). On the opposite, when the present variability is high (on the right), the interval is too optimistic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

volatility is predictable, one can estimate it thanks to some model. No ARMA model can as one should model the volatility of the WN as

$$\begin{aligned}\operatorname{Var}(X_t \mid X_{t-1}, X_{t-2}, \ldots) &= \operatorname{Var}(X_t - \mathbb{E}[X_t \mid X_{t-1}, X_{t-2}, \ldots] \mid X_{t-1}, X_{t-2}, \ldots) \\ &= \operatorname{Var}(Z_t \mid X_{t-1}, X_{t-2}, \ldots).\end{aligned}$$

It is the purpose of the next chapter.

Chapter 4

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

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

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

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 $Z_t \in \mathcal{F}_t = \sigma(W_s, s \leq t)$.

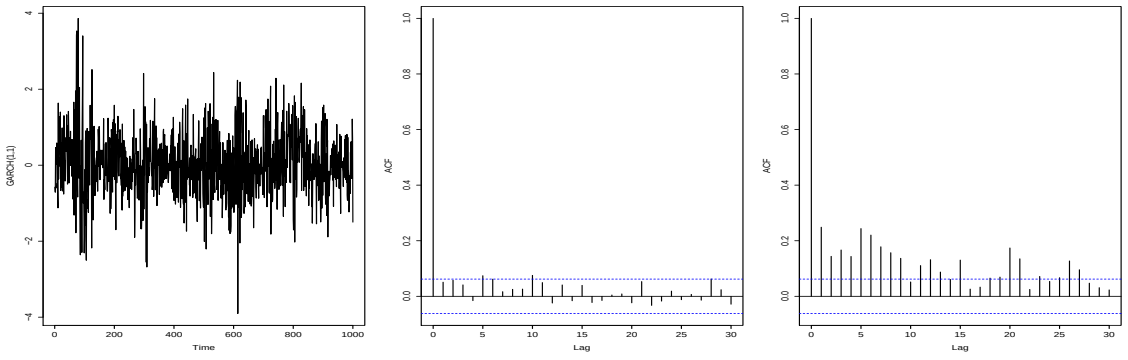


Figure 4.1: A trajectory and the corresponding ACF of the solution of a GARCH(1,1) model and its squares (to be compared with the SWN case)

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.

Exercise.
- Show that if $\alpha+\beta \geq 1$, there exists no second order stationary solution.
- Let $0 \leq \alpha+\beta<1$ and $1-\kappa\alpha^2-\beta^2-2\alpha\beta>0$, with $\kappa = \mathbb{E}\left[Z_t^4\right]$. Show that (σ_t^2) admits a second order stationary solution and determine the kurtosis $\frac{\mathbb{E}\left[X_t^4\right]}{\mathrm{Var}\,[X_t]^2}$.
- Show that if $1-\kappa\alpha^2-\beta^2-2\alpha\beta \leq 0$, then (σ_t^2) has no second order stationary solution.

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 $\mathbb{E}\left[\log(\beta+\alpha Z_0^2)\right] < 0$ and $\mathbb{E}\left[|\log(\beta+\alpha Z_0^2)|\right] < \infty$, 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. 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 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

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

as

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

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

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

which is observable for $t \geq 2$. We approximate $\sigma_t^2(\theta)$ with $\hat{\sigma}_t^2(\theta)$ such that

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

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

Definition. The QMLE is the M-estimator defined as

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

where $\Theta = (0,\infty) \times [0,\infty) \times [0,1)$ and $(\hat{\sigma}_t^2(\theta))$ 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*

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

If moreover $(\hat{\theta}_n)$ 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

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

This result will be proved in full generality for any GARCH(p,q) mode later. In particular we have the identities

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

The QMLE is efficient only if (W_t) gaussian WN(1). In this case $\mathbb{E}[W_0^4] - 1 = 1$ 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 $(\hat{\theta}_n)$, a natural issue is overfitting. Thus, one will test whether one can reject the null hypothesis

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

To do so, one will construct a region of reject of the form $\hat{\theta}_i > c_i$ 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 $\theta_i \approx (\mathrm{se}_i/\sqrt{n})\max\{\mathcal{N}(0,1), 0\}$ in distribution under the null hypothesis. For N standard gaussian r.v., the p-value of the test is $\mathbb{P}(\sqrt{n}\hat{\theta}_i/\mathrm{se}_i \geq \max\{N, 0\}) = \mathbb{P}(N \geq -\sqrt{n}\hat{\theta}_i/\mathrm{se}_i)$, the smallest level of the test that reject the null hypothesis, i.e. the probability to reject the null hypothesis abusively.

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 $\theta = \hat{\theta}_n$. To do so, we differentiate the recursive equation (4.1) followed by $\hat{\sigma}_t^2(\theta)$

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

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

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

invert the approximation and estimate

$$\mathbb{E}[W_0^4] - 1 \approx \frac{1}{n}\sum_{t=1}^{n} \hat{W}_t^2 - 1$$

where $\hat{W}_t = Z_t/\hat{\sigma}_t(\hat{\theta}_n)$ 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 $\hat{\theta}_0$ under the null $\alpha = 0$ as then the volatility is degenerate to $\omega/(1-\beta)$. 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 $\hat{\alpha}_n$.

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, we have $(\omega_0, \alpha_0, \beta_0) \in \overset{\circ}{\Theta}_0$ when $\alpha_0 + \beta_0 = 1$ and $\beta_0 > 0$, θ_0 is uniquely determined as the minimizer of the QLik risk and the asymptotic normality holds. We have

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

where

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

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 $\hat{\alpha}_n + \hat{\beta}_n - 1 > c$ for some constant c, is

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

because the region is one-sided

4.3.3 Invertibility test

If $\hat{\beta}_n \lesssim 1$ under the constraint $\beta < 1$, which is often the case in finance, it is legitimate to ask weather 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. 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). One can show that the QMLE is asymptotically normal when the model is well specified

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

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 $\hat{W}_t$ and plugging in $\hat{\beta}_n$. The p-value of the test with reject region of the form $\hat{\beta}_n < c$ is of the form

$$\mathbb{P}(N < \sqrt{n}(\hat{\beta}_n - 1)/se).$$

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. In that case, 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

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

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

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

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

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

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

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 (I_t) with the volatility model) a prediction interval on X_{n+h}

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

with some abuse of notation as there is two different $\hat{\theta}_n$, one for the ARMA and another for the GARCH. Actually, it is preferable to 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 $\hat{L}_n(\theta) = \sum_{t=1}^n \hat{\ell}_t(\theta)$ computed recursively as follows: Starting from arbitrary initial values, observing recursively X_t,

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

This one-step QMLE is strongly consistent

Theorem (Francq & Zakoïan)**.** *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.*

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. 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 R 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

$$\begin{cases} X_t = (1, 0, \dots, 0)^\top \mathbf{Y}_t + Z_t, & \text{Space equation,} \\ \mathbf{Y}_t = \begin{pmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 1 \\ \phi_r & \cdots & \phi_2 & \phi_1 \end{pmatrix} \mathbf{Y}_{t-1} + \begin{pmatrix} L_1^\top \\ \vdots \\ \vdots \\ L_r^\top \end{pmatrix} \mathbb{X}_{t-1} + \begin{pmatrix} \psi_1 \\ \vdots \\ \vdots \\ \psi_r \end{pmatrix} Z_{t-1}, & \text{State equation,} \end{cases}$$

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 L_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) but this parametrization is not the unique one and suffers over-parametrization. One could prefer the parametrization such that $L_i = L$ 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 $\hat{\mathbf{Y}}_\mathbf{0}$ and Ω_0. Assume now that we have a recursive procedure providing at each step $\hat{X}_n = \Pi_{n-1}(X_n)$, $R_n^L = \mathrm{Var}\,(I_n)$, $\hat{\mathbf{Y}}_n = \Pi_{n-1}(\mathbf{Y}_n)$ and $\Omega_n = \mathbb{E}[(\mathbf{Y}_n - \hat{\mathbf{Y}}_n)(\mathbf{Y}_n - \hat{\mathbf{Y}}_n)^\top]$, the covariance matrix of the prediction error of the state $\mathbf{Y}_n$.

Let us compute $\hat{X}_{n+1} = \Pi_n(X_n)$ 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

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

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

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

The first term is computed recursively using the space equation

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

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

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

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

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

Let us denote $\mathbf{K}_n = \mathbf{F}\Omega_n G/(G^\top\Omega_n G + R)$ 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

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

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

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

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

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

Finally, we have the following theorem

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

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

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 $\text{Var}\,(Z) = R = \sigma^2 = 1$. 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 $\mathcal{F}_t = \sigma(X_t, \ldots, X_1, \hat{\mathbf{Y}}_0, \Omega_0)$ 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 $\Pi_{t-1}(G^\top \mathbf{Y}_t + Z_t) = G^\top \hat{\mathbf{Y}}_t = \hat{X}_t(\theta)$ and variance $\text{Var}\,(G^\top(\mathbf{Y}_i - \hat{\mathbf{Y}}_i)) + 1 = r_t^L(\theta)$. As both the innovations $I_t(\theta) = X_t - \hat{X}_t(\theta)$ and their standardized variances $r_t^L(\theta)$ are computing by the Kalman's recursion, we derive the following recursion for computing the QLik contrast:

- *Initialization:* θ, initial values $\hat{\mathbf{Y}}_{\mathbf{0}}(\theta)$, $\Omega_0(\theta)$ and $L_0(\theta) = 0$
- *New observation* X_n*:*
  1. Compute the innovation $I_n(\theta) = X_n - \hat{X}_n(\theta)$,
  2. Update the QLik loss $L_n(\theta) = L_{n-1}(\theta) + I_n^2(\theta)/\sigma^2 r_n^L(\theta) + \log(\sigma^2 r_n^L(\theta))$,
  3. Compute the next linear prediction $\hat{X}_{n+1}(\theta)$ and the associated standardized risk $r_{n+1}^L(\theta)$ thanks to the Kalman's recursion.

