MATH 545 - Intro to Time Series Analysis

Pr. David A. Stephens

Course notes by
Léo Raymond-Belzile
Leo.Raymond-Belzile@mail.mcgill.ca

The current version is that of August 21, 2015
Fall 2012, McGill University

Please signal the author by email if you find any typo.
These notes have not been revised and should be read carefully.

Contents

Foreword

The objectives of time series analysis is the **modeling** of sequences of random variables by time, as $X_1, X_2, \ldots, X_t$ based on data $x_1, \ldots, x_n$. Typically, $(X_1, \ldots, X_n)$ will not be mutually independent and n is large[1]. There is much greater interest in forecasting and prediction: given data $x_1, \ldots, x_n$, we wish to make statements about $X_{n+1}, X_{n+2}, \ldots$

Definition (Time series)

A time series model is a probabilistic representation describing $\{X_t\}$ via their joint distribution or the moments, in particular expectation, variance and covariance.

It will often not be possible to specify a joint distribution for the observations. Also, one will want to impose restrictions on moments, which we refer to as simplifying assumptions.

Note

1. X_t can be a discrete or continuous random variable
2. X_t could be vector-valued.
3. The time index will most typically be discrete and represent constant time spacing - it could also be necessary to consider continuous-time indexing.

[1] That is usually greater than 50, contrary to longitudinal statistics where the number of time observations is small

Chapter 1
Introduction

Section 1.1: Simple stationary models

The joint CDF of $(X_1, \ldots, X_n)$ gives one description of the probabilistic relationship between the random variables (RVs). Using the standard notation $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \mathsf{P}(X_1, \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$, this function fully specifies the joint probability model for the RVs. If X_t is vector-valued, denoted $\boldsymbol{X}_t = (X_{t_1}, \ldots, X_{t_k})^\top$ (namely a $k \times 1$ vector), then "$X_t \leq x_t$" is to be interpreted component wise: $X_{t_1} \leq x_{t_1}, X_{t_2} \leq x_{t_2}, \ldots, X_{t_k} \leq x_{t_k}$. A special case of this is mutual independence.

$$X_1, \ldots, X_n \textbf{ are mutually independent } \Leftrightarrow \mathsf{P}(X_1 \leq x_1, \ldots, X_n \leq x_n) = \prod_{t=1}^{n} \mathsf{P}(X_t \leq x_t),$$

that is the observed values of $X_1, X_2, \ldots, X_{t-1}$ contain no information for modeling (or predicting) X_t. Beyond independence, at the other extreme, specifying a completely general joint model requires the specification of one marginal distribution and $n-1$ different conditional models[2]. Often, modeling is restricted to specification of moments of the process. Most typically, attention focuses on the first two moments (expectation and variance/covariance).

The emphasis of the lecture will be on stationarity models, which we will deal through in the beginning of the course. The concept is one to do with stability. Some process we dealt with last class were stable in mean or in variance or exhibiting periodic structure.

Definition 1.1 (Stationarity)
A time series process is **stationary** if $\{X_t, t \in \mathbb{Z}\}$ and $\{X_{t+h}, t \in \mathbb{Z}\}$ for $h = \pm 1, \pm 2, \ldots,$ have the precisely the same statistical properties (either in terms of joint distribution or moment properties).

In practical terms, t is arbitrary and it doesn't matter where we start collecting, we will always be able to dip in the stationary model. For most of the course, we will be dealing with **weak stationarity**, or moment stationarity, as joint distribution is intractable.

Definition 1.2 (Mean and covariance function)
Let $\{X_t\}$ be a time series process with $\mathsf{E}\left(X_t^2\right) < \infty \; \forall \; t$. Then $\mu_X(t) = \mathsf{E}\left(X_t\right)$ is the **mean** function and $\tilde{\gamma}_X(t, s) = \mathsf{Cov}(X_t, X_s) = \mathsf{E}\left((X_t - \mu_X(t))(X_s - \mu_X(s))\right)$ for integers t, s is the **covariance function**.

[2]This procedure can be intractable (if one is interested in a specific period and the implications of the model for that given interval) and may not be feasible analytically.

Definition 1.3 (Weak stationarity)
We say $\{X_t\}$ is **weakly stationary** if

1. $\mu_X(t)$ does not depend on t;
2. $\tilde{\gamma}_X(t+h,t)$ does not depend on t for all integers h.

A stronger form of stationarity imposes conditions on the form of the joint distribution of the process.

Definition 1.4 (Strong stationarity)
$\{X_t\}$ is **strongly stationarity** if $(X_{t+1},\ldots,X_{t+n})$ and $(X_{t+h+1},\ldots,X_{t+h+n})$. have the same joint distribution for all t,h,n.

Clearly, the second inequality implies the second, although it does not guarantee that the first two moments exist. Gaussian process is an example of the latter stronger conditions, with multivariate and a specific structure for the covariance matrix and the mean vector. For most part, we will be dealing with weak stationarity.

Definition 1.5 (Autocovariance function)
For a weakly stationary process, define the **autocovariance function** (ACVF) by

$$\gamma_X(h) \equiv \tilde{\gamma}_X(h,0) \equiv \tilde{\gamma}_X(t+h,t)$$

and the **autocorrelation function** (or ACF) by $\rho_X(h) = \gamma_X(h)/\gamma_X(0)$ for $h \in \mathbb{Z}$. [3]

So far, we have defined only properties of the process, not the data. For the moment, these are nonparametrically specified functions. We have by definition that $\gamma_X(h) = \mathsf{Cov}(X_{t+h},X_t)$ and $\gamma_X(0) = \mathsf{Cov}(X_t,X_t) = \mathsf{Var}(X_t)$ and $\rho_X(h) = \mathsf{Cor}(X_{t+h},X_t)$ for all t,h.

Some examples to illustrate the construction of these functions.

Example 1.1 (IID process)
Suppose we have an IID process where we suppose $\{X_t\}$ is a binary process, i.e. $X_t \in \{-1,1\}$ such that $\mathsf{P}(X_t=1)=p$ and $\mathsf{P}(X_t=-1)=1-p$ such that the X_t are mutually independent. We can easily compute the moments, as

$$\mathsf{E}(X_t) = p + (1-p)(-1) = 2p-1$$

and

$$\mathsf{Var}(X_t) = [p(1)^2 + (1-p)(-1)^2] - (2p-1)^2 = 1-(1-2p)^2.$$

By independence, $\mathsf{Cov}(X_t,X_s)=0$ if $t \neq s$.

[3] This function of h will be bounded on $[-1,1]$.

Example 1.2 (Random walk)

A more interesting process is the random walk, where $\{X_t\}$ is the same as in the previous example with the IID process and $\{S_t\}$ is a process constructed from $\{X_t\}$, defined by $S_t = \sum_{i=1}^{t} X_i = S_{t-1} + X_t$ assuming that $t \geq 1$ and the process start at 0 (that is $S_0 = 0$). Although the process is easy to construct, it has some interesting properties. If we take the simple case where $p = 1/2$, then $\mathsf{E}(X_t) = 0$ and $\mathsf{Var}(X_t) = 1$ and therefore $\mathsf{E}(S_t) = 0$, but $\mathsf{Var}(S_t) = t$. Since the variance of sum of IID random variable is the sum of the variance of each elements, so that the variance increase linearly with t and our process is not stationarity.

If $p \neq 1/2$, then $\mathsf{E}(X_t) = 2p - 1$ and $\mathsf{E}(S_t) = (2p-1)t$ and S_t also grows linearly with t and $\mathsf{Var}(X_t) = 1 - (1-2p)^2$ and $\mathsf{Var}(S_t) = t\mathsf{Var}(X_t)$. In this case, if $p < 1/2$ we have a downward "drift", while if $p > 1/2$, we have an upward drift. In all cases, the variance of S_t grows linearly with t. As such, $\{S_t\}$ is nonstationary (clearly).

Note

We could also define X_t as the difference of the random variable S_t, that is $X_t = S_t - S_{t-1}$. This gives a key insight on how to turn a nonstationary process into a stationary one, in this case by differencing.

Example 1.3 (Stationary Gaussian process)

Suppose, for all finite collections $X_1, \ldots, X_n$, the joint distribution is multivariate Gaussian (Normal) with $\mathsf{E}(X_t) = \mu_X$ (not depending on t) and covariance defined for X_t, X_s as $\mathsf{Cov}(X_t, X_s) = \gamma_X(|t-s|)$ for $t, s \in \{1, \ldots, n\}$.

This stationary version imposes extra conditions, which corresponds to a structured covariance matrix which has fewer than $\frac{1}{2}n(n+1)$ different elements.

Let $\mathbf{\Gamma}_X(n)$ denote the $(n \times n)$ matrix with $[\mathbf{\Gamma}_X(n)]_{t,s} = \gamma_X(|t-s|)$. Then $\mathbf{\Gamma}_X(n)$ is a symmetric positive definite [4] matrix ($\forall\, x \in \mathbb{R}^n, x^\top \mathbf{\Gamma}_X(n) x > 0$) with **Toeplitz** structure, constant among diagonals,

$$\mathbf{\Gamma}_X(n) = \begin{bmatrix} \gamma_X(0) & \gamma_X(1) & \cdots & \cdots & \gamma_X(n-1) \\ \gamma_X(1) & \gamma_X(0) & \gamma_X(1) & \cdots & \vdots \\ \gamma_X(2) & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \gamma_X(1) \\ \gamma_X(n-1) & \cdots & \cdots & \gamma_X(1) & \gamma_X(0) \end{bmatrix}$$

unlike the usual settings, with n free parameters in $\mathbf{\Gamma}_X(n)$. We write the vector $\boldsymbol{X} = (X_1, \ldots, X_n)^\top \sim \mathcal{N}_n(\mu_X \mathbf{1}_n, \mathbf{\Gamma}_X(n))$.[5] Note that all marginal distributions and all condi-

[4] In fact, is non-negative definite, but we will restrict our examples to positive definite matrix

[5] Again, these are parameters of the process, not any estimates derived from the data.

tional distributions are also multivariate Gaussian. Specifically, the one that we might be most interested in is $p(X_t|X_1, \ldots, X_{t-1})$, which is univariate Gaussian.

Exercise 1.1

Review the properties of the multivariate Normal and make sure you understand the last statement and are able to derive it.

The next few examples use similar construction methods, but are more general.

Example 1.4 (General IID noise)

Let $\{X_t\} \sim \text{IID}(0, \sigma^2)$, so that $\mathsf{E}(X_t) = 0$ and $\mathsf{E}\left(X_t^2\right) = \sigma^2 < \infty$ with all X_t mutually independent. From that, we get that the autocovariance function

$$\gamma_X(h) = \tilde{\gamma}_X(t+h, t) = \begin{cases} \sigma^2 & \text{if } h = 0 \\ 0 & \text{if } h \neq 0. \end{cases}$$

and

$$\rho_X(h) = \begin{cases} 1 & \text{if } h = 0 \\ 0 & \text{if } h \neq 0. \end{cases}$$

We now have a more arbitrary and general stationary process.

Example 1.5 (White noise)

Let $\{X_t\} \sim \text{WN}(0, \sigma^2)$. The white noise process has $\mathsf{E}(X_t) = 0$ and $\mathsf{E}\left(X_t^2\right) = \sigma^2 < \infty$ with X_t's uncorrelated[6]. The autocovariance function is still

$$\gamma_X(h) = \tilde{\gamma}_X(t+h, t) = \begin{cases} \sigma^2 & \text{if } h = 0 \\ 0 & \text{if } h \neq 0. \end{cases}$$

For most of what we will be doing in the course, we will restrict to white noise. However, if we wanted to estimate the model using maximum likelihood, we would need additional requirements (IID).

Example 1.6 (General random walk)

Let $\{X_t\} \sim \text{IID}(0, \sigma^2)$ and again $\{S_t\}$ is defined by $S_t = \sum_{i=1}^t X_i$. Using linear properties of the expectation, $\mathsf{E}(X_t) = 0$ implies that $\mathsf{E}(S_t) = 0$ and $\mathsf{E}\left(X_t^2\right) = \sigma^2$ and $\mathsf{E}\left(S_t^2\right) = t\sigma^2$ as $S_t^2 = \left(\sum_{i=1}^t X_i\right)^2 = \sum_{i=1}^t X_i^2 + 2\sum_{i=2}^t \sum_{j=1}^{i-1} X_i X_j$. The first term of the summation has expectation $t\sigma^2$ and the second term has expectation zero from independence. Again, $\{S_t\}$ is nonstationary.

[6] Independence would imply uncorrelatedness, but not conversely. This allows us to relax our assumptions, as the only calculations we are required to perform are moments so uncorrelatedness is really a moment requirement.

We also have

$$\begin{aligned}\tilde{\gamma}_X(t+h,t) &= \mathsf{Cov}(S_{t+h}, S_t)\\ &= \mathsf{Cov}(S_t + X_{t+1} + \cdots + X_{t+h}, S_t)\\ &= \mathsf{Cov}(S_t, S_t) + \mathsf{Cov}(X_{t+1} + \cdots X_{t+h}, S_t)\\ &= \mathsf{Cov}(S_t, S_t) + \sum_{i=1}^{h} \mathsf{Cov}(X_{t+i}, S_t)\\ &= \mathsf{Var}(S_t) + 0\\ &= t\sigma^2.\end{aligned}$$

and the function depends on t, but not on h. This calculation uses result for covariances which says that $\mathsf{Cov}(aX + bY + c, Z) = a\mathsf{Cov}(X, Z) + b\mathsf{Cov}(Y, Z)$.

Example 1.7 (Moving average process)
Let $\{X_t\} \sim \text{MA}(1)$ and suppose $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ where $\sigma^2 < \infty$. Let $\theta \in R$ and define

$$X_t = Z_t + \theta Z_{t-1}, \quad t \in \mathbb{Z} \tag{1.1}$$

We can easily verify stationarity by calculating the first two moments. By linearity

$$\mathsf{E}(X_t) = \mathsf{E}(Z_t + \theta Z_{t-1}) = \mathsf{E}(Z_t) + \theta \mathsf{E}(Z_{t-1}) = 0$$

We want to verify that $\tilde{\gamma}_X(t+h, t)$ doesn't depend on t

$$\mathsf{Cov}(X_{t+h}, X_t) = \mathsf{E}(X_{t+h} X_t)$$

Now

$$\begin{aligned}X_{t+h}X_t &= (Z_{t+h} + \theta Z_{t+h-1})(Z_t + \theta Z_{t-1})\\ &= Z_{t+h}Z_t + \theta(Z_{t+h}Z_{t-1} + Z_{t+h-1}Z_t) + \theta^2 Z_{t+h-1}Z_{t-1}.\end{aligned}$$

Since Z_t is a white-noise process, $\mathsf{E}(Z_r Z_s) = \sigma^2$ if $r = s$ and zero otherwise (if $r \neq s$). Then

$$\mathsf{E}(Z_{t+h}Z_t) = \begin{cases}\sigma^2 & \text{if } h = 0\\ 0 & \text{if } h \neq 0\end{cases} \qquad \mathsf{E}(Z_{t+h}Z_{t-1}) = \begin{cases}\sigma^2 & \text{if } h = -1\\ 0 & \text{if } h \neq -1\end{cases}$$

and

$$\mathsf{E}\left(Z_{t+h-1}, Z_t\right) = \begin{cases} \sigma^2 & \text{if } h = 1 \\ 0 & \text{if } h \neq 1 \end{cases} \qquad \mathsf{E}\left(Z_{t+h-1}, Z_{t-1}\right) = \begin{cases} \sigma^2 & \text{if } h = 0 \\ 0 & \text{if } h \neq 0. \end{cases}$$

Combining the above results, we obtain

$$\tilde{\gamma}_X(t+h,t) = \begin{cases} \sigma^2(1+\theta^2) & \text{if } h = 0 \\ \sigma^2\theta & \text{if } h = \pm 1 \\ 0 & \text{otherwise .} \end{cases}$$

Since $\tilde{\gamma}_X$ does not depend on t, this imply $\{X_t\}$ is stationary, $X_t \sim \text{MA}(1)$ with $\mathsf{Var}\left(X_t\right) = \tilde{\gamma}_X(t,t) = \sigma^2(1+\theta^2)$ with autocorrelation function

$$\rho_X(h) = \begin{cases} 1 & \text{if } h = 0 \\ \theta/(1+\theta^2) & \text{if } h = \pm 1 \\ 0 & \text{otherwise .} \end{cases}$$

We will see that any process defined as a linear combination of white noise process for arbitrary lag will be a stationary process.

Example 1.8 (Autoregression)

Let $X_t \sim \text{AR}(1)$. Again, suppose $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, with $\sigma^2 < \infty$. Let $|\phi| < 1$ and[7] **assume** $\{X_t\}$ is a stationary process such that X_t and X_s are uncorrelated if $s < t$, with the following recursive definition

$$X_t = \phi X_{t-1} + Z_t, \quad t = 0, \pm 1, \pm 2, \ldots.$$

This can be rewritten $X_t - \phi X_{t-1} = Z_t$. [8] Then

$$\mathsf{E}\left(X_t\right) = \mathsf{E}\left(\phi X_{t-1}\right) + \mathsf{E}\left(Z_t\right) = \phi \mathsf{E}\left(X_{t-1}\right).$$

Under the assumption of stationarity, we must have $\mathsf{E}\left(X_t\right) = 0$ as this is the only solution for an arbitrary ϕ. Under stationarity, we can go directly for the autocorrelation function and

$$\gamma_X(h) = \mathsf{E}\left(X_{t+h} X_t\right) = \mathsf{E}\left(X_t X_{t-h}\right) = \gamma_X(-h)$$

[7] Otherwise, if $\phi = 1$, we would have the random walk model.

[8] It is not straightforward to derive stationarity with this definition, we will prove it later. We can for now do some calculations.

so that γ_X is an even function. We can take this form here and substitute in for X_t, yielding

$$\begin{aligned}\gamma_X(h) &= \mathsf{E}\left((\phi X_{t-1} + Z_t)X_{t-h}\right) \\ &= \phi\mathsf{E}\left(X_{t-1}X_{t-h}\right) + \mathsf{E}\left(Z_t X_{t-h}\right) \\ &= \phi\gamma_X(h-1) + 0\end{aligned}$$

where $\mathsf{E}\left(Z_t X_{t-h}\right)$ as $t > t-h$. Hence, using the above recursion, we can get $\gamma_X(h) = \phi\gamma_X(h-1) = \phi^h\gamma_X(0)$ for $h > 0$ and using symmetry, $\gamma_X(h) = \gamma_X(-h) = \phi^{|h|}\gamma_X(0)$.

We now want to find $\gamma_X(0)$, equal to

$$\begin{aligned}\mathsf{Cov}(X_t, X_t) &= \mathsf{E}\left(X_t^2\right) \\ &= \mathsf{E}\left((\phi X_{t-1} + Z_t)(\phi X_{t-1} + Z_t)\right) \\ &= \phi^2\mathsf{E}\left(X_{t-1}^2\right) + 2\phi\mathsf{E}\left(X_{t-1}Z_t\right) + \mathsf{E}\left(Z_t^2\right).\end{aligned}$$

But, by assumption $\mathsf{E}\left(X_{t-1}Z_t\right) = 0$ (as current X are uncorrelated with future values of Z) and $\mathsf{E}\left(X_{t-1}^2\right) = \gamma_X(0)$ as $\{X_t\}$ is stationary. Therefore, $\gamma_X(0) = \phi^2\gamma_X(0) + \sigma^2$. This implicit formula can be rearranged to get and explicit form for $\gamma_X(0) = \sigma^2/(1-\phi^2)$.

It is possible to extend the two processes described above, thus MA(1) can be generalized to MA(q), of the form

$$X_t = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \cdots + \theta_q Z_{t-q} \tag{1.2}$$

and AR(1) to AR(p) as a regression on the previous p values of X_t as

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t \tag{1.3}$$

and also combine the two models. We would end up modeling

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t + \theta_1 Z_{t-1} + \cdots \theta_q Z_{t-q}, \tag{1.4}$$

which is the so called ARMA(p, q), or **autoregressive moving-average process**. Most of the models we will play will be ARMA-type models.

Generalization for MA is far more trivial than AR, since MA processes are very stable (we can take linear combinations of MA processes (with finite coefficients) which are stable as well), whereas for AR(p) it is not straightforward to extend them: verification is necessary. Since AR is close to the random walk, we need to think carefully which values for $\phi_i, i = 1, \ldots, p$ to choose to get stationarity. We will study these later.

Section 1.2: Trends and Seasonality

1.2.1. Trends

For many datasets, it is evident that the observed series does not arise from a process that is stationary or constant in mean. We might observe that

- perhaps a monotonic trend is present
- some other deterministic variation in mean is observed

We might want to decompose the process into a part that has a deterministic mean structure and on top of that have a stable or stationary component.

Thus, a model of the form $X_t = m_t + Y_t$ may be appropriate, where m_t is a deterministic function of t and $\{Y_t\}$ is a zero-mean (and often stationary) process. If $\{Y_t\}$ is assumed to have finite variance, and m_t is assumed to have some parametric form, say $m_t(\beta)$, then β may be estimated using least-squares. One could also use nonparametric or semiparametric regression methods to estimate the model.

If, say, $\{Y_t\} \sim \text{WN}(0, \sigma^2)$ is assumed then, we can determine the parameter estimate for β by minimizing the sum of squared residuals errors, that is

$$\widehat{\beta} = \arg\min_{\beta} \sum_{t=1}^{n} (x_t - m_t(\beta))^2;$$

this is linear (or nonlinear) regression with homoskedastic errors. Optimization is tractable analytically, (it is just convex optimization). In the easy case, we could have *e.g.* $m_t(\beta) = \beta_0 + \beta_1 t + \beta_2 t^2$ and $\boldsymbol{\beta} = (\beta_0 \;\; \beta_1 \;\; \beta_2)^\top$ can be estimated using ordinary least squares (OLS). If there was a correlation structure (such as AR) and the assumption of white noise did not hold, then OLS would not be statistically efficient and weighted least squares would be preferable.

Example 1.9 (Lake Huron data, example 135 B&D)
In this example X_t is the level (in feet) of Lake Huron from 1875-1972, with $n = 198$. In the illustration, a fitted linear trend is presented. One could also try to test for structural break around observation 20 and fit a model with two plateaus. Assuming $m_t = \beta_0 + \beta_1 t$, fitting this model is straightforward; in R, the lm function can be used. What we get is $\widehat{\beta}_0 = 10.204(0.230)$ and $\widehat{\beta}_1 = -0.024(0.004)$ with standard errors indicated in parenthesis. The slope is statistically significant.

We can verify some of the assumptions that we made about $\{Y_t\}$. Let $\widehat{y}_t = x_t - m_t(\widehat{\beta}) = x_t - \widehat{\beta}_0 - \widehat{\beta}_1 t$. Detrended data, that is $\{\widehat{y}_t\}$ looks like a zero mean process - this is a standard

residuals plots. We might suspect an increase in variance, but the assumptions are not too far off. We also need to verify that Y_t are uncorrelated, but it turns out that they are not; a scatter plot of $(\widehat{y}_{t-1}, \widehat{y}_t)$, for $t = 2, \ldots, n$ shows that the series exhibit positive correlation (0.775). In fact, Y_t is not likely a white-noise process.

1.2.2. Seasonality

Many (real) time series are influenced by seasonal factors, which may be due to climate, calendar, economic cycles or physical factors. In this section, we will consider only seasonality and not consider trend. This that we are studying is a deterministic seasonality. One model for "seasonal" behaviour is

$$X_t = s_t + Y_t, \qquad t \in \mathbb{Z},$$

but where $\{s_t\}$ is **periodic** with period d, that is $s_{t-d} = s_t \; \forall \, t$.[9]

An example is the "harmonic model", where $s_t = s_t(\beta)$ is modeled in a parametric fashion, that is s_t is of the form

$$s_t = \beta_0 + \sum_{j=1}^{k} [\beta_{1j} \cos(\lambda_j t) + \beta_{2j} \sin(\lambda_j t)]$$

where $\beta_0, \beta_{11}, \beta_{21}, \beta_{12}, \beta_{22}, \ldots$ are unknown constants, but $\lambda_1 < \lambda_2 < \cdots < \lambda_k$ are fixed constants (known).[10]

Example 1.10 (Accidental Deaths in the US: Jan 1973-Dec 1978)
We have $n = 72$ monthly observations. Choose $k = 2$ and set $\lambda_1 = 2\pi/12$, which gives a 12 monthly cycle and $\lambda_2 = 2\pi/6$ which is a 6 monthly cycle. [11]

The next test is to assess whether our assumptions where correct. We will

- estimate $\beta_0, \beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}$ using OLS
- form $\{\widehat{Y}_t\}$ series as $\widehat{y}_t = x_t - s_t(\widehat{\beta})$.
- assess WN$(0, \sigma^2)$ assumption about $\{Y_t\}$
- in this analysis, residual series $\{\widehat{y}_t\}$ is positively uncorrelated.

[9] Again we can think of this as a decomposition into a deterministic part and a zero mean process Y_t which is the random component.

[10] We treat them as known constant as not to have to estimate the λ_i, so as to be able to fit a linear model. λ are chosen so that the periodicity matches the observed data.

[11] This is easy to implement in R, building a $n \times 5$ design matrix and perform OLS, directly using the lm function. One can make some model selection and perform goodness of fit test, nested structure and F tests to evaluate the models.

- try to use AR, MA models for $\{Y_t\}$.

Before looking at the data, one doesn't know what the periodicities are and the λ are unknown. If they are estimated, one has to do NLS, which can be easily implemented as well. Another example is Montreal gas station prices, which tend to peak on Tuesday, and exhibit a periodic structure if you demean the data. Generally, to carry out assessment of model assumptions concerning a stationary process, we may examine standard moment-based estimators. We are looking for estimators of $\widehat{\mu} = \bar{x}$ and for the ACVF, we use

$$\widehat{\gamma}(h) = \frac{1}{n} \sum_{t=1}^{n-|h|} (x_{t+|h|} - \bar{x})(x_t - \bar{x}), \qquad -n < h < n.$$

$\widehat{\gamma}(h)$ is a **consistent** estimator of $\gamma(h)$, but it is **biased**. The advantage of using this estimator is that it does, however, guarantee that $\widehat{\Gamma}_n = [\widehat{\gamma}(|i-j|)]_{ij}$ is non-singular.

To assess the validity of the WN$(0, \sigma^2)$ assumption, we compute $\widehat{\gamma}(h)$ and examine whether $\widehat{\gamma}(h)$ is "significantly different" from zero when $h \neq 0$.

Note

Note that this is a "nonparametric" approach. Parametric approaches are also possible in some cases. [12]

Example 1.11 (Lake Huron data)

Recall the model we imposed on this data, which was of the form $X_t = m_t(\beta) + Y_t = \beta_0 + \beta_1 t + Y_t$. The fit yielded residuals $\{\widehat{y}_t\}$, where $\widehat{y}_t = x_t - \widehat{\beta}_0 - \widehat{\beta}_1 t$. We know that $\mathsf{Cor}(\widehat{y}_{t-1}, \widehat{y}_t) \approx 0.77$. We can try an AR(1) model, with $Y_t = \phi Y_{t-1} + Z_t$, where $\{Z_t\} \sim$ WN$(0, \sigma^2)$ (recall that $\{Y_t\}$ is zero mean process). ϕ and σ^2 can be estimated from $\{\widehat{y}_t\}$ series using OLS, regressing $\widehat{y}_t$ on $\widehat{y}_{t-1}$. This does not impose the stationarity assumption on $\{Y_t\}$. [13] Using the lm function in R returns $\widehat{\phi} = 0.791$. We can now go back to the proposed model and compute $\{\widehat{z}_t\}$ as

$$\widehat{z}_t = \widehat{y}_t - \widehat{\phi}\widehat{y}_{t-1}.$$

Is $\widehat{z}_t$ an uncorrelated sequence? Apparently not, there is still positive correlation between $\widehat{z}_t$ and $\widehat{z}_{t-1}$. This means that the AR(1) model is not adequate. At this stage, we could work with this residual or go back to the model and propose a more complicated model. We try an AR(2) model, *i.e.*

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + Z_t.$$

The fit via OLS yields $\widehat{\phi}_1 = 1.002$ and $\widehat{\phi}_2 = -0.283$. In this case, it is hard to see whether

[12] This function would be completely specified as a function of the parameters of the process, while here we impose no relation for $\gamma(h)$.

[13] We have assumed until now that our AR(1) model be stationary. Since this is 1 period lagged model, this corresponds to $|\phi| < 1$.

these values imply stationarity. The resulting residual series $\{\widehat{z}_t\}$ formed by taking

$$\widehat{z}_t = \widehat{y}_t - \widehat{\phi}_1 \widehat{y}_{t-1} - \widehat{\phi}_2 \widehat{y}_{t-2}$$

appears to be possibly WN$(0, \sigma^2)$.

The **general strategy** when faced with a real time-series model: for observed series $\{X_t\}$ with realization $\{x_t\}$,

- write $X_t = m_t + s_t + Y_t$, a combination of deterministic trend and seasonality components.
- estimate m_t, s_t using parametric (or possibly semi-parametric or nonparametric models)
- form the residual series $\{\widehat{y}_t\}$ as $\widehat{y}_t = x_t - \widehat{m}_t - \widehat{s}_t$
- check/model properties of $\{\widehat{y}_t\}$ and check whether they are IID/WN, or AR, or MA, etc.

1.2.3. Non-parametric trend removal

A non-parametric smoothing approach is constructed as follows:

a) Set

$$\widehat{m}_t = \frac{1}{2q-1} \sum_{j=-q}^{q} x_{t+j}, \quad q+1 \leq t \leq n-q$$

taking a local unweighted average of the points in the vicinity of t. This is a form of "low pass filtering". An extension of this is to enlarge the window centered on t, equivalent to the above by setting $a_j = 0$ if $|j| > q$, given by

$$\widehat{m}_t = \sum_{=-\infty}^{\infty} a_j x_{t-j}$$

b) "Exponential" smoothing, using $\alpha \in (0, 1)$ as a smoothing parameter

$$\widehat{m}_t = \begin{cases} \alpha x_t + (1-\alpha)\widehat{m}_{t-1} & \text{if } t \geq 2 \\ x_1 & \text{if } t = 1 \end{cases}$$

and it is evident that the closer to 0, the smoother the curve. An explicit form [14] for $\widehat{m}_t$ is

$$\widehat{m}_t = \sum_{j=0}^{t-2} \alpha(1-\alpha)^j x_{t-j} + (1-\alpha)^{t-1} x_1$$

In the presence of both trend and seasonality, the procedures for pre-processing must be applied as follows. [15]

If $X_t = m_t + s_t + Y_t$ with $\mathsf{E}(Y_t) = 0$, $s_{t+d} = s_t$, and $\sum_{j=1}^{d} s_j = 0$. The steps are outline next[16]

Step 1
- d is even: set $q = d/2$ and take

$$\widehat{m}_t = \left(\frac{x_{t-q} + x_{t+q}}{2d}\right) + \frac{1}{d} \sum_{j=-(q-1)}^{q-1} x_{t+j}$$

- d is odd: set $q = (d-1)/2$ and take

$$\widehat{m}_t = \frac{1}{d} \sum_{j=-q}^{q} x_{t+j}$$

This removes the trend whilst respecting seasonality.