Computing recursively the QLik loss, it is then simple to derive the Quasi Maximum Likelihood Estimator for state-space models:

Definition. The QMLE of a stat-space model is defined as a minimizer $\hat{\theta}_n \in \arg\min_\Theta L_n(\theta)$ where $L_n(\theta)$ is defined recursively thanks to the procedure described above assuming that $R = \sigma^2 = 1$. An estimator of σ^2 is provided by

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

Notice that, neglecting the optimization issues, one should write $\theta_n(\hat{\mathbf{Y}}_\mathbf{0})$ as the whole procedure depends on the initial state $\hat{\mathbf{Y}}_\mathbf{0}(\theta)$ chosen arbitrarily in practice, because the 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. 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

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

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 (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 transform with random coefficients. Notice that the volatility of the GARCH model satisfies such recursion too. The special case $\mu_0 = 1$ 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 μ_0 as big as 1.25 by choosing accordingly the value of σ_0. The stationary solution of such SRE exhibits heavy tails comparable to Pareto distribution

Theorem (Goldie)**.** *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 coefficients c_+ and c_- such that $c_+ + c_- > 0$ and

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

Goldie Theorem is very important as the SRE solution appears as natural heavy-tailed time series. 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 $p > \alpha$.

For $\mu_0 = 1$, 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. We have the following result:

Proposition (Klüppelberg & Pergamenchtchikov)**.** *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 phenomenum and its extremal behavior is really sensitive to the values of F and H.

Those models exhihbit heavy tails because $\mathbb{E}[\theta_0] \approx 1$: the random multiplicative coefficient is fluctuating round 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

$$\begin{cases} X_t = \theta_t^\top \mathbf{X}_t + Z_t & \text{Space equation,} \\ \theta_t = \mathbf{F}_t\theta_{t-1} + \mathbf{H}_t\eta_t & \text{State equation,} \end{cases}$$

where the coefficients $(\mathbf{F}_t)$, $(\mathbf{X}_t)$ and $(\mathbf{H}_t)$ are random and (η_t) and (Z_t) are SWN(I_k) and SWN(σ^2), respectivley. The main assumption is that $(\mathbf{F_t})$, $(\mathbf{X}_t)$ and $(\mathbf{H}_t)$ are stationary ergodic sequences adapted to the filtration $\mathcal{F}_t = \sigma(\eta_t, Z_t, \eta_{t-1}, Z_{t-1}, \ldots)$.

Under the gaussian assumption, working recursively conditionally on $\mathcal{F}_{t-1}$ and using that for normal vectors orthogonality and independence is equivalent, one can extend the Kalman's recursion

Theorem (Kalman)**.** *In a state-space model with random coefficients, under the normal condition and if $\hat{Y}_0$ and Ω_0 are well-chosen, one can compute recursively $\hat{X}_n = \Pi_{n-1}(X_n)$, $\sigma^2 r_n^L = \mathbb{E}[(X_n - \hat{X}_n)^2]$, $\hat{\theta}_n = \Pi_{n-1}(\theta_n)$ and $\sigma^2 v_n = \mathbb{E}[(\theta_n - \hat{\theta}_n)(\theta_n - \hat{\theta}_n)^\top]$ by the following recursion*

$$\begin{aligned} \hat{\theta}_{n+1} &= \mathbf{F}_n\hat{\theta}_n + \frac{\mathbf{F}_n v_n \mathbf{X}_n}{r_n^L}(X_n - \hat{X}_n) \\ \hat{X}_{n+1} &= \hat{\theta}_{n+1}^\top \mathbf{X}_n \\ v_{n+1} &= \mathbf{F}_n v_n \mathbf{F}_n^\top + \mathbf{H}_n\mathbf{H}_n^\top - \frac{\mathbf{F}_n v_n \mathbf{X}_n}{r_n^L}\mathbf{X}_n^\top v_n \mathbf{F}_n^\top \\ r_{n+1}^L &= \mathbf{X}_n^\top v_{n+1}\mathbf{X}_n + 1. \end{aligned}$$

Notice that under the gaussian conditional assumption we have $\hat{X}_{n+1} = \mathbb{E}[X_{n+1} \mid \mathcal{F}_n, \hat{\theta}_0, \Omega_0]$ and $R_{n+1}^L = \sigma^2 r_n^L = \mathrm{Var}\,(X_{n+1} \mid \mathcal{F}_{n+1}, \hat{\theta}_0, \Omega_0)$ when the arbitrary initial values for $\hat{\theta}_0$, and Ω_0 are included in the filtration. Thus 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 $\hat{\theta}_0$ be some starting coefficient corresponding to the (unique) most likely fit on the observations, i.e. the usual OMS. The QLik contrast L_n is then approximated computed recursively:

- *Initialization:* λ, $\hat{\theta}_0$, Ω_0 and $L_0 = 0$
- *New observation* X_n*:*
  1. Compute the innovation $I_n(\lambda) = X_n - \hat{X}_n(\lambda)$,
  2. Update the QLIK loss $L_n(\lambda) = L_{n-1}(\lambda) + I_n(\lambda)^2/\sigma^2 r_n^L(\lambda) + \log(\sigma^2 r_n^L(\lambda))$,
  3. Compute the next linear prediction $\hat{X}_{n+1}(\lambda)$ and the associated risk $r_{n+1}^L(\lambda)$ thanks to the Kalman's recursion.

One can optimize the QLik contrast as before. Notice that λ should stay of dimension relatively small.

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

$$\hat{\theta}_{n+1} = \hat{\theta}_n + \frac{v_n \mathbf{X}_n}{r_n^L}(X_n - \hat{X}_n)$$

coincides with a stochastic gradient algorithm. More precisely, if one consider the problem of minimization of the quadratic loss

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

then one can use a stochastic gradient approach based where

$$\nabla_\theta \ell_t(\theta) = -2\mathbf{X}_t(X_t - \theta^T \mathbf{X}_t).$$

Then the recursion is

$$\hat{\theta}_{n+1} = \hat{\theta}_n - \mu \nabla_\theta \ell_t(\hat{\theta}_n) = \hat{\theta}_n + 2\lambda \mathbf{X}_t(X_t - \theta^T \mathbf{X}_t)$$

where $\mu > 0$ is an hyperparameter called the learning rate. One can identify the learning rates of the Kalman filter as

$$\frac{1}{\sigma^2 r_n^L} v_n.$$

Thus, it is very flexible and reasonable to use. The hyperparameters $\lambda = \mathrm{Diag}(\mathbf{H})$ and σ^2 where and $\lambda \in \mathbb{R}^k$ is estimated thanks to the QMLE $\hat{\lambda}_n$ computed as above and σ^2 is estimated by

$$\hat{\sigma}_n^2 := \frac{1}{n} \sum_{t=1}^{n} \frac{(X_t - \hat{X}_t(\hat{\lambda}_n))^2}{r_t^L(\hat{\lambda}_n)}.$$

Part IV

Stability for stochastic recursions

Chapter 7

Stability of non-linear recursions

7.1 Motivation

Non-linear models are popular because they exhibit a wider range of behaviors than the linear ones. In many applications, some empirical facts are observed:

- The time series can be considered as stationary on some non too long period of observation,
- The stationary marginal distribution exhibits heavy tails, and thus extremal events appears that have much more importance than the usual average behavior,
- The extremal events appears in clusters, i.e. a few of them are observed consecutively in time.

Take a stationary linear model with the moving average representation $X_t = \sum_j \psi_{t-j} Z_j$, (Z_t) iid with $\psi_0 > \psi_j \geq 0$, $j \in \mathbb{Z}$ and (ψ_t) non negative and summable. The only way to obtain an heavy tailed marginal distribution is to assume that the distribution of the Z_t is heavy tailed. Then, by independence, one of the Z_t is much larger than the other, let say Z_0. in that case $X_0 \approx \psi_0 Z_0$ and thus $X_1 \approx \psi_1 Z_0 \approx \psi_1/\psi_0 X_0$. Thus, the dependence between X_1 consecutively of an extremely large X_0 is fixed. It means that if now X_t is the maximum of the observations, then $X_{t+1} \approx \psi_1/\psi_0 X_t$. If $\psi_1 > 0$ a luster of 2 consecutive extremes appears. But the relation between X_t and X_{t+1} is deterministic given that X_t is large, which is not realistic.

To obtain a better understanding of the clustering of the extremes, some nonlinearity is required. One of the most popular model is the Stochastic Recurrent Equation (SRE)

$$X_t = A_t X_{t-1} + B_t, \qquad t \in \mathbb{Z}, \tag{7.1}$$

with (A_t, B_t) iid. Applying the recurrence, one obtain approximatively that given X_0 extremely large then $X_1 \approx A_1 X_0$. Now the relationship between X_1 and X_0 given that X_0 is large is random and thus much wider than in linear models. It is then possible to capture much more different cluster behavior with that simple SRE than with a linear model.