Step 2 Set $w_k = \frac{1}{n_k} \sum_j (x_{k+jd} - \widehat{m}_{k+jd})$ for $k = 1, 2, \ldots, d$, where the sum extends over j such that $q < k + jd < n - q$ and n_k is the number of terms in the sum. Then

$$\widehat{s}_k = w_k - \bar{w}_k = w_k - \frac{1}{d} \sum_{i=1}^{d} w_k$$

for $k = 1, 2, \ldots, d$ and $\widehat{s}_t = \widehat{s}_{t-d}$ for $t > d$.[17]

Step 3 Remove the remaining trend: compute $\widehat{m}_t^*$ from the de-seasonalized data $x_t^* = x_t - \widehat{m}_t - \widehat{s}_t$ using a parametric or a non-parametric approach. This yields the decomposition $X_t = \widehat{m}_t^* + \widehat{s}_t + Y_t$

[14] This is a preprocessing set, but be careful as you can remove the structure of the data through the smoothing method.

[15] The above averaging does not respect the seasonality, especially if you take terms from the previous cycle and you will distort the estimate of $\widehat{m}_t$.

[16] if this is non-zero, without loss of generality we can include it in the m_t component.

[17] Everything we have done in this analysis is non-parametric. We have removed the seasonality with that step.

1.2.4. Differencing

Definition 1.6 (Differencing at lag 1)
The lag difference operator,

$$\nabla X_t = X_t - X_{t-1} = X_t - BX_t = (1-B)X_t$$

where B is the backshift operator.

Definition 1.7 (Seasonal differencing)
The lag d difference operator, ∇_d, is defined by

$$\nabla_d X_t = X_t - X_{t-d} = X_t - B^d X_t = (1-B^d)X_t$$

Note (Elementary algebraic properties)

1. $B^j X_t = X_{t-j} = BX_{t-j+1} = B^2 X_{t-j+2}$, *etc.*
2. $\nabla^j X_t = \nabla(\nabla^{j-1} X_t)$. So for example,

$$\begin{aligned}\nabla^2 X_t &= (1-B)(1-B)X_t \\ &= (1-2B+B^2)X_t \\ &= X_t - 2X_{t-1} + X_{t-2} \\ &= (X_t - X_{t-1}) - (X_{t-1} - X_{t-2})\end{aligned}$$

3. If $X_t = m_t + Y_t$, applying the first-difference operator to X_t, then

$$\nabla X_t = X_t - X_{t-1} = (m_t - m_{t-1}) + (Y_t - Y_{t-1})$$

   so if $m_t = m_t(\boldsymbol{\beta}) = \beta_0 + \beta_1 t$, then $m_t - m_{t-1} = \beta_1$, *i.e.* ∇ removes a linear trend. [18] To remove a polynomial trend of order k, one may apply k^{th} order differencing, that is we look at $\nabla^k X_t$. Note that if $\{Y_t\} \sim \text{WN}(0, \sigma^2)$, then $\nabla^k Y_t$ is not white-noise, but still stationary.

4. $\nabla\nabla^d X_t = (1-B)(1-B^d)X_t = (1-B^d)(1-B)X_t = \nabla^d \nabla X_t$.
5. ∇^d removes a seasonality with period d. If $X_t = m_t + s_t + Y_t$, then

$$\nabla^d X_t - (m_t - m_{t-d}) + (s_t - s_{t-d}) + (Y_t - Y_{t-d})$$

   and $s_t - s_{t-d} = 0$.

[18] $Y_t^* = (Y_t - Y_{t-1})$ is what is obtained. This **is different** from our first detrending approach, which used a linear parametric formulation for m_t. If Y_t is white-noise, then Y_t^* is no longer white noise, but MA(1). The nature of the Y_t process will change.

1.2.5. Assessing the white noise assumption

Some tests that are commonly used. If $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, then as $n \to \infty$,

$$\widehat{\rho}_n(h) \dot{\sim} \mathcal{N}\left(0, \frac{1}{n}\right), \quad \forall\, h \geq 1$$

where $\widehat{\rho}_n(h)$ is the estimator of $\rho_X(h)$ derived from data $Z_1, \ldots, Z_n$. This result is derived from Central limit theorem.

This allows us to carry out pointwise (in h) hypothesis tests of the form [19]

$$H_0 : \rho_X(h) = 0, \quad h \geq 1$$

Also, $(\widehat{\rho}_n(1), \ldots, \widehat{\rho}_n(l))$ are asymptotically independent under the white noise assumption. We can test $\widehat{\rho}_n(h), h = 1, \ldots, l$ individually or jointly (*i.e.* simultaneously testing: should make correction in the critical values in order to account for multiplicity). [20] An approximate 95% CI for $\rho_X(h)$ is

$$\widehat{\rho}_n(h) \pm \frac{1.96}{\sqrt{n}}$$

There are other tests, which also rely on asymptotics. They are presented next.

Proposition 1.8 (Portmanteau test)
The test statistic is $Q = n \sum_{j=1}^{h} \{\widehat{\rho}_n(j)\}^2$. Under $H_0 : \rho_X(j) = 0$ for $j = 1, \ldots, h$ and $Q \sim \chi^2_h$.[21]

Proposition 1.9 (Box-Ljung test)
This test attempts finite-sample bias correction

$$Q_{\text{BL}} = n(n+2) \sum_{j=1}^{h} \frac{\{\widehat{\rho}_n(j)\}^2}{n-j}$$

and under $H_0, Q_{\text{BL}} \dot{\sim} \chi^2(h)$

[19] In R, we have seen line for pointwise critical values for the test or simultaneous tests for more than one h.

[20] It would also be possible (maybe not in practice) to perform a permutation test for the white-noise case: if the data is uncorrelated, it will not be impacted if we permute it. One can then recompute the test statistic for these $n!$ permutations. This yield a discrete distribution on $n!$ elements. One can compare the true observed value from the sample with the distribution from the permutation.

[21] For most of the cases we are dealing with, this is a powerful test for simultaneous testing.

Proposition 1.10 (McLeod-Li test)

Let $W_t = Z_t^2$, then compute $\widehat{\rho}_n^W(h)$ for $\{w_t\}$. The test statistic is given by

$$Q_{\text{ML}} = n(n+2)\sum_{j=1}^{h}\frac{\{\widehat{\rho}_n^W(j)\}^2}{n-j}$$

and again under $H_0, Q_{\text{ML}} \dot{\sim} \chi^2_h$.[22]

1.2.6. Some general results and representations for stationary processes

1. **Stationarity and autocovariance**

Definition 1.11 (Non-negative definite)

A real-valued function $g : \mathbb{Z} \to \mathbb{R}$ is non-negative definite if, for all $n \geq 1$, if $\boldsymbol{a} = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n$, then

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i g(i-j) a_j \geq 0$$

Theorem 1.12

A function g defined on $\mathbb{Z}$ is the autocovariance function (acvf) of a stationary time series process if and only if g is **even** and **non-negative definite.**

Proof Suppose g is the acvf of a stationary process $\{X_t\}$. Then g is even by properties of covariance. Let $\boldsymbol{a} = (a_1, \ldots, a_n)^\top$ and $\boldsymbol{X}_{1:n} = (X_1, \ldots, X_n)^\top$, and let $\boldsymbol{\Gamma}_X(n)$ be the $n \times n$ matrix with $(i,j)^{\text{th}}$ element $\gamma_X(i-j)$ for $i, j = 1, \ldots, n$ and $n \geq 1$. By definition $g(i-j) \equiv \gamma_X(i-j)$. Then

$$\mathsf{Var}\left(\boldsymbol{a}^\top \boldsymbol{X}_{1:n}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \gamma_X(i-j) a_j.$$

But $\mathsf{Var}\left(a^\top \boldsymbol{X}_{1:n}\right) \geq 0$ by properties of variance. Thus g is non-negative definite.

Conversely, if g is even and non-negative definite, then $\boldsymbol{\Gamma}_X(n)$ formed by setting the $(i,j)^{\text{th}}$ element to $g(i-j)$ is symmetric and non-negative definite[23]. So consider $\{X_t\}$ the Gaussian process such that $\boldsymbol{X}_{i:n} \sim \mathcal{N}_n(0, \boldsymbol{\Gamma}_X(n))$. ■

[22] Notice that $\widehat{\mathsf{Cor}}(Z_t^2)$ has the same asymptotic distribution as $\widehat{\mathsf{Cor}}(W_t)$ and that the squaring insure that we sum non-negative correlation coefficient.

[23] By defining Γ in this way, this matrix has a Toeplitz structure, i.e. constant along the diagonal.

2. **Construction of stationary processes**
   If $\{Z_t\}$ is stationary, then the "filtered" process $\{X_t\}$ defined y

$$X_t = g(Z_t, Z_{t-1}, \ldots, Z_{t-q})$$

   for $q \geq 0$ for some function g is also stationary with X_t and X_s for $|t-s| > q$ uncorrelated if $\{Z_t\}$ is white-noise. $\{X_t\}$ is termed "q-dependent" in this case. [24]

3. **Linear processes**
   $\{X_t\}$ is a **linear process** if, for all t,

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$$

   where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, $\{\psi_j\}$ is a sequence of real constants such that $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. That is $X_t = \Psi(B)Z_t$ with $\Psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j$, an infinite order polynomial (generating function).

   A linear process is "non-anticipating" or "causal" if $\psi_j = 0 \; \forall \; j < 0$.

   Note

   The condition $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$ ensures that $\sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$ is well-defined, that is, that the sum converges in mean-square, *i.e.* in fact $X_t \overset{\text{M.S.}}{=} \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$ that is $\forall \; t$,

$$\lim_{n \to \infty} \mathsf{E}\left(\left(X_t - \sum_{j=-n}^{n} \psi_j Z_{t-j} \right)^2 \right) = 0.$$

   In this calculation, $\psi(B)$ is a linear filter acting on $\{Z_t\}$. From this, we can compute the moment properties of $\{X_t\}$ directly.

Recall $X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$. Note that

$$\mathsf{E}\left(|Z_t|\right) = \mathsf{E}\left(\sqrt{|Z_t|^2}\right) = \sqrt{\mathsf{E}\left(Z_t^2\right)} = \sigma$$

[24] Filtering generally means all observations up to current time in such case.

and

$$\begin{aligned}
\mathsf{E}\left(|X_t|\right) &= \mathsf{E}\left(\left|\sum_{j=-\infty}^{\infty}\psi_j Z_{t-j}\right|\right) \\
&\leq \mathsf{E}\left(\sum_{j=-\infty}^{\infty}|\psi_j||Z_{t-j}|\right) \\
&= \sum_{j=-\infty}^{\infty}|\psi_j|\mathsf{E}|Z_{t-j}| \\
&\leq \sigma\sum_{j=-\infty}^{\infty}|\psi_j| < \infty.
\end{aligned}$$

Now $\mathsf{E}\left(X_t\right) = 0$ and

$$\begin{aligned}
\mathsf{Var}\left(X_t\right) = \mathsf{E}\left(X_t^2\right) &= \mathsf{E}\left(\left(\sum_{j=-\infty}^{\infty}\psi_j Z_{t-j}\right)^2\right) \\
&= \mathsf{E}\left(\left(\sum_{j=-\infty}^{\infty}\psi_j Z_{t-j}\right)\left(\sum_{k=-\infty}^{\infty}\psi_k Z_{t-k}\right)\right) \\
&= \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\psi_j\psi_k\mathsf{E}\left(Z_{t-j}Z_{t-k}\right) \\
&= \sum_{j=-\infty}^{\infty}\psi_j^2\mathsf{E}\left(Z_{t-j}^2\right) \\
&= \sigma^2\sum_{j=-\infty}^{\infty}|\psi_j|^2 \\
&\leq \sigma^2\left(\sum_{j=-\infty}^{\infty}|\psi_j|\right)^2 < \infty.
\end{aligned}$$

We thus have a finite variance process and we have an expression for the variance of X_t in terms of the ψ_j. We can also, perhaps more importantly, compute the autocovariance, in exactly the same fashion as we did the previous calculation.

$$\begin{aligned}
\mathsf{E}\left(X_{t+h}X_t\right) &= \mathsf{E}\left(\left(\sum_{j=-\infty}^{\infty}\psi_j Z_{t+h-j}\right)\left(\sum_{k=-\infty}^{\infty}\psi_k Z_{t-k}\right)\right) \\
&= \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\psi_j\psi_k\mathsf{E}\left(Z_{t+h-j}Z_k\right) \\
&= \sum_{j=-\infty}^{\infty}\psi_j\psi_{j-h}\mathsf{E}\left(Z_{t+h-j}^2\right) \\
&= \sigma^2\sum_{j=-\infty}^{\infty}\psi_j\psi_{j+h} \equiv \gamma_X(h).
\end{aligned}$$

Note that if $\{Y_t\}$ is stationary with acvf $\gamma_Y(h)$, then if $X_t = \psi(B)Y_t = \sum_{j=-\infty}^{\infty}\psi_j Y_{t-j}$, we can replace $\{Z_t\}$ in the previous calculations to obtain results for $\mathsf{E}\left(X_t\right), \mathsf{Var}\left(X_t\right)$ and $\gamma_X(h)$. In particular, if $\mathsf{E}\left(Y_t\right) = 0$, then $\mathsf{E}\left(X_t\right) = 0$ and $\gamma_X(h)$ can be written as

$$\gamma_X(h) = \sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty}\psi_j\psi_k\gamma_Y(h-j+k).$$

Section 1.3: Autoregressive Time Series Processes

Recall the AR(1) process: let $\{Z_t\} \sim \text{WN}(0,\sigma^2)$ with $|\phi| < 1$ and $\{X_t\}$ is defined as the solution to $X_t = \phi X_{t-1} + Z_t$ with $\mathsf{E}\left(X_t Z_s\right) = 0$, $s > t$. That is $X_t - \phi X_{t-1} = Z_t$ or $(1-\phi B)X_t = Z_t$

By recursion,

$$\begin{aligned}
X_t &= \phi X_{t-1} + Z_t = \phi(\phi(X_{t-2} + Z_{t-1}) + Z_t \\
&= \phi^2 X_{t-2} + \phi Z_{t-1} + Z_t \\
\cdots &= \phi^n X_{t-n} + \sum_{j=0}^{n-1}\phi^j Z_{t-j} \\
\ldots &= \sum_{j=0}^{\infty}\phi^j Z_{t-j}
\end{aligned}$$

since the leading term vanishes if $\phi < 1$, $\phi^n \to 0$ as $n \to \infty$.

So we may write $X_t = \sum_{j=0}^{\infty}\psi_j Z_{t-j}$, where $\psi_j = \phi^j$. But $\sum_{j=0}^{\infty}|\phi^j| \leq \sum_{j=0}^{\infty}|\phi|^j < \infty$. We are thus using a linear filter with absolutely convergent coefficients series, and $\{X_t\}$ is

stationary, $\mathsf{E}(X_t) = 0$ and again by our earlier results,

$$\gamma_X(h) = \sum_{j=0}^{\infty} \psi_j \psi_{j+h} \mathsf{E}\left(Z_t^2\right) = \sigma^2 \sum_{j=0}^{\infty} \phi^j \phi^{j+h}$$

and so $\gamma_X(h) = \sigma^2 \phi^h \sum_{j=0}^{\infty} \phi^{2j} = \sigma^2 \phi^h / (1 - \phi^2)$.

Note

We may formally write that $\Psi(B) = (1 - \phi B)^{-1}$ as

$$\begin{aligned}
\Psi(B)(1 - \phi B) &= \left(\sum_{j=0}^{\infty} \psi_j B^j \right) (1 - \phi B) \\
&= \sum_{j=0}^{\infty} \psi_j B^j - \sum_{j=0}^{\infty} \phi \psi_j B^{j+1} \\
&= \sum_{j=0}^{\infty} \phi^j B^j - \sum_{j=1}^{\infty} \phi^j B^j = 1
\end{aligned}$$

so that $(1 - \phi B) X_t = Z_t \to X_t = (1 - \phi B)^{-1} Z_t$ and if we allow series expansion in B of this $(1 - \phi B)^{-1}$, then by definition $(1 - \phi B)^{-1} \equiv \sum_{j=0}^{\infty} \phi^j B^j$.

We have thus constructed a solution for the process, but still have to demonstrate its uniqueness. Now suppose that $\{Y_t\}$ is another stationary solution to the equation $Y_t = \phi Y_{t-1} + Z_t$. We show an explicit relationship between $\{X_t\}$ and $\{Y_t\}$. By the previous recursive approach, $Y_t = \sum_{j=0}^{\infty} \phi^j Z_{t-j}$. Then

$$\begin{aligned}
Y_t - \sum_{j=0}^{k} \phi^j Z_{t-j} &= \sum_{j=k+1}^{\infty} \phi^j Z_{t-j} \\
&= \phi^{k+1} \sum_{j=0}^{\infty} \phi^j Z_{t-j-k-1} \\
&= \phi^{k+1} Y_{t-k-1}
\end{aligned}$$

and

$$\lim_{k \to \infty} \mathsf{E}\left(\left(Y_t - \sum_{j=0}^{k} \phi^j Z_{t-j} \right)^2 \right) = \lim_{k \to \infty} \phi^{2k+2} \mathsf{E}\left(Y_{t-k-1}^2\right) = 0$$

as $\mathsf{E}\left(Y_{t-k-1}^2\right) < \infty$ by assumption. Therefore, we have $Y_t \overset{\text{MS}}{=} \sum_{j=0}^{\infty} \phi^j Z_{t-j} = X_t$. Therefore $\{X_t\}$ and $\{Y_t\}$ are equal in mean-square, thus $\{X_t\}$ is the unique stationary solution to the

AR(1) equation (up to mean-square equivalence).

If we choose to define $\{X_t\} \sim \text{AR}(1)$ with $|\phi| < 1$, then $\{X_t\}$ has a "unique" representation as an MA(∞) (or linear) process.

Note

If $|\phi| = 1$, *i.e.* $X_t = X_{t-1} + Z_t$ or $X_t = -X_{t-1} + Z_t$. Suppose (for the sake of contradiction) that $\{X_t\}$ is a stationary solution to the equation $X_t = \phi X_{t-1} + Z_t$ with $|\phi| = 1$. Then by the previous recursion,

$$X_t = \phi^{n+1} X_{t-n-1} + \sum_{j=0}^{n} \phi^j Z_{t-j}$$

and let $R_{t,n} = X_t - \phi^{n+1} X_{t-n-1}$. Then

$$\mathsf{Var}\left(R_{t,n}\right) = \sum_{j=0}^{n} \phi^{2j} \mathsf{Var}\left(Z_{t-j}\right) = \sum_{j=0}^{n} \mathsf{Var}\left(Z_{t-j}\right) = (n+1)\sigma^2.$$

The variance of that remainder go to infinity as $n \to \infty$. Therefore, the variance of X_t grows to infinity as if the converse was true and $\mathsf{Var}\left(X_t\right)$ was finite, then $\mathsf{Var}\left(X_t - \phi^{n+1} X_{t-n-1}\right) < \infty$. Therefore, no such stationary solution exists. What happens if $|\phi| > 1$? Clearly in this case $\sum_{j=0}^{\infty} \phi^j Z_{t-j}$ diverges because the coefficient of Z_{t-j} grows without bound. But note that we can also rewrite the AR(1) equation from its original form $X_t = \phi X_{t-1} + Z_t \Rightarrow X_{t+1} = \phi X_t + Z_{t+1}$ which imply that

$$\begin{aligned} X_t &= \frac{1}{\phi} X_{t+1} - \frac{1}{\phi} Z_{t+1} \\ &= \frac{1}{\phi^2} X_{t+2} - \frac{1}{\phi} Z_{t+1} - \frac{1}{\phi^2} Z_{t+2} \\ \ldots &= -\sum_{j=1}^{\infty} \phi^{-j} Z_{t+j} \end{aligned}$$

with $\sum_{j=1}^{\infty} |\phi^{-j}| \leq \sum_{j=0}^{\infty} |\phi^{-1}|^j < \infty$ as $|\phi^{-1}| < 1$.

We **do** have a stationary solution $X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$ where

$$\psi_j = \begin{cases} -\phi^{-j} & \text{if } j < 0 \\ 0 & \text{if } j \geq 0 \end{cases}$$

Since the white noise are residuals from the future observations, this is not of much use in practical terms.

To summarize, there are three situations for the AR(1): if

- $|\phi| < 1$, stationary and causal
- $|\phi| = 1$, non-stationary
- $|\phi| > 1$, stationary, non-causal.

Note

We were able in the AR(1) case to do the expansion $(1-\phi B)^{-1} \equiv \sum_{j=0}^{\infty} \phi_j B^j$ for a single factor $(1-\phi B)$ as an operator acting on $\{X_t\}$ to produce the MA(∞) or linear process representation. But suppose

$$\alpha(B) = \sum_{j=-\infty}^{\infty} \alpha_j B^j$$
$$\beta(B) = \sum_{j=-\infty}^{\infty} \beta_j B^j$$

where $\sum_{j=-\infty}^{\infty} |\alpha_j| < \infty$, $\sum_{j=-\infty}^{\infty} |\beta_j| < \infty$. Then

$$\begin{aligned}
\alpha(B)\beta(B)Z_t &= \left(\sum_{j=-\infty}^{\infty} \alpha_j B^j\right)\left(\sum_{j=-\infty}^{\infty} \beta_j B^j\right) Z_t \\
&= \left(\sum_{j=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} \alpha_j\beta_k B^{j+k}\right) Z_t \\
&= \left(\sum_{l=-\infty}^{\infty}\left\{\sum_{k=-\infty}^{\infty} \alpha_k\beta_{l-k}\right\} B^l\right) Z_t \\
&= \left(\sum_{l=-\infty}^{\infty} \psi_l B^l\right) Z_t
\end{aligned}$$

where $\psi_l = \sum_{k=-\infty}^{\infty} \alpha_k\beta_{l-k}$. The last line is a linear process representation, as it is just

$$\sum_{l=-\infty}^{\infty} \psi_l Z_{t-l}.$$

Note that we need

$$\begin{aligned}
\sum_{l=-\infty}^{\infty} |\psi_l| &\le \sum_{l=-\infty}^{\infty}\sum_{k=-\infty}^{\infty} |\alpha_k\beta_{l-k}| \\
&\le \left(\sum_{k=-\infty}^{\infty} |\alpha_k|\right)\left(\sum_{l=-\infty}^{\infty} |\beta_{l-k}|\right) < \infty
\end{aligned}$$

The two linear filters combine to yield another linear filter

$$\alpha(B)\beta(B) = \beta(B)\alpha(B) = \psi(B)$$

and we can consider the filters applied sequentially. *i.e.*

$$\alpha(B)\beta(B)Z_t = \alpha(B)\{\beta(B)Z_t\} = \alpha(B)Y_t$$

where $\{Y_t\}$ defined by $Y_t = \beta(B)Z_t$ is a stationary process.

1.3.1. Autoregressive model of order p (AR(p))

Suppose $\{X_t\}$ is the solution to the stochastic equation

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \cdots - \phi_p X_{t-p} = Z_t$$

$\{X_t\}$ is the autoregressive process of order p (denoted $\{X_t\} \sim \text{AR}(p)$) with real coefficients $\phi_1, \ldots, \phi_p$.

Example 1.12 (AR(2) process)
Take $X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} = Z_t$ for $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. By identical arguments to the AR(1) case. We have $\mathsf{E}(Z_t) = 0, \mathsf{Var}(Z_t) = \sigma^2 < \infty$, which leads to $\mathsf{E}(X_t) = 0$. Note that $\{X_t\}$ can be found as the solution to the equation

$$(1 - \phi_1 B - \phi_2 B^2)X_t = Z_t$$

– but can we write

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

for some collection of coefficients $\{\psi_j\}$ such that $\{X_t\}$ converges in mean square? If so, then the ψ_j must be the coefficients in the series expansion of $(1 - \phi_1 B - \phi_2 B^2)^{-1}$. First, note that

$$\Phi(B) = (1 - \phi_1 B - \phi_2 B^2) = (1 - \xi_1 B)(1 - \xi_2 B)$$

where we identified the complex-valued parameters ξ_1, ξ_2 by equating coefficients, that is $\phi_1 = \xi_1 + \xi_2$ and $\phi_2 = \xi_1 \xi_2$.

In general, ξ_1, ξ_2 will be complex-valued and in this case, they will form a complex conjugate pair (and have same modulus). However, in some cases, ξ_1 and ξ_2 will be entirely **real.** Note also that if $\Phi(z) = 1 - \phi_1 z - \phi_2 z^2$, then ξ_1 and ξ_2 solve $\Phi(z) = 0$ for $|z| \leq 1$ (*i.e.* ξ_1 and ξ_2

are the reciprocal roots of $\Phi(z) = 0$). Note finally that we could write

$$\Phi(z) = \phi_2 \left(\frac{1}{\phi_2} - \frac{\phi_1}{\phi_2} - z^2 \right)$$

Therefore $\eta_1 = \xi_1^{-1}$ and $\eta_2 = \xi_2^{-1}$ are the **roots** of

$$\begin{aligned} z^2 - \phi_1/\phi_2 z - 1/\phi_2 &= 0 \\ (z - \eta_1)(z - \eta_2) &= 0 \end{aligned}$$

So $\Phi(B) = (1 - \xi_1 B)(1 - \xi_2 B)$ and thus

$$\begin{aligned} \{\Phi(B)\}^{-1} &= \{(1 - \xi_1 B)(1 - \xi_2 B)\}^{-1} \\ &= (1 - \xi_1 B)^{-1}(1 - \xi_2 B)^{-1} \end{aligned}$$

as

$$\begin{aligned} \Phi(B)\{\Phi(B)\}^{-1} &= (1 - \xi_1 B)(1 - \xi_2 B)(1 - \xi_1 B)^{-1}(1 - \xi_2 B)^{-1} \\ &= (1 - \xi_1 B)(1 - \xi_1 B)^{-1}(1 - \xi_2 B)(1 - \xi_2 B)^{-1} \\ &= 1 \end{aligned}$$

as the operators commute.

What we have is that we can write $(1 - \phi_1 B - \phi_2 B^2)$ as

$$\begin{aligned} (1 - \xi_1 B)^{-1}(1 - \xi_2 B)^{-1} &= \left(\sum_{j=0}^{\infty} \xi_1^j B^j \right) \left(\sum_{k=0}^{\infty} \xi_2^k B^k \right) \\ &= \psi_1(B)\psi_2(B) \end{aligned}$$

with each sum convergent provided $|\xi_1|, |\xi_2| < 1$. Indeed, this is a necessary and sufficient condition. Thus

$$X_t = \Psi(B) Z_t = \psi_1(B)\psi_2(B) Z_t$$

where

$$\Psi(B) = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} \xi_1^j \xi_2^k B^{l+k}$$

with $\psi_r = \sum_{k=0}^{r} \xi_1^k \xi_2^{r-k}$ for $r \geq 0$.

This is a valid series expansion, we need to verify some conditions on ψ_r. We have $|\psi_r| \leq$

$\sum_{k=0}^{r} |\xi_1|^k |\xi_2|^{r-k}$. We require that $\sum_{r=0}^{\infty} |\psi_r| < \infty$. Explicitly, if ξ_1, ξ_2 are a complex pair, so that $|\xi_1| = |\xi_2|$, then $|\psi_r| \leq \sum_{k=0}^{r} |\xi_1|^r = (r+1)M^r$ say, where $M = |\xi_1| < 1$. Therefore,

$$\sum_{r=0}^{\infty} |\psi_r| \leq \sum_{r=0}^{\infty} (r+1)M^r < \infty$$

as $M < 1$, by standard results for convergent series.

If ξ_1, ξ_2 are real valued, take $M = \max(|\xi_1|, |\xi_2|) < 1$. The same result follows. The AR(2) process is stationary and causal provided that $|\xi_1|, |\xi_2| < 1$, that is the roots are inside the unit circle.

For the AR(p), in the more general case, we can apply the same results and construction. Write

$$\begin{aligned} \Phi(B) &= (1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3 - \cdots - \phi_p B^p) \\ &= \prod_{r=1}^{p} (1 - \xi_r B) \end{aligned}$$

The AR(p) process is stationary and causal if $|\xi_r| < 1$ for $r = 1, \ldots p$. The values $\xi_1, \ldots, \xi_p$ solve $\Phi(z) = 0$; as before, if $\eta_r = \xi_r^{-1}$, then $\eta_1, \ldots, \eta_p$ solve $\Phi(z) = 0$ also with $|\eta_r| > 1 \; \forall \; r$.[25]

What if $|\xi_r| = 1$ for some r? Or if $|\xi_r| > 1$? The first case will turn to be nonstationary, while for the second, one can construct a stationary process in a certain way, but which turns out to be non-causal.

Section 1.4: Moving Average Processes

Recall the MA(1) process $\{X_t\}$ defined by $X_t = Z_t + \theta Z_{t-1}$ for $t \in \mathbb{Z}$ and the more general MA(q) process, defined by

$$X_t = Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. We know that $\mathsf{E}(X_t) = 0 \; \forall \; t$ and also that $\{X_t\}$ is stationary. Also, we have

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$$

[25] Sometimes seen as the condition that all the roots lie outside the unit circle, contrary to ξ_r lying inside the unit circle).

where

$$\psi_j = \begin{cases} \theta_j, & \text{if } j = 1, 2, \ldots, q \\ 1, & \text{if } j = 0 \\ 0, & \text{if } j < 0 \end{cases}$$

so the MA(q) process for $q \geq 1$ has a straightforward linear process representation. By the earlier results,

$$\begin{aligned} \gamma_X(h) &= \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+|h|} \\ &= \sum_{j=0}^{q-|h|} \theta_j \theta_{j+|h|} \end{aligned}$$

where $\theta_0 \equiv 1$ with the convention that the sum is zero if $|h| > q$.

1.4.1. MA(q) Process as an AR process

Consider $\{X_t\} \sim \text{MA}(1)$ with

$$X_t = (1 + \theta B) Z_t.$$

To get an AR representation $\Pi(B) X_t = Z_t$, we must find $\Pi(B)$. We may formally write

$$\Pi(B) = (1 + \theta B)^{-1} = \sum_{j=0}^{\infty} (-\theta)^j B^j$$

(as $(1 + \theta B)^{-1}(1 + \theta B) = 1$ using this definition) provided $|\theta| < 1$, otherwise the sum $\Pi(B) X_t = \sum_{j=0}^{\infty} (-\theta)^j X_{t-j}$ would not be mean-square convergent.