When talking about heavy tail and dependent extreme, the natural question of the stability of the statistical inference is crucial. The normal condition, assuming that the marginals are normally distributed, is no longer relevant as the gaussian distribution is not

heavy tailed. Instead, we choose to focus here on the important concept of the conditional gaussian distributions, meaning that a process exhibiting heavy tails can still be gaussian distributed conditionally to the past. We will make the point clear later as it is crucial to understand how the conditional normal condition does not exclude extremes and actually most of the time generates extremes in a nice non-linear way. This second part of the lecture notes focusses also on exhibiting statistical inferences that are likely to be stable to the occurrence of extreme and in asserting the stability of those procedures. Non-linearity is essential as we have seen that most of the efficient statistical procedures are not linear, even if the models are.

7.2 Stability in statistics

In order to deal with the stability of statistical methods, one has to define some objects on a parametric space Θ that is a compact set (closed and bounded) of $\mathbb{R}^d$, $d \geq 1$.

Definition. A loss function ℓ_n is a random element of $\mathbb{C}(\Theta, \mathbb{R})$, the space of $\mathbb{R}$ valued continuous functions. The randomness of ℓ comes from the one of the observations $X_1, \ldots, X_n$.

We equip $\mathbb{C}(\Theta, \mathbb{R})$ with the supremum norm $\|f\|_\infty = \max_\Theta |f|$.

Proposition. *The space* $(\mathbb{C}(\Theta, \mathbb{R}), \|\cdot\|_\infty)$ *is a Banach space, i.e. a linear space on* $\mathbb{R}$ *that is complete.*

The interest of considering Banach spaces is that all the stability results from the previous section extend to that setting. In particular, assume that (ℓ_t) is a sequence of stationary and ergodic losses such that $\mathbb{E}[\|\ell_0\|_\infty] < \infty$, then the SLLN holds

$$\mathbb{P}\Big(\lim_\infty \Big\| \frac{1}{n} \sum_{t=1}^n \ell_t(\theta) - \mathbb{E}[\ell_0](\theta) \Big\|_\infty = 0 \Big) = 1.$$

Most of the estimator in statistics are constructed as follows: an unknown parameter θ_0 satisfying $\mathbb{E}[\ell_0(\theta_0)] = \min_\Theta \mathbb{E}[\ell_0(\theta)]$ is approximated by the minimizer of the empirical approximation of the asymptotic contrast $\mathbb{E}[\ell_0]$:

7.3 Random Iterated Lipschitz Maps

Before introducing the notion of stability useful in these notes, one needs to restrict ourselves to sufficiently regular recursions.

Definition. A Lipschitz function ϕ on a metric space (E, d) has a finite Lipschitz coefficient

$$\Lambda(\phi) := \sup_{x \neq y} \frac{d(\phi(x), \phi(y))}{d(x, y)} < \infty.$$

If the function ϕ is differentiable, the Lipschitz coefficient is an upper bound of the norm of the first derivative. The Lipschitz property is very useful to ensure stability, even in the deterministic case. Consider the fixed point problem $x = \phi(x)$ then

Theorem (Banach fixed point)**.** *Assume that* (E, d) *is complete (i.e. any Cauchy's sequence converge) then there exists a unique fixed point* $x^* = \phi(x^*)$ *when* ϕ *is contracting, i.e.* $\Lambda(\phi) < 1$*. Moreover* x^* *is exponentially stable that for any* $x_0 \in E$ *it exists* $C > 0$ *such that the iterations* $x_{t+1} = \phi(x_t)$*,* $t \geq 0$ *satisfies*

$$d(x_t, x^*) \leq C\Lambda(\phi)^t d(x_0, x^*) \to 0, \qquad t \to \infty.$$

Notice that the contraction property is necessary. We want to generalize this result to the stochastic case.

Definition. A sequence of Random Iterated Lipschitz Maps on (E, d) is defined as (ϕ_t) such that their Lipschitz coefficients $\Lambda(\phi_t)$ are positive random variables a.s. finite. We associate the forward recursion

$$x_{t+1} = \phi_{t+1}(x_t), \qquad t \geq 0,$$

from an initial value $x_0 \in E$. The random recursion has a solution if it exists a random element x_t that satisfies the forward-backward recursion

$$x_{t+1} = \phi_{t+1}(x_t), \qquad t \in \mathbb{Z}.$$

The forward recursion is constructive and defined a random sequence $(x_t)_{t\geq 0}$ depending on the initial value x_0. In general, it does not provide a solution to the random recursion that has to satisfy

$$x_t = \phi_t(x_{t-1}) = \phi_t \circ \cdots \circ \phi_{t-n+1}(x_{t-n}) =: \phi_t^{(n)}(x_{t-n}),$$

for all $n \geq 0$. In particular, the backward recurrence and the existence of a solution implies that the limit exists a.s as $n \to \infty$.

Definition. The top-Lyapunov coefficient of a sequence of Random Iterated Lipschitz Maps is defined as the limit of $n^{-1}\log(\Lambda(\phi_t^{(n)}))$ if it exists.

The existence of the top-Lyapunov coefficient is ensured by the following result:

Proposition (Kesten-Furstenberg)**.** *Assume that (ϕ_t) is a stationary ergodic sequence such that $\mathbb{E}[\log^+(\Lambda(\phi_0))] < \infty$. Then*

$$\frac{1}{n}\log(\Lambda(\phi_0^{(n)})) \to \inf_{k\geq 1} \frac{\mathbb{E}[\log(\Lambda(\phi_0^{(k)}))]}{k}, \qquad a.s.$$

Proof. Notice that $g_n = \log(\Lambda(\phi_0^{(n)}))$ is a subbadditive sequence of functions of the stationary ergodic process (ϕ_t). □

7.4 Exponential Almost Sure stability

The stability of the recursion will depend on the negativity of the top-Lyapunov coefficient

Theorem (Bougerol)**.** *Let (ϕ_t) be a stationary and ergodic sequence of Lipschitz maps on a complete metric space E. Suppose that*

(S1) *it exists $x \in E$ such that $\mathbb{E}[\log^+ d(\phi_0(x), x)] < \infty$, $\mathbb{E}[\log^+ \Lambda(\phi_0)] < \infty$,*

(S2) *the top-Lyapunov coefficient of the sequence is negative,*

then the recursion $x_t = \phi_t(x_{t-1})$, $t \in \mathbb{Z}$ admits a unique stationary ergodic solution and admits the causal representation

$$x_t = \lim_{n\to\infty} \phi_t^{(n)}(y), \qquad t \in \mathbb{Z}, \quad y \in E.$$

Moreover, the recursion is EAS stable as any approximation $\hat{x}_t = \phi_t(\hat{x}_{t-1})$, $t \geq 1$ starting from any arbitrary $\hat{x}_0$ is such that there exists $\gamma > 1$ satisfying

$$\gamma^t d(\hat{x}_t, x_t) \to 0, \qquad a.s. \text{ (denoted } d(\hat{x}_t, x_t) \xrightarrow{e.a.s.} 0).$$

Proof. We show that $(\phi_0^{(n)}(y))$ converges under (S2). Let $\log \rho < 0$ denotes the top-Lyapunov coefficient. Under (S1), by the Borel-Cantelli Lemma we obtain

$$\frac{1}{n} \log^+ d(\phi_{-n}(x), x) \to_{n\to\infty} 0, \qquad a.s.$$

Thus we have

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log d(\phi_0^{(n+1)}(x), \phi_0^{(n)}(x)) &\leq \limsup_{n\to\infty} \frac{1}{n} \log \Lambda(\phi_0^{(n)}) + \limsup_{n\to\infty} \frac{1}{n} \log^+ d(\phi_{-n}(x), x) \\ &\leq \log \rho.\end{aligned}$$

since $\rho < 1$, the series $(\phi_0^{(n)}(x))$ satisfies the Cauchy criteria and converges. Then $x_t = \lim_{n\to\infty} \phi_t^{(n)}(x)$ is a causal stationary ergodic solution as $x_t = \phi_t(\lim_{n\to\infty} \phi_{t-1}^{(n-1)}(x))$ by continuity.

By stationarity of the (ϕ_t) we also have $\hat{x}_t = \phi_t^{(t)}(\hat{x}_0) =^d \phi_0^{(t)}(\hat{x}_0)$ so that

$$\frac{1}{n} \log(\Lambda(\phi_n^{(n)})) \to \log \rho, \qquad a.s.$$

So there exists a random N sufficiently large and $C > 0$ such that for $n \geq N$ $\Lambda(\phi_n^{(n)}) \leq C\rho^n$ a.s.. As $\rho < 1$, we obtain

$$d(\hat{x}_t, x_t) = d(\phi_t^{(t)}(\hat{x}_0), \phi_t^{(t)}(x_0)) \leq \Lambda(\phi_t^{(t)}) d(\hat{x}_0, x_0) \xrightarrow{e.a.s} 0.$$

The uniqueness follows similarly. □

Notice that (S1) is a very weak condition of moment and is applied by the existence of moments of any order $\varepsilon > 0$ as small as possible because $\log^+(x) \leq x^\varepsilon$, $x > 0$.

We can apply the EAS stability to linear models

Example. Consider the state-space model

$$\begin{cases} X_t = G'\mathbf{Y}_t + W_t, & \text{Space equation,} \\ \mathbf{Y}_t = \mathbf{F}\mathbf{Y}_{t-1} + \mathbf{V}_t, & \text{State equation.} \end{cases}$$

Assume that the WN (V_t) is stationary and ergodic. We want to apply Bougerol theorem on the random state recursion $\phi_t(y_t) = \mathbf{F}y_{t-1} + \mathbf{V}_t$, $t \in \mathbb{Z}$. As $\mathbb{E}[\|V_0\|^2] < \infty$ and $\Lambda(\phi_0^{(n)}) = \|F^n\|$ deterministic, (S1) is satisfied. By Gelfand's formula, $\|F^n\|^{1/n} \to \rho(F)$ and the top-Lyapunov coefficient is $\log(\rho(F))$ It is negative when the spectral radius is smaller than 1, i.e. when the state-space model is stable. Thus, under very weak assumptions, a stable state-space model is also EAS stable. Notice that the same holds automatically to the corresponding observations (X_t).

The stability of the model implies that the model can be nicely simulated; any arbitrary initial values is forgotten exponentially fast in the simulation. However, it does not guaranty the stability of the inference.

7.5 Application to GARCH models

Let us detail the two different random recursions underlying the GARCH(1,1) models. The recursion on the volatility is driven by the Lipschitz map $\phi_t(x) = \omega_0 + (\alpha_0 Z_{t-1}^2 + \beta_0)x$. It is an affine random function from $(0,\infty)$ to $(0,\infty)$. The non-completeness of the space (think of the Cauchy sequence $(1/n)$ tending to 0) can be solved by considering $E = [\omega_0/(1-\beta_0),\infty)$. Notice that ϕ_t is a function of Z_{t-1} and the shift will explain the predictability of the process.

The condition (S1) is satisfied as soon as Z_0 has finite moments. The top Lyapunov coefficients is $\mathbb{E}[\log(\alpha_0 Z_0^2 + \beta_0)]$ by an application of the SLLN to the iid sequence $(\log(\alpha_0 Z_t^2 + \beta_0))$. So (S2) coincides with the condition of stationarity

$$\mathbb{E}[\log(\alpha_0 Z_0^2 + \beta_0)] < 0.$$

We obtain the existence of a unique causal stationary solution EAS stable, even for heavy tailed Z_0 (actually the theorem also applies to any stationary ergodic (Z_t)). It means that knowing $\theta_0 = (\omega_0, \alpha_0, \beta_0)'$, one can simulate the solution of the corresponding GARCH(1,1) model starting from any initial value $\hat{\sigma}_0^2$ and using the recursion

$$\hat{\sigma}_t^2 = \omega_0 + (\alpha_0 Z_{t-1}^2 + \beta_0)\hat{\sigma}_{t-1}^2.$$

It provides an EAS approximation of the volatility (and then of the stationary solution $\hat{X}_t = \hat{\sigma}_t Z_t$).

One cannot use this random recursion to infer the GARCH(1,1) model for two reasons: θ_0 is unknown and Z_t is not observed. Assume that the observations X_i, $1 \le i \le n$, come from a stationary ergodic time series (X_t). Let us use the inverted recursion

$$\phi_t^\theta(x) = \omega + \alpha X_{t-1}^2 + \beta x$$

for any $\theta = (\omega, \alpha, \beta)' \in \Theta$. Notice that now ϕ_t^θ is a function of the past observation X_{t-1} and is then observed for $2 \le t \le n+1$. Consider first the recursion on the complete space $[\omega/(1-\beta),\infty)$. The condition (S1) is satisfied under the existence of finite moments on the observations. The top-Lyapunov coefficient is $\log\beta$ and one can apply Bougerol theorem for any $\beta < 1$. In particular, we obtain that starting from any arbitrary initial value $\hat{\sigma}_0^2(\theta)$ the volatility approximation

$$\hat{\sigma}_t^2(\theta) = \omega + \alpha X_{t-1}^2 + \beta\hat{\sigma}_{t-1}^2(\theta)$$

gets exponentially fast close to some stationary ergodic volatility $\sigma_t^2(\theta)$. Again from Bougerol theorem, we know that this volatility process is the unique solution of the random recursion and can be expressed as the limit of the backward recursion

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

Notice that $\sigma_t^2(\theta)$ coincides with the volatility of (X_t) iff (X_t) is the solution of the GARCH(1,1) model with parameter θ_0 and $\theta = \theta_0$ (the model is invertible). One can check that actually in that case the inverted recursion is always stable as $\log(\beta_0) < E[\log(\alpha_0 Z_0^2 + \beta_0)] < 0$.

The inverted recursion of the stationary GARCH(1,1) model is exponentially stable so that knowing θ_0 one can predict the volatility starting from $\hat{\sigma}_0^2(\theta_0)$ arbitrary and using the recursion

$$\hat{\sigma}_t^2(\theta_0) = \omega_0 + \alpha_0 X_{t-1}^2 + \beta_0\hat{\sigma}_{t-1}^2(\theta_0).$$

We then have the approximation $|\hat{\sigma}_t^2(\theta_0) - \sigma_t^2| \xrightarrow{e.a.s.} 0$.

More generally, for GARCH(p,q) models where

$$\begin{cases} X_t = \sigma_t Z_t \\ \sigma_t^2 = \omega_0 + \beta_{0,1}\sigma_{t-1}^2 + \cdots + \beta_{0,p}\sigma_{t-p}^2 + \alpha_{0,1}X_{t-1}^2 + \cdots + \alpha_{0,q}X_{t-q}^2 \end{cases}$$

we have to consider a random recursion in $\mathbb{R}^{p+q-1}$ with

$$\phi_t(x) = A_{t-1}x + (\omega_0, 0 \ldots, 0)', \qquad t \in \mathbb{Z}$$

and

$$A_t = \begin{pmatrix} \alpha_{0,1}Z_t^2 + \beta_{0,1} & \beta_2 & \cdots & \cdots & \beta_{0,p} & \alpha_{0,2} & \cdots & \cdots & \alpha_{0,q} \\ 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & \ddots & & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & & \vdots \\ 0 & \cdots & 0 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\ Z_t^2 & 0 & \cdots & \cdots & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 1 & \ddots & & \vdots \\ \vdots & & & & & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 0 \end{pmatrix}.$$

The stationary solution will be $(\sigma_t^2, \ldots, \sigma_{t-p+1}^2, X_{t-1}^2, \ldots, X_{t-q}^2)$.

Theorem (Bougerol & Picard)**.** *The GARCH(p,q) model admits a unique stationary ergodic non-anticipative solution iff the top-Lyapunov coefficient*

$$\log \rho = \lim_{n\to\infty} \frac{1}{n} \log \|A_1 \cdots A_n\|$$

is negative. Moreover, the model is EAS stable.

The condition that Z_t is SWN is required for the identity $\sigma_t^2 = \mathrm{Var}\,(X_t \mid X_{t-1}, X_{t-2} \ldots)$.

The inverted recursion holds in $\mathbb{R}^p$ and is driven by

$$\phi_t^\theta(x) = \begin{pmatrix} \beta_1 & \cdots & \cdots & \cdots & \beta_p \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} x + \begin{pmatrix} \omega + \alpha_1 X_{t-1}^2 + \cdots + \alpha_q X_{t-q}^2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

If the time series (X_t) is stationary ergodic, so is (ϕ_t^θ) for any $\theta = (\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)' \in [0,\infty)^{p+q+1}$. Moreover, the model is invertible as soon as the recursion is stable, i.e. when the roots of the lag polynomial $\beta(z) = 1 - \beta_1 z - \cdots - \beta_p z$ are outside the unit disc. However notice that by positivity of the coefficients it is equivalent to $\sum_{j=1}^p \beta_j < 1$. We have the following proposition where $\alpha(z) = 1 + \alpha 1 z + \cdots + \alpha_p z^p$:

Proposition. *Under the stationary condition, the GARCH(p,q) model satisfies necessarily $\sum_{j=1}^p \beta_j < 1$. Then $\sigma_t^2 = \beta_0^{-1} \circ \alpha_0(T) X_{t-1}^2$ is a linear combination of the past observations. Moreover, the recursion is EAS stable.*

So if θ_0 is known one can always predict the volatility of a GARCH model with a recursion admitting an error decreasing exponentially fast to 0 regardless the initial value.

7.6 Other volatility models

We extend the previous discussion to other stochastic volatility models of the form $X_t = \sigma_t Z_t$ with (Z_t) iid centered and normalized so that $\sigma_t^2 = \mathrm{Var}\,(X_t \mid X_{t-1}, X_{t-2}, \ldots)$. The model will specify the dynamic of the volatility. We will distinguish two random recursions, either or not inverted. We have to distinguish when the dynamic holds on σ_t^2 itself or on $\log(\sigma_t^2)$.

7.6.1 Extension of GARCH models

A volatility model will be of GARCH type when the recursion holds on σ_t^2

$$\sigma_t^2 = f_{\theta_0}(\sigma_{t-1}^2, \cdots, \sigma_{t-p}^2, X_{t-1}^2, \ldots, X_{t-q}^2), \qquad t \in \mathbb{Z}.$$

Such extension of the GARCH will be used to model two stylized facts: the possible memory in the volatility by considering p large enough and the leverage effect, i.e. the asymmetry of the response of the volatility to a large positive or negative observation. The volatility should be more impacted by a negative observation.

All those models share with the GARCH one several features; the stationary solution might exist under weak (but not tractable) assumptions. On the opposite, the stability of the inverted recursion will easily follows from the dynamic above if f_θ is a Lipschitz function of its p first coordinates with Lipschitz constants $\Lambda(f_\theta(, X_{t-1}^2, \ldots, X_{t-q}^2))_j$, $1 \le j \le p$, satisfying (S2). One rough but explicit (and optimal in the deterministic case) is

$$\sum_{j=1}^{p} \max_{\theta\in\Theta} \Lambda(f_\theta(\cdot, X_{t-1}^2, \ldots, X_{t-q}^2))_j \le \rho < 1, \qquad a.s.$$

Optimizing on a compact set Θ from an arbitrary positive initial value sufficiently far from 0, the QLIK approach produces reliable estimator $\hat\theta_n$ even if the model is misspecified, for instance when the model corresponding to $\hat\theta_n$ does not have a stationary solution. However the statistician can still use the stable recursion to compute $\hat\sigma_t^2(\hat\theta_n)$ and a reasonable prediction $\hat\sigma_{n+1}^2(\hat\theta_n)$ of a risk measure of the second order. Moreover one can rely on the residuals of the model $X_t/\hat\sigma_t^2(\hat\theta_n)$ to do model adequacy diagnostic.