If $|\theta| < 1$, $\{X_t\} \sim \text{MA}(1)$ admits the representation

$$\Pi(B) X_t = Z_t$$

where $\Pi(B) = \sum_{j=-\infty}^{\infty} \pi_j B^j$ with $\pi_j = (-\theta)^j$ if $j \geq 0$ and zero otherwise, that is $\{X_t\} \sim \text{AR}(\infty)$.

If $\{X_t\} \sim \text{MA}(q)$, say $X_t = \Theta(B) Z_t$, where $\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$, then we

may factorize

$$\Theta(B) = \prod_{l=1}^{q}(1 - \omega_l B).$$

Hence $\{X_t\}$ admits an AR(∞) representation, i.e.

$$\Pi(B)X_t = Z_t$$

with $\Pi(B) = \sum_{j=-\infty}^{\infty} \pi_j B^j$ provided that $|\omega_l| < 1$, for $l = 1, \ldots, q$ where π_j is the coefficient of B^j in the expansion of

$$\begin{aligned}(1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q)^{-1} &= \prod_{l=1}^{q}(1 - \omega_l B)^{-1} \\ &= \prod_{l=1}^{q}\left(\sum_{j=-\infty}^{\infty} \omega_l^j B^j\right)\end{aligned}$$

If $\{X_t\}$ admits an AR(∞) representation, that is $|\omega_l| < 1$ for $l = 1, \ldots, q$, we say that $\{X_t\}$ is **invertible** (note that the AR(∞) representation is **causal.**)

If $\{X_t\} \sim$ MA(1) with $|\theta| > 1$, X_t is still stationary, but it does not admit a causal AR(∞) representation.

However, we may write $Z_t = -\sum_{j=1}^{\infty}(-\theta)^{-j} X_{t+j}$ which is a mean-square convergent AR(∞) process that is **non-causal**.

Section 1.5: Forecasting Stationary Processes

We aim to predict X_{n+h} on the basis of $X_1, \ldots, X_n$. Optimal prediction in general is largely intractable, so we focus on linear predictors[26].

Criterion : In the context of weakly stationary processes, with finite first and second moments, a natural criterion to optimize with respect to is **minimum mean squared error.** That is, for linear predictors, we aim to choose coefficients $a_1, \ldots, a_n$ to minimize the expected value of the squared difference between the true value of the variable to be predicted and the predictor, that is

$$\mathsf{E}\left((X_{n+h} - \widehat{X}_{n+h})^2\right)$$

[26]For practicality reasons and due to the analogy with the Gaussian case

where

$$\widehat{X}_{n+h} = a_0 + \sum_{i=1}^{n} a_i X_{n-i+1}$$

i.e. solve the minimization problem over a

$$\min_a \mathsf{E}\left(\left(X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n-i+1}\right)^2\right) = \min_a \text{MSE}(a)$$

for variable a. Note that the expectation is over all the random variables, which are $X_1, \ldots, X_n, X_{n+h}$. We may solve this optimization (minimization) problem analytically.

$$\frac{\partial \text{MSE}(a)}{\partial a_0} = \mathsf{E}\left(2\left(X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n-i+1}\right)\right) = 0 \tag{1.5}$$

$$\frac{\partial \text{MSE}(a)}{\partial a_j} = \mathsf{E}\left(2\left(X_{n+h} - a_0 - \sum_{i=1}^{n} a_i X_{n-i+1}\right) X_{n-j+1}\right) = 0 \tag{1.6}$$

for $j = 1, \ldots, n$. From (1.5), we have

$$\mathsf{E}\left(X_{n+h}\right) - \sum_{i=1}^{n} a_i \mathsf{E}\left(X_{n-i+1}\right) = a_0.$$

But under stationarity, $\mathsf{E}\left(X_t\right) = \mu$ (say) for all t. Therefore,

$$a_0 = \mu\left(1 - \sum_{i=1}^{n} a_i\right)$$

From (1.6) and substituting this expression for a_0, we get

$$\mathsf{E}\left(X_{n+h}X_{n-j+1}\right) - \mu\left(1 - \sum_{i=1}^{n} a_i\right)\mathsf{E}\left(X_{n-j+1}\right) - \sum_{i=1}^{n} a_i \mathsf{E}\left(X_{n-i+1}X_{n-j+1}\right) = 0$$

We can rewrite this as

$$\mathsf{E}\left(X_{n+h}X_{n-j+1} - \mu^2\right) - \sum_{i=1}^{n} a_i \mathsf{E}\left(X_{n-i+1}X_{n-j+1} - \mu^2\right) = 0$$

imply that

$$\gamma_X(h+j-1) - \sum_{i=1}^{n} a_i \gamma_X(i-j) = 0$$

for $j = 1, \ldots, n$. Let

$$\boldsymbol{\gamma}_X(n,h) \equiv (\gamma_X(h), \ldots, \gamma_X(h+n-1))^\top$$

and again denote the covariance matrix

$$\boldsymbol{\Gamma}_X(n) = [\gamma_X(i-j)]_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}}$$

and similarly, $\boldsymbol{a}_{1:n} = (a_1, \ldots, a_n)^\top$. Then the system of equations to be solved becomes

$$\boldsymbol{\Gamma}_X(n)\boldsymbol{a}_{1:n} = \boldsymbol{\gamma}_X(n,h)$$

an $n \times 1$ system of equations. The minimum MSE is achieved when $\boldsymbol{a}_{1:n}$ solves this equation; the solution depends on both n and h, so write $\boldsymbol{a}_{1:n}(h) = (a_1(h), \ldots, a_n(h))^\top$ as the optimal value. The optimal forecast (prediction) is then $\widehat{X}_{n+h} = \mu + \sum_{i=1}^{n} a_i(h)(X_{n-i+1} - \mu)$. [27]
The optimal forecast is $\widehat{X}_{n+h} = \mu + \sum_{i=1}^{n} a_i(h)(X_{n-i+1} - \mu)$ and the minimum MSE is

$$\begin{aligned}
\mathsf{E}\left((X_{n+h} - \widehat{X}_{n+h})^2\right) &= \gamma_X(0) - \boldsymbol{a}_{1:n}(h)^\top \boldsymbol{\gamma}_X(n,h) \\
&= \gamma_X(0) - 2\sum_{i=1}^{n} a_i \gamma_X(h+i-1) + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i \gamma(i-j) a_j
\end{aligned}$$

Note

Finding $\boldsymbol{a}_{1:n}(h)$from

$$\boldsymbol{\Gamma}_X(n)\boldsymbol{a}_{1:n} = \boldsymbol{\gamma}_X(n,h)$$

is a straightforward numerical exercise as $\boldsymbol{\Gamma}_X(n)$ is non-negative definite, symmetric, Toeplitz. [28] [29]

[27] The optimal construction depends upon the covariance of the process. Clearly, $\boldsymbol{a}_{1:n}(h)$ is a function of the covariance sequence.

[28] In R, using the QR decomposition (through solve). For $n > 2000$, the computation becomes prohibitive. In general, Γ_X is fully parametrized in most cases, and the matrix can be sparse. Choleski, eigenvalue decomposition may work effectively.

[29] If we are dealing with a MA process, we can use the AR(∞) approximation and truncate the coefficients for forecasting.

Example 1.13

In the AR(1) case, $X_t = \phi X_{t-1} + Z_t$ where $Z_t \sim \text{WN}(0, \sigma^2)$ and $|\phi| < 1$. Then

$$\mathbf{\Gamma}_X(n) = \frac{\sigma^2}{1-\phi^2} \begin{bmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{n-1} \\ \phi & 1 & \phi & \cdots & \cdots \\ \phi^2 & \phi & 1 & \cdots & \cdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \phi^{n-1} & \cdots & \cdots & \phi & 1 \end{bmatrix}$$

and

$$\begin{aligned} \boldsymbol{\gamma}_X(n,h) &= (\gamma_X(h), \cdots, \gamma_X(n+h-1))^\top \\ &= \frac{\sigma^2}{1-\phi^2}(\phi^h, \phi^{h+1}, \cdots, \phi^{n+h-1})^\top \end{aligned}$$

For $h = 1$, we have $\mathbf{\Gamma}_X(n)\boldsymbol{a}_{1:n}(h) = \gamma_X(n,1)$ yields $\boldsymbol{a}_{1:n}(1) = (\phi, 0, \cdots, 0)^\top$. Thus,

$$\begin{aligned} \widehat{X}_{n+1} &= \mu + \sum_{i=1}^{n} a_i(i)(X_{n-i+1} - \mu) \\ &= \mu + \phi(X_n - \mu) \end{aligned}$$

which equals ϕX_n for a zero mean process.

By similar methods, we have that

$$\widehat{X}_{n+h} = \phi^h X_n$$

almost surely for $h \geq 1$.

Note

As $h \to \infty, \widehat{X}_{n+h} \to 0$ almost surely, which corresponds to the mean of the process.

Example 1.14

Consider the AR(p) case, with a general model of the form $\Phi(B)X_t = Z_t$

$$X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p} = Z_t$$

one can go through the same arguments and conclude that the optimal forecast for $\widehat{X}_{n+1}$ is

$$\widehat{X}_{n+1} = \sum_{i=1}^{n} \phi_i X_{n-i}$$

where $\phi_i = 0$ for $i > p$.

Proposition 1.13 (Solving for $\boldsymbol{a}_{1:n}(h)$)

If $\boldsymbol{\Gamma}_X(n)$ is non-singular, we may write $\boldsymbol{a}_{1:n}(h) = \{\boldsymbol{\Gamma}_X(n)\}^{-1}\gamma_X(n,h)$, but computing the matrix inverse may be prohibitive.

Sufficient conditions for non-singularity of $\boldsymbol{\Gamma}_X(n)$

1. $\gamma_X(0) > 0$
2. $\gamma_X(h) \to 0$ as $h \to \infty$.

Note

For real time series data, $\gamma_X(h)$ is not known, so sample-based estimates $\widehat{\gamma}_X(h)$ are used,

$$\widehat{\gamma}_X(h) = \frac{1}{n}\sum_{t=1}^{n-h}(X_{t+h} - \bar{X}_n)(X_t - \bar{X}_n)$$

where $\bar{X}_n = n^{-1}\sum_{t=1}^{n} X_t$.[30]

Two algorithms can be used to compute the optimal coefficients recursively; this is useful for large n, or when data are being observed on an ongoing basis.

Sum of correlated observations are harder to deal with, although the form is nice. The Levinson-Durbin provides a recursive algorithm that uses the residuals.

Algorithm 1.1 (Levinson-Durbin)

Without loss of generality, take $\mathsf{E}(X_t) = 0$. Write

$$\begin{aligned}\widehat{X}_{n+1} &= \sum_{i=1}^{n}\varphi_{n,i}X_{n-i+1}\\ &= \boldsymbol{\varphi}_n^\top X_{n:1}\end{aligned}$$

where $\boldsymbol{\varphi}_n = (\varphi_{n,1}, \cdots, \varphi_{n:n})^\top$, which corresponds in our prediction forecast to $\boldsymbol{a}_{1:n}$ The optimal choice solves

$$\boldsymbol{\Gamma}_n\boldsymbol{\varphi}_n = \boldsymbol{\gamma}(n,1),$$

which is an $n \times 1$ system of equations, yielding minimum MSE

$$v_n = \gamma(0) - \boldsymbol{\varphi}_n^\top\boldsymbol{\gamma}(n,1)$$

[30] One assumes that the order of the process is finite, so that some terms be zero in the sum. Note that the last term, since we are dealing with an autocorrelation, then the variance of the sum will be larger than the variance for the IID case. Each sample observation brings less information about the sample then an IID counterpart.

Recursion: Define $\boldsymbol{\varphi}_n$ in terms of $\boldsymbol{\varphi}_{n-1}$ and v_{n-1} as follows:

$$\boxed{\varphi_{n,n} = \frac{1}{v_{n-1}}\left[\gamma(n) - \boldsymbol{\varphi}_{n-1}^{\top}\boldsymbol{\gamma}^{\mathrm{R}}(n-1,1)\right]}$$

where $\boldsymbol{\gamma}^{\mathrm{R}}(k,1) = (\gamma(k), \gamma(k-1), \ldots, \gamma(1))^{\top}$ and assuming γ is known, and

$$\boxed{\boldsymbol{\varphi}_{n,1:(n-1)} = \boldsymbol{\varphi}_{n-1} - \varphi_{n,n}\boldsymbol{\varphi}_{n-1}^{\mathrm{R}}}$$

for $\boldsymbol{\varphi}_{n-1}, \boldsymbol{\varphi}_{n-1}^{\mathrm{R}}$ is a $(n-1)\times 1$ vector, $\varphi_{n,n}$ a scalar and $\boldsymbol{\varphi}_{n-1}^{\mathrm{R}} = (\varphi_{n-1,n-1}, \varphi_{n-1,n-2}, \cdots, \varphi_{n-1,1})^{\top}$.[31]

Then set

$$\boxed{v_n = v_{n-1}(1 - \varphi_{n,n}^2).}$$

Initialization: For $n = 1, \varphi_{1,1} = \frac{\gamma(1)}{\gamma(0)} = \rho(1)$ with $v_0 = \gamma(0)$ and $v_1 = \gamma(0)(1 - \rho^2(1))$.

To see why this works, let

$$\mathbf{P}_n = \frac{1}{\gamma(0)}\boldsymbol{\Gamma}_n$$

be the autocorrelation matrix and let

$$\boldsymbol{\rho}_{1:n} = \frac{1}{\gamma(0)}(\gamma(1), \ldots, \gamma(n))^{\top}$$
$$\boldsymbol{\rho}_{n:1} = \boldsymbol{\rho}_{1:n}^{\mathrm{R}} = (\rho(n), \cdots, \rho(1))^{\top}$$

Proof Note that $\boldsymbol{\Gamma}_n\boldsymbol{\varphi}_n = \boldsymbol{\gamma}(n,1) \Rightarrow \mathbf{P}_n\boldsymbol{\varphi}_n = \boldsymbol{\rho}_{1:n}$. We need to verify that $\mathbf{P}_n\boldsymbol{\varphi}_n = \boldsymbol{\rho}_{1:n} \Rightarrow \mathbf{P}_{n+1}\boldsymbol{\varphi}_{n+1} = \boldsymbol{\rho}_{1:(n+1)}$.

For $n = 1$, $\mathbf{P}_1 = 1, \boldsymbol{\varphi}_1 = \varphi_{1,1} = \rho(1) = \boldsymbol{\rho}_{1,1}$. By induction on n. Assume that $\mathbf{P}_k\boldsymbol{\varphi}_k = \boldsymbol{\rho}_{1:k}$. Now

$$\mathbf{P}_{k+1} = \begin{bmatrix} \mathbf{P}_k & \boldsymbol{\rho}_{k:1} \\ \boldsymbol{\rho}_{k:1}^{\top} & 1 \end{bmatrix}$$

a block matrix and also $\mathbf{P}_k\boldsymbol{\varphi}_{\boldsymbol{k}} = \boldsymbol{\rho}_{1:k}$, therefore

$$\mathbf{P}_k\boldsymbol{\varphi}_k^{\mathrm{R}} = \boldsymbol{\rho}_{k:1}$$

[31] The calculations depend on an inner product, of $O(n)$, rather than $n\log(n)$. This is thus more computationally efficient.

using the symmetry of $\mathbf{P}_k$. By the proposed recursion, we have that

$$\mathbf{P}_{k+1}\boldsymbol{\varphi}_{k+1} = \begin{bmatrix} \mathbf{P}_k & \boldsymbol{\rho}_{k:1} \\ \boldsymbol{\rho}_{k:1}^\top & 1 \end{bmatrix} \begin{pmatrix} \boldsymbol{\varphi}_k - \varphi_{k+1,k+1}\boldsymbol{\varphi}_k^{\mathrm{R}} \\ \varphi_{k+1,k+1} \end{pmatrix} = \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix}$$

where $\mathbf{a}$ is equal to a $k \times 1$ system of the form

$$\begin{aligned} \mathbf{a} &= \mathbf{P}_k\boldsymbol{\varphi}_k - \varphi_{k+1,k+1}\mathbf{P}_k\boldsymbol{\varphi}_k^{\mathrm{R}} + \varphi_{k+1,k+1}\boldsymbol{\rho}_{k:1} \\ &= \mathbf{P}_k\boldsymbol{\varphi}_k = \boldsymbol{\rho}_{1:k} \end{aligned}$$

by the induction hypothesis. For $\mathbf{b}$, we have

$$\mathbf{b} = \boldsymbol{\rho}_{k:1}^\top\boldsymbol{\varphi}_k - \varphi_{k+1,k+1}\boldsymbol{\rho}_{k:1}^\top\boldsymbol{\varphi}_k^{\mathrm{R}} + \varphi_{k+1,k+1}$$

$$= \boldsymbol{\rho}_{k:1}^\top\boldsymbol{\varphi}_k - \varphi_{k+1,k+1}(1 - \boldsymbol{\rho}_{k:1}^\top\boldsymbol{\varphi}_k^{\mathrm{R}})$$

We have now

$$\varphi_{k+1,k+1} = \frac{1}{v_k}\left[\gamma(k+1) - \boldsymbol{\varphi}_k^\top\boldsymbol{\gamma}^{\mathrm{R}}(k,1)\right]$$

where

$$\begin{aligned} v_k &= \gamma(0) - \boldsymbol{\varphi}_k^\top\boldsymbol{\gamma}(k,1) \\ &= \gamma(0)\left(1 - \boldsymbol{\varphi}_k^\top\boldsymbol{\rho}_{1:k}\right) \\ &= \gamma(0)(1 - \boldsymbol{\rho}_{k:1}^\top\boldsymbol{\varphi}_k^{\mathrm{R}}) \end{aligned}$$

From the above formula, we have

$$\varphi_{k+1,k+1} = \frac{1}{\gamma(0)(1 - \boldsymbol{\rho}_{k:1}^\top\boldsymbol{\varphi}_k^{\mathrm{R}})}\left[\gamma(k+1) - \boldsymbol{\varphi}_k^\top\gamma^{\mathrm{R}}(k,1)\right] \tag{1.7}$$

changing the order of the inner product and from $\mathbf{b}$, we have

$$\begin{aligned} &\boldsymbol{\rho}_{k:1}^\top\boldsymbol{\varphi}_k + \frac{\left[\gamma(k+1) - \boldsymbol{\varphi}_k^\top\gamma^{\mathrm{R}}(k,1)\right]}{\gamma(0)(1 - \boldsymbol{\rho}_{k:1}^\top\boldsymbol{\varphi}_k^{\mathrm{R}})}(1 - \boldsymbol{\rho}_{k:1}^\top\boldsymbol{\varphi}_k^{\mathrm{R}}) \\ &= \boldsymbol{\rho}_{k:1}^\top\boldsymbol{\varphi}_k + \rho(k+1) - \boldsymbol{\varphi}_k^\top\boldsymbol{\rho}_{k:1} \\ &= \rho(k+1) \end{aligned}$$

■

We have effectively performed a matrix inversion by considering a block decomposition

$$\mathbf{P}_{k+1} = \begin{bmatrix} \mathbf{P}_k & \boldsymbol{\rho}_{k:1} \\ \boldsymbol{\rho}_{k:1}^\top & 1 \end{bmatrix}$$

and then computed $\mathbf{P}_{k+1}^{-1}$ using $\mathbf{P}_k^{-1}$. [32] For v_n, we have

$$\begin{aligned} v_n &= \mathsf{E}\left((X_{n+1} - \boldsymbol{\varphi}_n^\top X_{n:1})^2\right) \\ &= \gamma(0) - \boldsymbol{\varphi}_n^\top \gamma(n-1) \\ &= \gamma(0) - \boldsymbol{\varphi}_{n-1}^\top \boldsymbol{\gamma}(n-1,1) + \varphi_{n,n}(\boldsymbol{\varphi}_{n-1}^R)^\top \boldsymbol{\gamma}(n-1,1) - \varphi_{n,n}\gamma(n) \\ &= v_{n-1} + \varphi_{n,n}\big((\boldsymbol{\varphi}_{n-1}^R)^\top \boldsymbol{\gamma}(n-1,1) - \gamma(n)\big) \\ &= v_{n-1} - \varphi_{n,n}^2\big(\gamma(0) - (\boldsymbol{\varphi}_{n-1}^R)^\top \boldsymbol{\gamma}(n-1,1)\big) \\ &= v_{n-1} - \varphi_{n,n}^2 v_{n-1} \\ &= v_{n-1}(1 - \varphi_{n,n}^2) \end{aligned}$$

plugging in (1.7). That is, $v_n = v_{n-1}(1-\varphi_{n,n}^2)$ and as $v_n \geq 0$, imply $\varphi_{n,n}^2 < 1, |\varphi_{n,n}| < 1$ thus $v_n \leq v_{n-1}$, *i.e.* the $\{v_n\}$ form a non-increasing sequence.

Note

The Levinson-Durbin recursion solves the equation $\boldsymbol{\Gamma}_n \boldsymbol{\varphi}_n = \boldsymbol{\gamma}(n,1)$ without matrix inversion, in order n^2 operations. Most matrix inversion procedures are order n^3; L-D exploits the Toeplitz structure.[33]

Section 1.6: Partial autocorrelation

Define the function $\alpha_X(h)$ by

$$\alpha_X(h) = \begin{cases} 1, & \text{if } h = 0 \\ \phi_{h,h}, & \text{if } h \geq 1 \end{cases}$$

for $h = 0, 1, 2, \ldots$ This $\alpha_X(h)$ is the **partial autocorrelation function** (PACF). For random variables X, Y, Z, the partial correlation is defined by

$$\mathsf{Cor}(X - \mathsf{E}(X|Z)), (Y - \mathsf{E}(Y|Z)))$$

- this is the partial correlation for X and Y given (accounting for) Z; it is the correlation between residuals obtained by regressing in turn X and on Z and Y on Z. The PACF

[32] This works because of the nature of the nature of the matrix (Toeplitz structure). When we get to estimation by maximum likelihood estimates, we will get the likelihood in terms of matrix inverses.

[33] One can also use singular value, eigenvalue decomposition or Choleski decomposition; these method have special implementation for positive-definite Toeplitz structure matrices.

computes the correlation between

$$X_t - \mathsf{E}\left(X_t | X_{(t+1):(t+h-1)}\right) \tag{1.8a}$$
$$X_{t+h} - \mathsf{E}\left(X_{t+h} | X_{(t+1):(t+h-1)}\right) \tag{1.8b}$$

Example 1.15

Consider an causal stationary AR(p) process; for $h > p$,

$$\mathsf{E}\left(X_{t+h} | X_{(t+1):(t+h-1)}, X_{1:(t-1)}\right) = \phi X_{t+h-1} + \cdots + \phi_p X_{t+h-p}$$
$$\mathsf{E}\left(X_t | X_{(t+1):(t+h-1)}, X_{1:(t-1)}\right) = \phi_1 X_{t+h-1} + \cdots + \phi_p X_{t+h-p}$$

where now (1.8a) equals Z_t and (1.8b) is Z_{t+h}, therefore the correlation $\mathsf{Cor}(Z_t, Z_{t+h}) = 0$ and $\alpha_X(h) = 0$ for $h > p$, while for $h \leq p, \alpha_X(h) \neq 0$. That is, we can diagnose an AR(p) structure by inspecting whether the partial autocorrelation drops to zero at some finite lag.

Algorithm 1.2 (Innovations algorithm)

Let $\{X_t\}$ be zero-mean process with $\mathsf{E}\left(X_t^2\right) < \infty$ and let $\tilde{\gamma}_X(t,s) = \mathsf{E}\left(X_t X_s\right)$ (not necessarily stationary). Let

$$\widehat{X}_n = \begin{cases} 0 & \text{if } n = 1 \\ \mathsf{E}\left(X_n | \boldsymbol{X}_{1:(n-1)}\right) & \text{if } n \geq 2 \end{cases}$$

Let $U_n = X_n - \widehat{X}_n$ be the error in prediction and let $\boldsymbol{U}_{1:n}$ be the vector of prediction errors (computed in a one-step ahead fashion). As $\mathsf{E}\left(X_n | \boldsymbol{X}_{1:(n-1)}\right)$ is linear in $X_1, \ldots, X_{n-1}$, we may write $\boldsymbol{U}_{1:n} = \boldsymbol{A}_n \boldsymbol{X}_{1:n}$ where A_n is given by

$$\boldsymbol{A}_n = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ a_{1,1} & 1 & 0 & 0 & 0 & 0 \\ a_{2,2} & a_{2,1} & 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & \\ a_{n-1,n-1} & \cdots & \cdots & \cdots & a_{n-1,1} & 1 \end{bmatrix}$$

as

$$U_1 = X_1$$
$$U_2 = X_2 - \widehat{X}_2 = X_2 + a_{1,1} X_1$$
$$U_3 = X_3 - \widehat{X}_3 = X_3 + a_{2,2} X_2 + a_{2,1} X_1$$
$$\cdots$$

Let $\boldsymbol{C}_n = \boldsymbol{A}_n^{-1}$, again a lower triangular matrix, where

$$\boldsymbol{C}_n = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ \vartheta_{1,1} & 1 & 0 & 0 & 0 & 0 \\ \vartheta_{2,2} & \vartheta_{2,1} & 1 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & \\ \vartheta_{n-1,n-1} & \cdots & \cdots & \cdots & \vartheta_{n-1,1} & 1 \end{bmatrix}$$

Then $\boldsymbol{X}_{1:n} = \boldsymbol{C}_n \boldsymbol{U}_{1:n} = \boldsymbol{C}_n(\boldsymbol{X}_{1:n} - \widehat{\boldsymbol{X}}_{1:n})$.

Also, $\widehat{\boldsymbol{X}}_{1:n} = \boldsymbol{X}_{1:n} - \boldsymbol{U}_{1:n} = \boldsymbol{C}_n \boldsymbol{U}_{1:n} - \boldsymbol{U}_{1:n} = (\boldsymbol{C}_n - \mathbf{I}_n)\boldsymbol{U}_{1:n}$, that is

$$\widehat{\boldsymbol{X}}_{1:n} = (\boldsymbol{C}_n - \mathbf{I}_n)(\boldsymbol{X}_{1:n} - \widehat{\boldsymbol{X}}_{1:n}) = \boldsymbol{\Theta}_n(\boldsymbol{X}_{1:n} - \widehat{\boldsymbol{X}}_{1:n})$$

where

$$\boldsymbol{\Theta}_n = \boldsymbol{C}_n - \mathbf{I}_n = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \vartheta_{1,1} & 0 & 0 & 0 & 0 & 0 \\ \vartheta_{2,2} & \vartheta_{2,1} & 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & \\ \vartheta_{n-1,n-1} & \cdots & \cdots & \cdots & \vartheta_{n-1,1} & 0 \end{bmatrix}$$

and so

$$\widehat{X}_{n+1} = \begin{cases} 0 & \text{if } n = 1 \\ \sum_{i=1}^{n} \vartheta_{n,i}(X_{n+1-i} - \widehat{X}_{n+1-i}) & \text{if } n \geq 2 \end{cases}$$

i.e. $\widehat{X}_{n+1}$ is a linear combination of $X_1 - \widehat{X}_1$, $X_2 - \widehat{X}_2$, $\ldots$, $X_n - \widehat{X}_n$, the one-step ahead prediction errors which are **uncorrelated.**

Recursion

- Initialize

$$\boxed{v_0 = \tilde{\gamma}_X(1,1) = \mathsf{E}\left(X_1^2\right)}$$

- Recursion at step n

$$\boxed{\vartheta_{n,n-k} = \frac{1}{v_k}\left[\tilde{\gamma}_X(n+1,k+1) - \sum_{j=0}^{k-1} \vartheta_{k,k-j}\vartheta_{n,n-j}v_j\right]}$$

for $0 \leq k < n$

$$\boxed{v_n = \tilde{\gamma}_X(n+1, n+1) - \sum_{j=0}^{n-1} \vartheta_{n,n-j}^2 v_j}$$

Compute: $v_0, \vartheta_{1,1}, v_1, \vartheta_{2,2}, \vartheta_{2,1}, v_2, \vartheta_{3,3}, \vartheta_{3,1}, v_3, \ldots$.

1.6.1. Wold Decomposition

A process $\{X_t\}$ is **deterministic** (or **predictable**) if, for all n, $X_n - \widehat{X}_n \overset{\text{M.S.}}{=} 0$ so that the prediction variance

$$\mathsf{E}\left((X_n - \widehat{X}_n)^2\right) = 0$$

Example 1.16

Suppose we have random variables A, B such that $\mathsf{E}(A) = \mathsf{E}(B) = 0$ and $\mathsf{Var}(A) = \mathsf{Var}(B) = 1$ and further $\mathsf{E}(AB) = 0$, that is A, B are uncorrelated.

Let $\{X_t\}$ be defined by

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

for some frequency $\omega \in (0, 2\pi)$. For each integer n,

$$X_n = A\cos(\omega n) + B\sin(\omega n)$$

But

$$\begin{aligned}\cos(\omega n) &= \cos(\omega(n-1))\cos(\omega) - \sin(\omega(n-1))\sin(\omega)\\ \sin(\omega n) &= \sin(\omega(n-1))\cos(\omega) + \cos(\omega(n-1))\sin(\omega),\end{aligned}$$

the double angle formula and therefore

$$X_n = 2\cos(\omega)X_{n-1} - X_{n-2}$$

Thus $\widehat{X}_n = 2\cos(\omega)X_{n-1} - X_{n-2}$ and $X_n - \widehat{X}_n \overset{\text{M.S.}}{=} 0$. Thus $\{X_t\}$ is a deterministic process.[34]

[34] Brockwell and Davis use this terminology, although it is still stochastic in nature.

The Wold decomposition states that if $\{X_t\}$ is stationary and non-deterministic, then $\{X_t\}$ can be written as

$$X_t \overset{\text{M.S.}}{=} \sum_{j=0}^{\infty} \psi_j Z_{t-j} + V_t$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, $\{V_t\}$ is predictable and $\{Z_t\}, \{V_t\}$ are uncorrelated.[35] Furthermore, $\psi_0 = 1, \sum_{j=1}^{\infty} \psi_j^2 < \infty$ with

$$\psi_j = \frac{\mathsf{E}\left(X_t Z_{t-j}\right)}{\mathsf{E}\left(Z_t^2\right)}, \quad j \geq 1$$

[35] Notice that the linear process is causal, so there is no contribution of the negative terms. The result goes beyond the ARMA case.

Chapter 2
ARMA models

Section 2.1: Basic properties

All indicates that we can combine the class of filters to get a richer class of models.