To quote a few of such models, TGARCH, APARCH, GJR-GARCH, GAS models,...

7.6.2 Log-GARCH models

A volatility model will be of the log-GARCH type when the recursion holds on $\log(\sigma_t^2)$

$$\log(\sigma_t^2) = f_{\theta_0}(\log(\sigma_{t-1}^2), \cdots, \log(\sigma_{t-p}^2), X_{t-1}^2, \ldots, X_{t-q}^2), \qquad t \in \mathbb{Z}.$$

The Log-GARCH type models can model the same stylized facts than the GARCH type models with out the constraint of positivity of f_{θ_0}.

The most popular model of this type is the EGARCH model of Nelson. It is not invertible and should be avoided for any practical purpose. Consider only the symmetric EGARCH(1,1) case for simplicity

$$\log(\sigma_t^2) = \omega_0 + \beta_0 \log(\sigma_{t-1}^2) + \alpha_0 Z_{t-1}, \qquad t \in \mathbb{Z}.$$

The model is a simple AR(1) on the log-volatilities and admits a stationary ergodic non-anticipative solution for $|\beta_0| < 1$. One invert the model plugging in the identity $Z_{t-1} = X_{t-1} e^{-\log(\sigma_{t-1}^2)/2}$ and considering the recursion

$$\phi_t^\theta(x) = \omega + \beta x + \alpha X_{t-1} e^{-x/2}, \qquad t \ge 1.$$

The instability comes from the fact that the function $x \to e^{-x/2}$ is not Lipschitz on $\mathbb{R}$ as the absolute value of the derivative $|x|e^{-x/2}$ is not bounded when $x \to -\infty$. The model is not stable and the arbitrary choice of the initial value $\log(\hat{\sigma}_t^2(\theta))$ impacts the QLIK approach in a non predictable way (depending on the value and sign of the observations X_t, multiplicative coefficients of the unstable exponential term). Worst, even if (X_t) satisfies an EGARCH model with θ_0 known, the inverted recursion does not provide any good approximation of the volatility.

In order to circumvent the unstability the EGARCH representation, the simplest way is to stabilize the inverted recursion. The Log-GARCH models follows an inverted recursion of the form (considering the Log-GARCH(1,1) for simplicity)

$$\phi_t^\theta(x) = \omega + \beta x + \alpha \log(X_{t-1}^2), \qquad t \geq 1.$$

The inverted recursion is stable whenever $|\beta| < 1$ on $E = \mathbb{R}$. The only drawback compared with the GARCH case is that the volatility is not bounded away from 0 (i.e $\log(\hat{\sigma}_t^2(\theta))$ can take possibly very large negative values). In order that the QLIK loss $\ell_t(\theta) = \log(\hat{\sigma}_t^2(\theta)) + X_t^2 e^{-\log(\hat{\sigma}_t^2(\theta))/2}$ is still integrable, one has to assume that the stationary ergodic time series (X_t) is a.s. different from 0 so that $\mathbb{E}[|\log(X_0^2)|]$. The model is driven by the recursion

$$\phi_t(x) = \omega_0 + (\beta_0 + \alpha_0)x + \alpha_0 \log(Z_{t-1}^2), \qquad t \in \mathbb{Z},$$

and a stationary solution exists when $|\beta_0 + \alpha_0| < 1$. The Log-GARCH models are extended to the AS-Log-GARCH(p, q) models in order to capture the leverage effects. Actually, one can show in general that

Proposition (Francq, Wintenberger & Zakoïan)**.** *The volatility of any EGARCH model is a.s. equal to the volatility of an AS-Log-GARCH model. The converse is not true.*

However, it is not true that the observations coincides as the SWN are different in both representations. To conclude, the AS-Log-GARCH is preferable to infer than the EGARCH model.

7.6.3 Stochastic Volatility models

The SV models are volatility models requiring an extra independent SWN (Z_t'). The most common one provides the same dynamic than the EGARCH model replacing Z_t by Z_t'

$$\phi_t(x) = \omega_0 + \beta_0 x + \alpha_0 Z_t', \qquad t \in \mathbb{Z}.$$

As Z_t' is independent of the past observations $(X_{t-1}, X_{t-2}, \ldots)$ it is an alternative way of circumventing the unstability of the EGARCH model. However, the addition of an extra SWN makes the inference more complicated as (Z_t') are not observable and so is

$$\log(\sigma_t^2) = \sum_{j=1}^{\infty} \beta_0^j (\omega_0 + \alpha_0 Z_{t-j}')$$

under the stationary condition $|\beta_0| < 1$. The model is of Hidden Markov Chain type and admits the state-space representation

$$\begin{cases} \log(X_t^2) = \log(\sigma_t^2) + \log(Z_t^2), & \text{Space equation,} \\ \log(\sigma_t^2) = \omega_0 + \beta_0 \log(\sigma_{t-1}^2) + \alpha_0 Z_t', & \text{State equation.} \end{cases}$$

As the $\log x^2$ transform is a bijection on $\mathbb{R}^+$ only, this representation makes sense only when the observations X_t are symmetric. Under this restrictive assumption, the inference will be done using the QLIK contrast computed thanks to the Kalman's recursion. The stability of the procedure is studied in the next chapter. Finally notice that the state space representation is always strongly misspecified as $\log(Z_0^2)$ is far from being normally distributed (usually the normal condition is reasonable on Z_0 itself). It is a major drawback of this approach and the AS-Log-GARCH is preferable to infer than the SV model in practice.

7.7 Stability of state-space models

Definition. A state-space model with the state equation $\mathbf{Y}_t = \mathbf{F}\mathbf{Y}_{t-1} + \mathbf{V}_t$ is stable iff the spectral radius of F is smaller than 1: $\rho(\mathbf{F}) < 1$.

It is indeed a notion of stability; By Gelfand formula we have

$$\lim_{k\to\infty} \|\mathbf{F}^k\|^{1/k} = \rho(\mathbf{F})$$

and so the the series $(\mathbf{F}^k)$ is absolutely convergent by the Cauchy criteria. Then, one can apply the state space recursively and obtain

$$\mathbf{Y}_t = \sum_{j=0}^{k-1} \mathbf{F}^j \mathbf{V}_{t-j} + \mathbf{F}^k \mathbf{Y}_{t-k} \to \sum_{j=0}^{\infty} \mathbf{F}^j \mathbf{V}_{t-j}.$$

There is a unique solution of the state equation that is a causal linear filter. Under the gaussian assumption, it is also non-anticipative, stationary and ergodic. Moreover the state $\mathbf{Y}_t$ forgets its initial position $\mathbf{Y}_0$ exponentially fast.

Let us consider an ARMA model in his canonical state-space representation

$$\begin{cases} X_t = (1,0,\dots,0)'\mathbf{Y}_t + Z_t, & \text{Space equation,} \\ \mathbf{Y}_t = \begin{pmatrix} \phi_1 & \phi_2 & \cdots & \phi_r \\ 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 1 & 0 \end{pmatrix} \mathbf{Y}_{t-1} + \begin{pmatrix} \psi_r \\ \vdots \\ \vdots \\ \psi_1 \end{pmatrix} Z_{t-1}, & \text{State equation.} \end{cases}$$

We rewrite the space equation as $\mathbf{Y}_t = \mathbf{F}\mathbf{Y}_{t-1} + HZ_{t-1}$. Noticing that $\mathbf{V}_t = HZ_{t-1}$ we obtain

$$\mathbf{Y}_t = \sum_{j=1}^{\infty} \mathbf{F}^{j-1} H Z_{t-j},$$

so that

$$X_t = Z_t + \sum_{j=1}^{\infty} G'\mathbf{F}^{j-1} H Z_{t-j}, \qquad t \in \mathbb{Z}.$$

By unicity, it proves that under the stability condition, the (ψ_j) of the linear representation of any ARMA process satisfies $\psi_0 = 1$ and

$$\psi_j = G'\mathbf{F}^{j-1}H, \qquad j \geq 1.$$

It proves that the ψ_j depends only on H and are geometrically decreasing.

The stability assumption is equivalent to the fact that all the eigenvalues of $\mathbf{F}$ are inside the unit circle, i.e. that the characteristic polynomial $\det(\lambda I_r - \mathbf{F})$ has no root outside the unit circle. In the ARMA case, one can compute explicitly the characteristic polynomial $\lambda^n - \phi_1\lambda^{n-1} - \cdots - \phi_n = \phi(\lambda^{-1})\lambda^n$. Thus, it is equivalent that the polynomial ϕ has no root inside the unit disc. We recognize the causal condition for ARMA processes. However, it is not enough to ensure that $\psi_j = 0$ for $j < 0$. By a symmetric argument, one also imposes that the polynomial γ does not have roots inside the unit circle so that $Z_t = \gamma \circ \phi(T)X_t$. Then X_t is invertible under the gaussian assumption and the existence of the ψ_j with $\psi_j = 0$ is ensured. In view of the condition of existence of a unique stationary solution, we will always work under the slightly more restrictive condition

Definition (Hannan)**.** The QMLE will be the minimizer of the QLIK contrast over the set Θ that is the compactification of

$\{\phi$ and γ do not have roots inside the unit circle and do not have common roots $\}$.