2.1.1. ARMA(p, q)

The **Autoregressive-Moving average model** of order (p, q) (or ARMA(p, q)) for time-series X_t specifies that $\{X_t\}$ is a solution of

$$\Phi(B)X_t = \Theta(B)Z_t$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ and

$$\Phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$
$$\Theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$$

Suppose we study the causal, invertible case, *i.e.*

$$\Phi(B) = \prod_{j=1}^{p}(1 - \xi_j B), \quad |\xi_j| < 1 \; \forall \; j$$
$$\Theta(B) = \prod_{i=1}^{q}(1 - \omega_i B), \quad |\omega_i| < 1 \; \forall \; i$$

Assume that $\Phi(B)$ and $\Theta(B)$ have no common factors *i.e.* all ξ_j are different from all ω_j (so that no cancellation is possible). Recall that $|\xi_j| \neq 1 \; \forall \; j$ imply stationarity, $|\xi_j| < 1 \; \forall \; j$ implies causal and $|\omega_i| < 1 \; \forall \; i$ implies invertible.

For a linear process representation, we can write

$$X_t = \Psi(B)Z_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j};$$

we must have for this representation to be valid

$$\Phi(B)\Psi(B) = \Theta(B)$$

that is $\Psi(Z) = \frac{\Theta(Z)}{\Phi(Z)}$ for arbitrary complex value Z. We can compute the $\psi_j' s$ by equating

coefficients, that is

$$(1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^p)(\psi_0+\psi_1 B+\psi_2 B^2+\cdots)=1+\theta_1 B+\theta 2B^2+\cdots+\theta_q B^q$$

Equating the coefficients, we get for

$$\begin{aligned} B^0: &\quad \psi_0=1 \\ B^1: &\quad \psi_1-\phi_1\psi_0=\theta_1 \Rightarrow \psi_1=\phi_1+\theta_1 \end{aligned}$$

$$B^j: \quad \begin{cases} \psi_j-\sum_{k=1}^{j}\phi_k\psi_{j-k}=\theta_j, & 0\le j\le \max(p,q+1) \\ \psi_j-\sum_{k=1}^{p}\phi_k\psi_{j-k}=0, & j\ge \max(p,q+1) \end{cases}$$

Therefore

$$\psi_j=\begin{cases} \theta_j+\sum_{k=1}^{j}\phi_k\psi_{j-k} & \text{if } 0\le j\le \max(p,q+1), \theta_0=1 \\ \sum_{k=1}^{p}\phi_k\psi_{j-k}=0 & \text{if } j\ge \max(p,q+1) \end{cases}$$

2.1.2. Autocovariance function

For the ACVF,

$$\gamma_X(h)=\mathsf{E}\left(X_t X_{t+h}\right)=\sigma^2\sum_{j=0}^{\infty}\psi_j\psi_{j+|h|}$$

from previous result.

This may be easy theoretically, but it may be difficult to do in practice. We will try to find another method to compute the autocovariance.

Example 2.1 (Autocovariances of ARMA(1,1))
Consider an ARMA(1,1), $|\phi|<1$ and the model

$$(1-\phi B)X_t=(1+\theta B)Z_t$$

or equivalently

$$X_t-\phi X_{t-1}=Z_t+\theta Z_{t-1}.$$

By the previous formulation,

$$\psi_j = \begin{cases} 1 & \text{if } j = 0 \\ (\theta + \phi)\phi^{j-1} & \text{if } j \geq 1 \end{cases}$$

We can compute using the previous formula

$$\begin{aligned} \gamma_X(0) &= \sigma^2 \sum_{j=0}^{\infty} \psi_j^2 \\ &= \sigma^2 \left[1 + \sum_{j=1}^{\infty} \psi_j^2 \right] \\ &= \sigma^2 \left[1 + (\theta + \phi)^2 \sum_{j=1}^{\infty} (\phi^{j-1})^2 \right] \\ &= \sigma^2 \left[1 + \frac{(\theta + \phi)^2}{1 - \phi^2} \right] \end{aligned}$$

and

$$\begin{aligned} \gamma_X(1) &= \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+1} \\ &= \sigma^2 \left[(\theta + \phi) + (\theta + \phi)^2 \phi \sum_{j=1}^{\infty} (\phi^{j-1})^2 \right] \\ &= \sigma^2 (\theta + \phi) \left[1 + \frac{(\theta + \phi)\phi}{1 - \phi^2} \right] \end{aligned}$$

and similarly,

$$\begin{aligned} \gamma_X(h) &= \sigma^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+h} \\ &= \phi \, \gamma_X(h-1) \\ &= \phi^{h-1} \gamma_X(1), \quad h \geq 2 \end{aligned}$$

For ARMA(p, q), direct computation may be more tractable.

Example 2.2

Consider an ARMA(3,1) case, which is of the form

$$X_t - \phi_1 X_{t-1} - \phi_2 X_{t-2} - \phi_3 X_{t-3} = Z_t - \theta_1 Z_{t-1}. \tag{2.9}$$

We wish to compute $\gamma_X(h)$ for $h \in \mathbb{Z}^+$. First write $X_t = \Psi(B)Z_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}$

1. Multiply through by X_t, take expectations in (2.9).
   On the left hand side, we have

$$\begin{aligned}&\mathsf{E}\left(X_t^2\right) - \phi_1 \mathsf{E}\left(X_t X_{t-1}\right) - \phi_2 \mathsf{E}\left(X_t X_{t-2}\right) - \phi_3 \mathsf{E}\left(X_t X_{t-3}\right)\\ &= \gamma_X(0) - \phi_1\gamma_X(1) - \phi_2\gamma_X(2) - \phi_3\gamma_X(3)\end{aligned}$$

   while on the right hand side,

$$\mathsf{E}\left(X_t Z_t\right) = \mathsf{E}\left(\left(\sum_{j=0}^{\infty} \psi_j Z_{t-j}\right) Z_t\right) = \sum_{j=0}^{\infty} \psi_j \mathsf{E}\left(Z_{t-j} Z_t\right) = \sigma^2\psi_0$$

$$\mathsf{E}\left(X_t Z_{t-1}\right) = \sum_{j=0}^{\infty} \mathsf{E}\left(Z_{t-j} Z_{t-1}\right) = \sigma^2\psi_1$$

2. Multiply through by X_{t-1}, take expectations. The left hand side is

$$\gamma_X(1) - \phi_1\gamma_X(0) - \phi_2\gamma_X(1) - \phi_3\gamma_X(2)$$

   and the right hand side

$$\begin{aligned}\mathsf{E}\left(X_{t-1} Z_t\right) &= 0\\ \mathsf{E}\left(X_{t-1} Z_{t-1}\right) &= \sigma^2\psi_0\end{aligned}$$

   and the right hand side is $\sigma^2\theta_1\psi_0$. We do not have yet enough information.

3. Multiply by X_{t-2}, take expectations. On the right hand side, $\mathsf{E}\left(X_{t-2} Z_t\right) = \mathsf{E}\left(X_{t-2}, Z_{t-1}\right) = 0$. Therefore, the equation becomes

$$\gamma_X(2) - \phi_1\gamma_X(1) - \phi_2\gamma_X(0) - \phi_3\gamma_X(1)$$

4. Multiply by X_{t-3}, take expectations, which leaves us with

$$\gamma_X(3) - \phi_1\gamma_X(2) - \phi_2\gamma_X(1) - \phi_3\gamma_X(0) = 0$$

   We now have four equations and four unknowns $\gamma_X(0), \gamma_X(1), \gamma_X(2), \gamma_X(3)$ which we can solve simultaneously. For X_{t-k}, where $k \geq 4$ and

$$\gamma_X(k) - \phi_1\gamma_X(k-1) - \phi_2\gamma_X(k-2) - \phi_3\gamma_X(k-3) = 0$$

   for $k \geq 4$.

Example 2.3 (Autocovariances of ARMA(1,3))
Consider an ARMA(1,3) model of the form

$$X_t - \phi_1 X_{t-1} = Z_t + \theta_1 Z_{t-1} + \theta_2 Z_{t-2} + \theta_3 Z_{t-3}$$

Premultiply by X_{t-k} and take expectations. Using the same procedure as before, for $k = 0$, we have

$$\gamma_X(0) - \phi_1 \gamma_X(1) = \sigma^2(\psi_0 + \theta_1\psi_1 + \theta_2\psi_2 + \theta_3\psi_3)$$

and for $k = 1$

$$\gamma_X(1) - \phi_1 \gamma_X(0) = \sigma^2(\theta_1\psi_0 + \theta_2\psi_1 + \theta_3\psi_2)$$

and iterating this procedure, we will not this time get a homogeneous equation (the right hand side doesn't vanish). For $k = 2$, we have

$$\gamma_X(2) - \phi_1 \gamma_X(1) = \sigma^2(\theta_2\psi_0 + \theta_3\psi_1),$$

and subsequently

$$\begin{aligned}
k = 3: &\quad \gamma_X(3) - \phi_1\gamma_X(2) = \sigma^2\theta_3\psi_0 \\
k = 4: &\quad \gamma_X(4) - \phi_1\gamma_X(3) = 0 \\
k \geq 5: &\quad \gamma_X(j) - \phi_1\gamma_X(k-1) = 0
\end{aligned}$$

We can compute the autocovariances numerically by substitution.

This can be used in a variety of examples.
Example 2.4 (Autocovariances of ARMA(2,1))

$$\begin{aligned}
k = 0: &\quad \gamma_X(0) - \phi_1\gamma_X(1) - \phi_2\gamma_X(2) = \sigma^2(\psi_0 + \theta_1\psi_1) \\
k = 1: &\quad \gamma_X(1) - \phi_1\gamma_X(0) - \phi_2\gamma_X(1) = \sigma^2\theta_1\psi_0 \\
k = 2: &\quad \gamma_X(2) - \phi_1\gamma_X(1) - \phi_2\gamma_X(0) = 0 \\
k = 3: &\quad \gamma_X(3) - \phi_1\gamma_X(2) - \phi_2\gamma_X(1) = 0
\end{aligned}$$

using the fact that $\gamma(k) = \gamma(-k)$ in the previous calculations.

The case of an ARMA(1,2) is left as an exercise. The calculation and the values of the autocovariances are necessary to compute the likelihood.

For the general ARMA case, say ARMA(p, q) where the form of the equation is

$$X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = Z_t + \sum_{j=1}^{q} \theta_j Z_{t-j}$$

Multiply by X_{t-k} and take expectations. From the previous calculations, the left hand side is

$$\gamma_X(k) - \sum_{j=1}^{p} \phi_j \gamma_X(|k-j|)$$

and on the right hand side,

$$\mathsf{E}\left(X_{t-k} Z_{t-j}\right) = \begin{cases} 0 & \text{if } k > j \\ \sigma^2 \psi_{j-k} & \text{if } k \leq j. \end{cases}$$

Write $\theta_0 = 1$; the the RHS becomes for $k \leq q$

$$\sigma^2 \sum_{j=k}^{q} \theta_j \psi_{j-k}$$

that is for $k = 0, 1, 2, \ldots, \max(p, q+1)$, we have that the LHS

$$\begin{aligned} &\gamma_X(k) - \phi_1 \gamma_X(k-1) - \cdots - \phi_p \gamma_X(k-p) \\ &= \begin{cases} \sigma^2 \sum_{j=k}^{q} \theta_j \psi_{j-k} & \text{if } k \leq q \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

and for $k < \max(p, q+1)$

$$\gamma_X(k) = \phi_1 \gamma_X(k-1) + \cdots + \phi_p \gamma_X(k-p)$$

and we solve the linear system of the first $\max(p, q+1) + 1$ equations, then use the final recursion.

Note

Analytical solution is possible, and is complicated, relies on the mathematics of difference equations and rely on the roots of the inverse polynomial (with boundary values).

In R, there is a function that allows you to do this fairly routine calculation.

2.1.3. Autocovariance Generating function (ACVGF)

The ACVGF is another tool for computing the ACVF. Suppose $\{X_t\}$ is defined by

$$X_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$$

with $\{Z_t\} \sim \text{WN}(0, \sigma^2)$. Suppose that there exists $r > 1$ such that $\sum_{j=-\infty}^{\infty} \psi_j z^j$ is convergent for $z \in \mathbb{C}$, $\frac{1}{r} < |z| < r$.

Define the ACVGF, G_X, by $G_X(z) = \sum_{h=-\infty}^{\infty} \gamma_X(h) z^h$. Now, we know from previous results that

$$\gamma_X(h) = \mathsf{Cov}(X_t, X_{t+h}) = \sigma^2 \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+|h|}$$

therefore

$$\begin{aligned}
G_X(z) &= \sigma^2 \sum_{h=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \psi_j \psi_{j+|h|} z^h \\
&= \sigma^2 \left[\sum_{j=-\infty}^{\infty} \psi_j^2 + \sum_{h=1}^{\infty} \psi_j \psi_{j+h} (z^h + z^{-h}) \right] \\
&= \sigma^2 \left(\sum_{j=-\infty}^{\infty} \psi_j z^j \right) \left(\sum_{h=-\infty}^{\infty} \psi_h z^{-h} \right) \\
&= \sigma^2 \Psi(z) \Psi(z^{-1})
\end{aligned}$$

where $\Psi(z) = \sum_{j=-\infty}^{\infty} \psi_j z^j$. For ARMA$(p, q)$, $\Phi(B) X_t = \Theta(B) Z_t$. For stationary processes, we have $\Psi(B) = \prod_{j=1}^{p} (1 - \xi_j B)$ with $|\xi_j| \neq 1$ for all j. Therefore, $\Psi(z) = \sum_{j=1}^{p} (1 - \xi_j z) \neq 0$ when $|z| = 1$. We now have

$$X_t = \Psi(B) Z_t$$

with $\Psi(B) = \sum_{j=-\infty}^{\infty} \psi_j B^j$ where $\Phi(z)\Psi(z) = \Theta(z)$, $\forall\ z$. This means that $\Psi(z) = \Theta(Z)/\Phi(z)$ is well-defined since there is a neighborhood around 1 when z is in the annulus A_r for some r and where $A_r = \{z : \frac{1}{r} < |z| < r\}$. For the ARMA$(p, q)$, we have

$$G_X(z) = \sigma^2 \Psi(z) \Psi(z^{-1}) = \frac{\sigma^2 \Theta(z) \Theta(z^{-1})}{\Phi(z) \Phi(z^{-1})}$$

A series expansion of $G_X(z)$ in z has coefficients $\gamma_X(h)$ for $h = 0, \pm 1, \pm 2, \ldots$.

Example 2.5

Let $Z_t \sim \text{WN}(0, \sigma^2)$, then $G_Z(z) = \sum_{h=-\infty}^{\infty} \gamma_X(h) z^h = \sigma^2$ and thus $G_Z(z)$ is constant, does not depend on z and is the only generating process for which the ACVGF is constant[36].