Under the Hannan condition, the ψ_j are well-defined, they satisfy $\psi_j = 0$ for $j < 0$,

$$\psi_j - \sum_{0<k\leq j} \phi_k \psi_{j-k} = \gamma_j, \qquad 0 \leq j \leq r,$$

and $\psi_j = G'\mathbf{F}^{j-1}H$ for $j > r$ (the recursive formula has been obtain through the identity $\psi\phi = \gamma$).

We will see that those conditions on the unknown parameter θ are enough for proving the consistency of the QMLE. We showed that they ensured the stability of the state-space model. However, they do not ensure the stability of the Kalman's recursion

$$\Omega_{n+1} = \mathbf{F}\Omega_n\mathbf{F}' + \mathbf{Q} - \frac{\mathbf{F}\Omega_n G}{G'\Omega_n G + R} G'\Omega_n\mathbf{F}'.$$

is much more complicated (i.e. non-linear) and does not have an explicit solution. Before treating the stability of non-linear system in generality, let us introduce useful notions for describing state-space models.

Chapter 8

Asymptotic properties of the QMLE under continuous invertibility

8.1 Continuous invertibility

In practice θ_0 remains unknown and one has to estimate it. From the stability of the inverted recursion, we can construct an approximation of the QLIK loss called the inverted QLIK loss $\hat{\ell}_t(\theta) = (X_t - \hat{X}_t(\theta))^2/R_t^L(\theta) + \log(R_t^L(\theta))$ from arbitrary initial values. From Pfanzagl theorem, the QMLE θ_n is strongly consistent if it is the minimizer of the QLIK contrast L_n. However, one actually had to consider the inverted QLIK contrast $\hat{L}_n = \sum_{t=1}^n \hat{\ell}_t$ instead. Let us define

Definition. A model is continuously invertible if it is invertible and if the inverted QLIK loss satisfies $\|\hat{\ell}_t - \ell_t\|_\infty \xrightarrow{e.a.s.} 0$ whatever is the arbitrary (continuous) choice of $\hat{\ell}_0$. It implies that the inverted QLIK contrast $\hat{L}_n$ is continuous and a minimizer over Θ is called the QMLE and denoted $\hat{\theta}_n$.

We then have the following consequence

Theorem (Wintenberger)**.** *Assume that (X_t) is a stationary ergodic time series. Assume that the model is continuously invertible on Θ. Then the QMLE $\hat{\theta}_n$ is robust and strongly convergent if $\Theta_0 = \{\theta_0\}$.*

Bougerol theorem is useful to check the continuous invertibility condition.

Example. Consider the inverted recursion of the GARCH(1,1) model

$$\phi_t^\theta(x) = \omega + \alpha X_{t-1}^2 + \beta x$$

on the space $E = \mathbb{C}(\Theta, [\underline{\omega}, \infty))$, the complete space of continuous functions $\hat{\sigma}_t^2$ from the compact set $\Theta = [\underline{\omega}, \bar{\omega}] \times [0, \bar{\beta}] \times [0, \bar{\alpha}]$ to $[b = \underline{\omega}, \infty)$. If the stationary ergodic time series (X_t) has some finite moments then (S1) is satisfied. If $\bar{\beta} < 1$ then (S2) is satisfied uniformly and Bougerol theorem applies. It provides the EAS stability uniformly over Θ automatically and thus the continuous invertibility.

The same holds for the GARCH(p, q) model: as soon as the model is invertible, it is continuously invertible on some compact set Θ. We are now ready to assert the strong consistency of the QMLE:

Theorem (Francq & Zakoïan)**.** *Assume that the GARCH(p,q) model is invertible on the compact set Θ and that (X_t) is the stationary solution of the model for $\theta_0 \in \Theta$ such that the top-Lyapunov is negative. Then the QMLE is strongly consistent if α_0 and β_0 do not have common roots.*

The last condition ensures the identifiability of the model, the model being regular by assumption. We also obtain a similar result for ARMA(p,q) models:

Theorem. *Assume that (X_t) is a stationary ergodic time series. Under the Hannan's condition on the compact set Θ for the ARMA model then the QMLE is robust and strongly convergent if $\{\theta_0\} = \Theta_0$, in particular strongly consistent if (X_t) satisfies the ARMA model with $\theta_0 \in \Theta$ and (Z_t) SWN.*

8.2 Inference in ARMA-GARCH models

One of the most common procedure is to fit an ARMA model on the observations and then a volatility on the residuals if necessary. However, this two steps procedure (applying two QMLE successively) should be replaced by the following one-step procedure. As soon as the diagnostic on the residuals suggested some correlations in the squares (or in the absolute values), if one seek at fitting a volatility model such as the GARCH(1,1) for instance, one should consider the process

$$\begin{aligned} X_t &= \phi_{0,1}X_{t-1} + \cdots + \phi_{0,p}X_{t-p} + \gamma_{0,1}\varepsilon_{t-1} + \cdots + \gamma_{0,q}\varepsilon_{t-p}, \\ \varepsilon_t &= \sigma_t Z_t, \qquad t \in \mathbb{Z} \\ \sigma_t^2 &= \omega_0 + \beta_0\sigma_{t-1}^2 + \alpha_0\varepsilon_{t-1}^2. \end{aligned}$$

Here (Z_t) is SWN(1) and the unknown parameter

$$\theta_0 = (\phi_{0,1}, \ldots, \phi_{0,p}, \gamma_{0,1}, \ldots, \gamma_{0,q}, \omega_0, \alpha_0, \beta_0)' \in \mathbb{R}^d, \qquad d = p + q + 3,$$

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

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

This one-step QMLE is strongly consistent

Theorem (Francq & Zakoïan)**.** *If $\theta_0 \in \Theta$ satisfies the condition of stationarity of the GARCH model, if any $\theta \in \Theta$ satisfies the Hannan's condition and $\beta < 1$, then the QMLE is strongly consistent.*

The one-step estimation is more robust to possibly heavy tailed observations as the condition of finite second moments on (Z_t) does not imply second finite moments on (ε_t) and thus also on (X_t).

More generally, we have also the following corollary that holds in both volatility and state-space models

Corollary. *Assume that (X_t) satisfies the stationary ergodic state-space model for $\theta_0 \in \Theta$ with SWN (with $\mathbb{E}[X_1 \mid X_0, X_{-1}, \ldots] = 0$). Then $\hat{X}_{n+1}(\hat{\theta}_n)$ and $R^L_{n+1}(\hat{\theta}_n)$ (or $\hat{\sigma}^2_{n+1}(\hat{\theta}_n)$) are strongly consistent predictions of $\Pi_n(X_{n+1})$ and R^L_{n+1} or $Var(X_{n+1} \mid X_n, X_{n-1}, \ldots)$, regardless the arbitrary initial choices in the recursion.*

It is very important for applications

1. Analysis of the residuals: Compute the residuals $\hat{Z}_t = (X_t - \hat{X}_t(\hat{\theta}_n))/\sqrt{R^L_t(\hat{\theta}_n)}$ or $\hat{Z}_t = X_t/\hat{\sigma}_t(\hat{\theta}_n)$. Then apply the diagnostic of the adequacy of the model and check if the residuals are correlated, if the squares are correlated, if the gaussian assumption is likely to hold,...

2. Quantitative prediction and risk management: After a positive diagnostic, the prediction $\hat{X}_{n+1}(\hat{\theta}_n)$ and its risk $R^L_{n+1}(\hat{\theta}_n)$ can be used to construct intervals of prediction. The range of the interval informs about the uncertainty degree of the prediction. In finance, the volatility prediction $\hat{\sigma}_t(\hat{\theta}_n)$ is used as a quantitative measure of risks.

8.3 Asymptotic normality of the QMLE

The asymptotic normality of the QMLE follows in most of the cases under extra assumptions. Assume that $\Theta_0 = \{\theta_0\} \subset \mathbb{R}^d$. If $\hat{L}_n$ is sufficiently regular (2-times continuously differentiable) then a Taylor expansion gives

$$\partial_\theta \hat{L}_n(\hat{\theta}_n) = \partial_\theta \hat{L}_n(\theta_0) + \partial^2_\theta \hat{L}_n(\tilde{\theta}_n)(\hat{\theta}_n - \theta_0)$$

with $\tilde{\theta}_n \in [\theta_0, \hat{\theta}_n]$. Notice that as $\hat{\theta}_n$ is strongly convergent, then $[\theta_0, \hat{\theta}_n] \to \{\theta_0\}$ a.s. Morreover, if $\hat{\theta}_n \in \overset{\circ}{\Theta}$ the interior of the compact set then $\partial_\theta \hat{L}_n(\hat{\theta}_n) = 0$ as the QMLE is the minimizer of the QLIK contrast by assumption. So we have to study the properties of the two first derivative of the contrast $\hat{L}_n$. Let us first show that the two first derivatives of L_n have nice properties at θ_0:

Definition. The score vector is defined as the gradient of the QLIK loss

$$S_t = -\nabla_\theta \log(f_{\theta_0, t-1}(X_t)).$$

The Fisher's information is $\mathcal{I}(\theta_0) = \mathbb{E}[\partial^2_\theta(-\log(f_{\theta_0,0}(X_1)))]$.

We have the following property, deriving from the definition of θ_0 as the unique minimizer of $-\mathbb{E}[\log(f_{\theta,0}(X_1)) \mid \mathcal{F}_0]$:

Proposition. *If $\theta_0 \in \overset{\circ}{\Theta}$ the score vector is centered $\mathbb{E}[S_1 \mid \mathcal{F}_0] = 0$ and $\mathcal{I}(\theta_0)$ is a symmetric definite positive matrix.*

Thus the sequence of score vectors (S_t) constitutes a difference of martingale. Under the gaussian assumption, when $h_\theta(z, X_0, X_1, \ldots) = \mathbb{E}_\theta[X_1 \mid \mathcal{F}_0] + \sqrt{\mathrm{Var}_\theta(X_1 \mid \mathcal{F}_0)}z$ and so we have a specific form of the score vector

$$\begin{aligned} S_1 =& \frac{\mathrm{Var}_{\theta_0}(X_1 \mid \mathcal{F}_0) - (X_1 - \mathbb{E}_{\theta_0}[X_1 \mid \mathcal{F}_0])^2}{\mathrm{Var}_{\theta_0}(X_1 \mid \mathcal{F}_0)^2} \nabla_\theta \mathrm{Var}_{\theta_0}(X_1 \mid \mathcal{F}_0) \\ &+ 2\frac{\mathbb{E}_{\theta_0}[X_1 \mid \mathcal{F}_0] - X_1}{\mathrm{Var}_{\theta_0}(X_1 \mid \mathcal{F}_0)} \nabla_\theta \mathbb{E}_{\theta_0}[X_1 \mid \mathcal{F}_0] \end{aligned}$$

Under the assumption made above that the conditional distribution of X_1 given $\mathcal{F}_0$ is $\mathbb{E}[X_1 \mid \mathcal{F}_0] + \sqrt{\mathrm{Var}\,(X_1 \mid \mathcal{F}_0)} Z_1$ for some misspecified Z_1 and that the model is rich enough to have $\mathbb{E}_{\theta_0}[X_1 \mid \mathcal{F}_0] = \mathbb{E}[X_1 \mid \mathcal{F}_0]$ and $\mathrm{Var}_{\theta_0}(X_1 \mid \mathcal{F}_0) = \mathrm{Var}\,(X_1 \mid \mathcal{F}_0)$, we rewrite the score vector

$$S_1 = \frac{1 - Z_1^2}{\mathrm{Var}\,(X_1 \mid \mathcal{F}_0)} \nabla_\theta \mathrm{Var}_{\theta_0}(X_1 \mid \mathcal{F}_0) - 2 \frac{Z_1}{\sqrt{\mathrm{Var}\,(X_1 \mid \mathcal{F}_0)}} \nabla_\theta \mathbb{E}_{\theta_0}[X_1 \mid \mathcal{F}_0]$$

When the SWN (Z_t) is centered and normalized one can check directly that (S_t) is a martingale difference. We then compute, shortening the notation,

$$\begin{aligned}\mathrm{Var}\,(S_1) =& (\mathbb{E}[Z_1^4] - 1)\mathbb{E}\left[\frac{\nabla_\theta \mathrm{Var}_{\theta_0} \nabla_\theta \mathrm{Var}'_{\theta_0}}{\mathrm{Var}^2_{\theta_0}}\right] + 4\mathbb{E}\left[\frac{\nabla_\theta \mathbb{E}_{\theta_0} \nabla_\theta \mathbb{E}'_{\theta_0}}{\mathrm{Var}_{\theta_0}}\right] \\ & - 2\mathbb{E}[Z_1^3]\mathbb{E}\left[\frac{\nabla_\theta \mathrm{Var}_{\theta_0} \nabla_\theta \mathbb{E}'_{\theta_0} + \nabla_\theta \mathbb{E}_{\theta_0} \nabla_\theta \mathrm{Var}'_{\theta_0}}{\mathrm{Var}^{3/2}_{\theta_0}}\right].\end{aligned}$$

Under the same assumptions, we have

$$\mathcal{I}(\theta_0) = \mathbb{E}\left[\frac{\nabla_\theta \mathrm{Var}_{\theta_0} \nabla_\theta \mathrm{Var}'_{\theta_0}}{\mathrm{Var}^2_{\theta_0}}\right] + 2\mathbb{E}\left[\frac{\nabla_\theta \mathbb{E}_{\theta_0} \nabla_\theta \mathbb{E}'_{\theta_0}}{\mathrm{Var}_{\theta_0}}\right].$$

We can check that the Fisher's information is a symmetric definite positive matrix. We obtain the identity $\mathrm{Var}\,(S_1) = 2\mathcal{I}(\theta_0)$ under the normal condition or more generally when $\mathbb{E}[Z_1^3] = 0$ and $\mathbb{E}[Z_1^4] = 3$.

Those formulas apply on the asymptotic contrast for L_n and not the inverted $\hat{L}_n$. However, we have

Definition. A model is $\mathbb{C}^2$-invertible on a compact set Θ if it is continuously invertible on Θ and $\partial_\theta \hat{\ell}_t$ is twice continuously differentiable on $\overset{\circ}{\Theta}$ such that $\|\partial_\theta \hat{\ell}_t - \partial_\theta \ell_t\|_\infty \xrightarrow{e.a.s.} 0$ and $\|\partial^2_\theta \hat{\ell}_t - \partial^2_\theta \ell_t\|_\infty \xrightarrow{e.a.s.} 0$ on any compact subsets of $\overset{\circ}{\Theta}$.

Models that are continuously invertible and regular are most of the time $\mathbb{C}^2$-invertible. We then obtain the uniform approximations

$$\frac{1}{n}\|\hat{L}_n - L_n\|_\infty \xrightarrow{a.s.} 0, \qquad \frac{1}{\sqrt{n}}\|\partial_\theta \hat{L}_n - \partial_\theta L_n\|_\infty \xrightarrow{a.s.} 0 \quad \text{and} \quad \frac{1}{n}\|\partial^2_\theta \hat{L}_n - \partial^2_\theta L_n\|_\infty \xrightarrow{a.s.} 0.$$

We are now ready to state

Theorem. *Assume that (X_t) is a stationary ergodic time series. If the model is $\mathbb{C}^2$-invertible on a compact set Θ, if $\Theta_0 = \{\theta_0\} \subset \overset{\circ}{\Theta}$ and $Var(S_1) < \infty$ then the QMLE is asymptotic normal*

$$\sqrt{n}(\hat{\theta}_n - \hat{\theta}_0) \xrightarrow{d.} \mathcal{N}_d(0, \mathcal{I}(\theta_0)^{-1} Var(S_1) \mathcal{I}(\theta_0)^{-1}).$$

Proof. From the Taylor's expansion

$$0 = \frac{1}{\sqrt{n}} \partial_\theta \hat{L}_n(\theta_0) + \frac{1}{n} \partial^2_\theta \hat{L}_n(\tilde{\theta}_n) \sqrt{n}(\hat{\theta}_n - \theta_0)$$

The first term approximates a.s. $\frac{1}{\sqrt{n}}\partial_\theta L_n(\theta_0)$. It is the renormalized sum of the vector scores on which we apply the CLT for martingale of Billingsley and we obtain

$$\frac{1}{\sqrt{n}}\partial_\theta \hat{L}_n(\theta_0) \xrightarrow{d.} \mathcal{N}_d(0, \mathrm{Var}\,(S_1)).$$

On the other hand, the second term $\frac{1}{n}\partial^2_\theta \hat{L}_n(\tilde{\theta}_n)$ is an a.s. approximation of $\frac{1}{n}\partial^2_\theta L_n(\tilde{\theta}_n)$ that converges to $\mathcal{I}(\theta_0)$ because $\tilde{\theta}_n \in [\theta_0, \hat{\theta}_n] \to \{\theta_0\}$ and because $\mathrm{Var}\,(S_1) < \infty$ ensures that $\mathcal{I}(\theta_0) < \infty$. Finally, we obtain the desired result by applying Slutsky Lemma. □

The condition of moments $\mathrm{Var}\,(S_1) < \infty$ is restrictive in terms of finite moments on (Z_t) and (X_t). In particular, it always requires that $\mathbb{E}[Z_1^4] < \infty$. If $\mathbb{E}[Z_1^3] = 0$ and $\mathbb{E}[Z_1^4] = 3$, the asymptotic variance is $2\mathcal{I}(\theta_0)^{-1}$. We recover the Cramer-Rao bound $\mathcal{I}(\theta_0)^{-1}$ up to a factor 2 due to the dependence.

In the dependent case, it is in general difficult to assert the efficiency, i.e. the optimality of the asymptotic variance. If one knows in advance the distribution of (Z_t), it is always more efficient to use it in the likelihood (whenever one can still compute it recursively).

For GARCH models, one can identify the Fisher's information and we obtain

Theorem (Francq & Zakoïan)**.** *Assume that the GARCH(p,q) model is invertible on the compact set Θ and that (X_t) is the stationary solution of the model for $\theta_0 \in \overset{\circ}{\Theta}$ such that the top-Lyapunov is negative. Then the QMLE is asymptotically normal if α_0 and β_0 do not have common roots and $\mathbb{E}[Z_1^4] < \infty$:*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d.} \mathcal{N}_{p+q+1}\left(0, (\mathbb{E}[Z_1^4] - 1)\mathbb{E}\left[\frac{\nabla_\theta \sigma_0^2(\theta_0)\nabla_\theta \sigma_0^2(\theta_0)'}{\sigma_0^4(\theta_0)}\right]^{-1}\right).$$

Notice that one can always estimate the asymptotic variance thanks to the EAS stable inverted recursions satisfied by $\sigma_0^2(\hat{\theta}_n)$ and $\nabla_\theta \sigma_0^2(\hat{\theta}_n)$ starting form arbitrary initial values. One also needs to estimate $\mathbb{E}[Z_1^4]$ thanks to the residuals.

For ARMA models, one can also identify the Fisher's information and we obtain

Theorem (Hannan)**.** *Under the Hannan's condition on the compact set Θ, (Z_t) is SWN and that (X_t) is the stationary solution of the model for $\theta_0 \in \overset{\circ}{\Theta}$ then the QMLE is asymptotically normal*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d.} \mathcal{N}_{p+q}\left(0, \sigma^2\, Var\,(AR_1, \ldots, AR_p, MA_1, \ldots, MA_q)^{-1}\right)$$

where (AR_t) and (MA_t) are the stationary AR(p) and AR(q) time series driven by the coefficient θ_0 and the same SWN(1) (Z_t):

$$\phi_0(T)AR_t = Z_t, \qquad \gamma_0(T)MA_t = Z_t, \qquad t \in \mathbb{Z}.$$

Moreover, R^L_{n+1} is a strongly consistent estimator of σ^2.

Notice that the asymptotic variance (except σ^2) is an explicit function of θ_0. It can be estimated by replacing θ_0 by $\hat{\theta}_n$ in this explicit formula.

The unknown parameter μ_0, ω_0 and σ_0^2 can be estimated under the normal condition thanks to the general QMLE approach of Francq and Zakoïan for ARMA-GARCH models described above. However, the estimation procedure of ω_0 and σ_0^2 is still stable in the unit root case $\mu_0 = 1$:

Theorem (Nielsen & Rabhek)**.** *The QMLE is strongly consistent and asymptotically normal on any compact set* $\Theta \subset (0,\infty)\times[0,\infty)$ *when* (X_t) *is the solution of the SRE for* $\theta_0 \in \mathring{\Theta}$ *satisfying the stationary condition. Moreover the asymptotic variance is the Cramer-Rao bound* $\mathcal{I}(\theta_0)^{-1}$*.*

As in the iid case, the asymptotic variance can be estimated thanks to the second order properties of the minimization routine; the QMLE is asymptotically normal even when the stationary solution of the SRE does not have finite variance.

Chapter 9

Stability of the Kalman's recursion

9.1 Controllability

The controllability is another notion describing the state-space model that depends only on the state equation $\mathbf{Y}_t = \mathbf{F}\mathbf{Y}_{t-1} + \mathbf{V}_t$ that we rewrite $\mathbf{Y}_t = \mathbf{F}\mathbf{Y}_{t-1} + \mathbf{H}\eta_t$ where $\mathbf{H}$ is a full rank matrix such that $\mathbf{H}\mathbf{H}' = \mathbf{Q}$

Definition. The state-space model is controllable iff for any two vectors A and B there exists an integer k and values $\eta_1, \ldots, \eta_k$ such that $\mathbf{Y}_0 = A$ and $\mathbf{Y}_k = B$.

One says that the state can reach any possible value in finite time starting from any other possible value. When the η are SWN, then the states $\mathbf{Y}_t$ constitute a Markov chain and the notion of controllability is related to irreducibility properties. The following proposition provides a checkable equivalent definition

Proposition. *The state space model is controllable iff the matrix*

$$(\mathbf{H} \quad \mathbf{F}\mathbf{H} \quad \cdots \quad \mathbf{F}^{r-1}\mathbf{H})$$

has full rank r, the dimension of the state space model.

Controllability is an important notion because of the following result

Proposition. *If the state-space model is not controllable then there exists another representation of lower dimension that is controllable.*

9.2 Observability

The observability notion relies on the state X_t satisfying

$$\begin{cases} X_t = G'\mathbf{Y}_t + W_t, & \text{Space equation,} \\ \mathbf{Y}_t = \mathbf{F}\mathbf{Y}_{t-1} + \mathbf{V}_t, & \text{State equation.} \end{cases}$$

Definition. The state-space model is observable iff the initial state $\mathbf{Y}_0$ can be completely determined from all possible future observations $X_1, X_2, \ldots$ $(n \to \infty)$ when $W_t = 0$ and $\mathbf{V}_t = 0$.

The following proposition provides a checkable equivalent definition

Proposition. *The state space model is observable iff the matrix*

$$(G \quad \mathbf{F}G \quad \cdots \quad \mathbf{F}^{r-1}G)$$

has full rank r, the dimension of the state space model.

Controllability is an important notion because of the following result

Proposition. *If the state-space model is not observable then there exists another representation of lower dimension that is observable.*

We are now ready to state the Theorem showing the importance of those notions

Theorem. *The state-space model is of minimal dimension iff it is controllable and observable.*

One can check that the canonical representation of the ARMA model is actually minimal. It is crucial for the application of the Kalman's recursion and the QLIK criteria, proving that the whole procedure is relying on the sparsest possible representation. We will also see that it is fundamental for the stability of the Kalman's recursion (that does not have to be mixed up with the stability of the state-space model).

9.3 Stability of the Kalman's recursion

Consider the state space model

$$\begin{cases} X_t = G'\mathbf{Y}_t + W_t, & \text{Space equation,} \\ \mathbf{Y}_t = \mathbf{F}\mathbf{Y}_{t-1} + \mathbf{V}_t, & \text{State equation.} \end{cases}$$

where the variance R of the WN (W_t) is assumed to be 1. Let θ be the parameters of the model determining G, $\mathbf{F}$ and $\mathbf{Q}$. We want to assert the invertibility of the model thanks to the EAS stability of the inverted recursion that coincides here with the Kalman's recursion. The stability of this random recursion depends only on the stability of the Riccati's equation

$$\Omega_{n+1}(\theta) = \mathbf{F}\Omega_n(\theta)\mathbf{F}' + \mathbf{Q} - \frac{\mathbf{F}\Omega_n(\theta)G}{G'\Omega_n(\theta)G + 1} G'\Omega_n(\theta)\mathbf{F}'.$$

It is a deterministic recursion starting from $\Omega_0(\theta)$ arbitrary and driven by the non linear transform

$$\phi_t^\theta(x) = \mathbf{F}x\mathbf{F}' + \mathbf{Q} - \frac{\mathbf{F}xG}{G'xG + 1} G'x\mathbf{F}'.$$

We have the following fundamental result:

Theorem (Kalman)**.** *The Kalman's recursion is EAS stable iff the state-space model is controllable and observable.*

It implies that the state-space model has minimal dimension. As a corollary, we have the continuous invertibility of the ARMA models with SWN.

Corollary. *Under the Hannan's condition, the canonical representation of the ARMA model generates an EAS stable Kalman recursion.*

The EAS stability of the Kalman's recursion is not implied by the invertibility of the ARMA model only. It is because the state-space model representation requires the condition of causality as any stable sate-space model admits a causal linear representation. Moreover, it implies that the state space representation is stable and also the ARMA model. One excludes the explosive cases such as the (widely) invertible AR(1) $X_t = \phi X_{t-1} + Z_t$, $t \in \mathbb{Z}$ with $|\phi| > 1$.

As the Riccati's recursion is deterministic, we face a fixed point problem with a unique solution $\Omega^*(\theta)$ known as the steady state and satisfying

$$\Omega^*(\theta) = \mathbf{F}\Omega^*(\theta)\mathbf{F}' + \mathbf{Q} - \frac{\mathbf{F}\Omega^*(\theta)G}{G'\Omega^*(\theta)G + 1} G'\Omega^*(\theta)\mathbf{F}'.$$

The steady state can be approximated numerically and it is the best choice for the initial value $\Omega_0(\theta) = \Omega^*(\theta)$. Indeed, we then ensure that $\Omega_n(\theta) = \Omega^*(\theta)$ for any $n \geq 0$ saving computation time in the Kalman's recursion.

Notice that it is possible to have an EAS stable Kalman recursion for a state-space model that is not stable. It is for instance the case of ARIMA processes (integrated ARMA) that admits an unstable representation for which the Kalman's recursion is still EAS stable. Then, it is likely to use a non-informative a-priori, i.e. a value $\Omega_0(\theta)$ big such that the first $\hat{\mathbf{Y}}_\mathbf{t}(\theta)$s in the inverted recursion have large variances and explore different states. The procedure will then stabilize by itself after some runs but cannot be stabilized a priori and the steady state is not a good choice (Koopman).

9.4 Time varying coefficient stability

The stability of the inverted recursion will depend on weakened controllability and observability notions.

Definition. The state-space model with random coefficients is weakly controllable iff for any two vectors A and B there exists an integer k and values $\eta_1, \ldots, \eta_k$ such that $\mathbf{Y}_0 = A$ and $\mathbf{Y}_k = B$ with some positive probability. It is weakly observable iff the initial state $\mathbf{Y}_0$ can be determined with some positive probability from all possible future observations $X_1, X_2, \ldots$ $(n \to \infty)$ when $W_t = 0$ and $\eta_t = 0$, $t \geq 1$.

Under weak controllability and stability the Kalman's recursion is stable in a random environment:

Theorem (Bougerol)**.** *The Kalman's recursion is EAS stable if the state space model is weakly controllable and observable and if* $\mathbf{F}_t$ *is an invertible random matrix for* $t \geq 1$.

It is a complete extension of the Kalman's theorem with constant coefficients except the invertibility assumption on the matrices $\mathbf{F}_t$'s. Notice that the stability of the state-space model is not required.

We also assume implicitly that they have finite log-moments. We can extend the notions of stability, controllability and observability of the deterministic case:

Definition. The state space model with random coefficients is stable if the top-Lyapunov coefficient

$$\log(\rho) = \lim_{n\to\infty} \frac{1}{n} \log(\|\mathbf{F}_1 \cdots \mathbf{F}_n\|)$$

is negative.

Applying Bougerol theorem on the state equation, we obtain

Proposition (Bougerol)**.** *If the state-space model is stable, then it admits a unique stationary ergodic solution that is not anticipative.*