Converting non-causal/non-invertible processes

Let $X_t \sim \text{ARMA}(p, q)$, that is $\Phi(B)X_t = \Theta(B)Z_t$ where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ and

$$\Phi(z) = \prod_{j=1}^{p}(1 - \xi_j z)$$
$$\Theta(z) = \prod_{j=1}^{q}(1 - \omega_j z)$$

where $\Phi(z) \neq 0$ and $\Theta(z) \neq 0$ when $|z| = 1$. and where we relax the assumption that all ξ are less than one in modulus; we allow

$$0 < \xi_1 \leq \xi_2 \leq \cdots \leq \xi_r < 1 < \xi_{r+1} \leq \cdots \leq \xi_p$$
$$0 < \omega_1 \leq \omega_2 \leq \cdots \leq \omega_s < 1 < \omega_{s+1} \leq \cdots \leq \omega_q$$

that is $\{X_t\}$ is **non-causal** if $r < p$ and **non-invertible** is $s < q$. However, define

$$\Phi^*(z) = \Phi(z) \prod_{j=r+1}^{p} \left(\frac{1 - \xi_j^{-1} z}{1 - \xi_j z} \right)$$
$$\Theta^*(z) = \Theta(z) \prod_{j=s+1}^{q} \left(\frac{1 - \omega_j^{-1} z}{1 - \omega_j z} \right)$$

Let $\{Z_t^*\}$ be defined by

$$\begin{aligned} Z_t^* &= \Phi^*(B) \{\Theta^*(B)\}^{-1} X_t \\ &= \Phi^*(B) \{\Theta^*(B)\}^{-1} \Theta^*(B) \{\Phi^*(B)\}^{-1} Z_t \\ &= \prod_{j=r+1}^{p} \left(\frac{1 - \xi_j^{-1} B}{1 - \xi_j B} \right) \left[\prod_{j=s+1}^{q} \left(\frac{1 - \omega_j^{-1} B}{1 - \omega_j B} \right) \right]^{-1} Z_t \\ &= \Psi^*(B) Z_t \end{aligned}$$

[36] Identifies Z up to mean-square

Therefore $G_{Z^*}(z) = \sigma^2\Psi^*(z)\Psi^*(z^{-1})$, therefore we are left with

$$\sigma^2\left\{\prod_{j=r+1}^{p}\frac{1-\xi_j^{-1}z}{1-\xi_j z}\right\}\left\{\prod_{j=s+1}^{q}\frac{1-\omega_j z}{1-\omega_j^{-1}z}\right\}\left\{\prod_{j=r+1}^{p}\frac{1-\xi_j^{-1}z^{-1}}{1-\xi_j z^{-1}}\right\}\left\{\prod_{j=s+1}^{q}\frac{1-\omega_j z^{-1}}{1-\omega_j^{-1}z^{-1}}\right\}.$$

But

$$\left(\frac{1-\xi_j^{-1}z}{1-\xi_j z}\right)\left(\frac{1-\xi_j^{-1}z^{-1}}{1-\xi_j z^{-1}}\right) = |\xi_j|^{-2}$$

for each j and similarly

$$\left(\frac{1-\omega_j z}{1-\omega_j^{-1}z}\right)\left(\frac{1-\omega_j z^{-1}}{1-\omega_j^{-1}z^{-1}}\right) = |\omega_j|^{2}$$

and so

$$G_{Z^*}(z) = \sigma^2\left(\prod_{j=r+1}^{p}|\xi_j|^{-2}\right)\left(\prod_{j=s+1}^{q}|\omega_j|^{2}\right) = \sigma^{*2}$$

and we therefore conclude that $\{Z_t^*\} \sim \text{WN}(0, \sigma^{*2})$.[37] Therefore $\Phi^*(B)X_t = \Theta^*(B)Z_t^*$ where $\{Z_t^*\} \sim \text{WN}(0, \sigma^{*2})$ therefore $\{X_t\} \sim \text{ARMA}(p, q)$ defined with respect to $\{Z_t^*\}$. But note

$$\Phi^*(B) = \prod_{j=1}^{r}(1-\xi_j B)\prod_{j=r+1}^{p}(1-\xi_j^{-1}B)$$

and all roots are less than one in modulus, therefore this representation is causal. Similarly,

$$\Theta^*(B) = \prod_{j=1}^{s}(1-\omega_j B)\prod_{j=s+1}^{q}(1-\omega_j^{-1}B)$$

therefore this process is invertible, since all the roots lie inside the unit circle.

Example 2.6

Let $\{X_t\} \sim \text{MA}(1)$ of the form

$$X_t = Z_t - 2Z_{t-1} = (1-2B)Z_t$$

[37] For this argument to work, we need stationarity to hold, for the expansion to be valid. In the case where the original process was both non-causal and non-invertible, the resulting variance may be smaller or lower, depending on the roots, how far they are from the unit circle

Let $Z_t^* = \left(1 - \frac{1}{2}B\right)^{-1}(1-2B)Z_t$, that is

$$\left(1 - \frac{1}{2}B\right) Z_t^* = (1-2B)Z_t = X_t$$

where Z_t^* is an ARMA(1,1) process in terms of Z_t. We could rewrite this as

$$\tilde{\Phi}(B)Z_t^* = \tilde{\Theta}(B)Z_t$$

where $\tilde{\Phi}(z) = 1 - \frac{1}{2}z$ and $\tilde{\Theta}(z) = 1 - 2z$ and therefore $Z_t^* = \Psi^*(B)Z_t$ where

$$\tilde{\Psi}(z) = \frac{\tilde{\Theta}(z)}{\tilde{\Phi}(z)} = \frac{1-2z}{1-\frac{1}{2}z}$$

and therefore

$$G_{Z^*}(z) = \sigma^2\tilde{\Psi}(z)\tilde{\Psi}(z^{-1}) = \sigma^2\left(\frac{1-2z}{1-\frac{1}{2}z}\right)\left(\frac{1-2z^{-1}}{1-\frac{1}{2}z^{-1}}\right) = 4\sigma^2$$

In the general definition, this correspond to the case $q = 1, s = 0, \omega_1 = 2$ and $Z_t^* \sim \text{WN}(0, 4\sigma^2)$ and

$$X_t = \left(1 - \frac{1}{2}B\right) Z_t^* = Z_t^* - \frac{1}{2}Z_{t-1}^*$$

so $\{X_t\}$ is invertible with respect to $\{Z_t^*\}$.

The morale of this story is that for any non-causal and/or non-invertible, a legitimate formulation can be found such that X_t is causal and invertible. Therefore, all the restrictions imposed earlier that restricted the roots of the $\Theta(B)$ and $\Phi(B)$ to be less than one in modulus are really for ease and that these are the only cases we need to consider.

Partial autocorrelation

Calculations for ARMA(p, q) follow the Levinson-Durbin general algorithm: no more transparent calculations are available.

2.1.4. Forecasting for ARMA(p, q)

Forecasting is possible using the **innovations algorithm** based on the ACV sequence.

Section 2.2: Estimation and model selection for ARMA(p,q)

2.2.1. Moment-based estimation

For general stationary processes, moment-based estimation of $\mu_x, \gamma_X(h), \rho_X(h), \sigma^2$ is used. For ARMA(p,q), we seek to estimate $(\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q)$ and σ^2 which then determine $\gamma_X(h)$. An elementary form of estimation involves finding $(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)$ such that the matrix $\boldsymbol{\Gamma}_X(n, \boldsymbol{\phi}, \boldsymbol{\theta})$ (the $n \times n$ covariance matrix implied by $(\boldsymbol{\phi}, \boldsymbol{\theta})$) most closely matches $\widehat{\boldsymbol{\Gamma}}_n$ (the sample covariance matrix) that is

$$(\widehat{\boldsymbol{\phi}}, \widehat{\boldsymbol{\theta}}) = \arg\min_{\boldsymbol{\phi}, \boldsymbol{\theta}} d(\boldsymbol{\Gamma}_X(n, \boldsymbol{\phi}, \boldsymbol{\theta}), \widehat{\boldsymbol{\Gamma}}_n)$$

this may not be straightforward.

We now target the estimation of parameters from the data, which we believe is a realization of the series.

Yule-Walker Method for AR(p)

Suppose $\{X_t\} \sim \text{AR}(p)$, which means $\Phi(B)X_t = Z_t$ or

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + Z_t$$

Using our previous strategy, multiply through by X_{t-k} and take expectations.

$$\mathsf{E}\left(X_t X_{t-k}\right) = \sum_{j=1}^{p} \phi_j \mathsf{E}\left(X_{t-j} X_{t-k}\right)$$

which gives us the equation

$$\gamma_X(k) = \sum_{j=1}^{p} \phi_j \gamma_X(|j-k|)$$

When $k = 0, 1, \ldots, p-1$, are considered, we have p (linear) equations in $\phi_1, \ldots, \phi_p$ and $\gamma_X(h), h = 0, 1, 2, \ldots$, which we may solve analytically.

- substituting $\widehat{\gamma}_X(h)$ for $\gamma_X(h)$ yields the estimates $\widehat{\phi}_1, \widehat{\phi}_2, \ldots, \widehat{\phi}_p$.

Note

1. For the AR(p), we could use OLS and form the design matrix with the lagged values and perform the regression on this matrix. However, the Yule-Walker approach guarantees a stationary and causal solution.

2. We can choose to look at $k = 1, \ldots, p$ to derive the Yule-Walker equations: in this case, we must solve

$$\boldsymbol{\Gamma}_X(p)\boldsymbol{\phi}_{1:p} = \boldsymbol{\gamma}_X(p, 1)$$

as in the case of optimal forecasting. We also have that

$$\sigma^2 = \gamma_X(0) - \boldsymbol{\phi}_{1:p}^\top \boldsymbol{\gamma}_X(p, 1)$$

so $\widehat{\sigma^2}$ is obtained by plugging in $\widehat{\gamma}_X$ and $\widehat{\boldsymbol{\phi}}_{1:p}$. [38]

3. This strategy also works for the ARMA(p, q). We solve

$$\gamma_X(k) - \phi_1\gamma_X(k-1) - \cdots - \phi_p\gamma_X(k-p) = \sigma^2 \sum_{j=k}^{q} \theta_j\psi_{j-k}$$

for $0 \leq k \leq p+q$ as we need $p+q+1$ equations to solve for the $p+q+1$ unknowns – but this is non-linear in $(\phi_1, \ldots \phi_p, \theta_1, \theta_q, \sigma^2)$ as $\{\psi_j\}$ are the coefficients in the MA(∞) representation of $\{X_t\}$.

- Burg's Algorithm for AR(p) This is an algorithm that uses the sample PACFs to estimate $\phi_1, \ldots, \phi_p$.
- For the MA(q), moment based estimation can be achieved using the Innovations algorithm.

2.2.2. Maximum Likelihood Estimation

In order to do this, we need parametric assumptions in order to do exact MLE, we need to make assumptions about the series. ML estimation is the most efficient method of inference in the "regular" case, [39] but it requires parametric assumptions. [40]

Gaussian case: Suppose $\{X_t\}$ is a Gaussian time-series with zero mean and autocovariance given by

$$\kappa(i, j) = \tilde{\gamma}_X(i, j) = \mathsf{E}\left(X_i X_j\right)$$

that is $X_{1:n} \sim \mathcal{N}_n(\mathbf{0}_n, \boldsymbol{\kappa}_n)$, where $\boldsymbol{\kappa}_n$ is the $(n \times n)$ covariance matrix. If we make a second parametric assumption that $\{X_t\}$ follows an ARMA(p, q) process, then $\boldsymbol{\kappa}_n = \boldsymbol{\kappa}_n(\boldsymbol{\phi}, \boldsymbol{\theta})$ we

[38] Depending on the starting p equations, we get different estimates which get closer and closer to each other as the sample size increase.

[39] Meaning that they produce estimates that have the lowest variance, at least asymptotically

[40] Econometricians are not happy with the Gaussianity assumption, and they rather rely on different techniques, using asymptotically justified likelihood from CLT and so on.

need to calculate the inverse of $\boldsymbol{\kappa}_n$ and the determinant of $\boldsymbol{\kappa}_n$. We can use the Innovations algorithm in order to get a decomposition of $\boldsymbol{\kappa}$ which avoids direct calculations of the determinant and the inverse of $\boldsymbol{\kappa}$.

If $\widehat{X}_t$ is the one-step ahead forecast,

$$\widehat{X}_t = \mathsf{E}\left(X_t | X_{1:(t-1)}\right)$$

then as $X_t | X_{1:(t-1)} \sim \mathcal{N}(\mu_{t-1}, \sigma^2_{t-1})$ where

$$\begin{aligned}\mu_{t-1} &= \boldsymbol{\kappa}_{t,1:(t-1)} \boldsymbol{\kappa}_{t-1}^{-1} \boldsymbol{X}_{1:(t-1)} \\ \sigma^2_{t-1} &= \kappa_{t,t} - \boldsymbol{\kappa}_{t,1:(t-1)}, \boldsymbol{\kappa}_{t-1} \boldsymbol{\kappa}_{1:(t-1),t}\end{aligned}$$

where

$$\boldsymbol{\kappa}_t = \left[\begin{array}{c|c} \boldsymbol{\kappa}_{t-1} & \boldsymbol{\kappa}_{1:(t-1),t} \\ \hline \boldsymbol{\kappa}_{t,1:(t-1)} & \kappa_{t,t} \end{array}\right]$$

we have $\widehat{X}_t = \mu_{t-1}$. The likelihood is given by

$$\mathcal{L}(\boldsymbol{\kappa}_n(\boldsymbol{\phi}, \boldsymbol{\theta})) = (2\pi)^{-\frac{n}{2}} |\det \boldsymbol{\kappa}_n(\boldsymbol{\phi}, \boldsymbol{\theta})|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \boldsymbol{x}_{1:n}^\top \boldsymbol{\kappa}_n^{-1}(\boldsymbol{\phi}, \boldsymbol{\theta}) \boldsymbol{x}_{1:n}\right)$$

Direct optimization of the log likelihood $\log \mathcal{L}(\boldsymbol{\kappa}_n(\boldsymbol{\phi}, \boldsymbol{\theta}))$, denoted ℓ, is possible but computationally expensive when n is large as we need to compute

$$\det |\boldsymbol{\kappa}_n(\boldsymbol{\phi}, \boldsymbol{\theta})| \quad \text{and} \quad \boldsymbol{\kappa}_n^{-1}(\boldsymbol{\phi}, \boldsymbol{\theta})$$

However, the Innovations algorithm allows to achieve a decomposition. Recall that

$$\boldsymbol{C}_n = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ \vartheta_{1,1} & 0 & 0 & 0 & 0 & 0 \\ \vartheta_{2,2} & \vartheta_{2,1} & 0 & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & \\ \vartheta_{n-1,n-1} & \cdots & \cdots & \cdots & \vartheta_{n-1,1} & 0 \end{bmatrix}$$

and $\boldsymbol{X}_{1:n} = \boldsymbol{C}_n(\boldsymbol{X}_{1:n} - \widehat{\boldsymbol{X}}_{1:n})$ where the elements of $\boldsymbol{X}_{1:n} - \widehat{\boldsymbol{X}}_{1:n}$ are uncorrelated (by construction) and are normally distributed with covariance matrix $\boldsymbol{D}_n = \operatorname{diag}(v_0, v_1, \ldots, v_{n-1})$, which is the diagonal matrix of one step forecast variances. Thus, $\boldsymbol{\kappa}_n = \boldsymbol{C}_n \boldsymbol{D}_n \boldsymbol{C}_n^\top$ and

$\det|\boldsymbol{\kappa}_n| = \det|\boldsymbol{D}_n| = \prod_{t=1}^{n} v_{t-1} = v_0 v_1 v_2 \cdots v_{n-1}$ and

$$\boldsymbol{X}_{1:n}^{\top}\boldsymbol{\kappa}_n^{-1}\boldsymbol{X}_{1:n} = (\boldsymbol{X}_{1:n} - \widehat{\boldsymbol{X}}_{1:n})^{\top}\boldsymbol{D}_n^{-1}(\boldsymbol{X}_{1:n} - \widehat{\boldsymbol{X}}_{1:n}) = \sum_{t=1}^{n} \frac{(X_t - \widehat{X}_t)^2}{v_{t-1}}$$

Thus the likelihood $\mathcal{L}$ is

$$\mathcal{L}(\boldsymbol{\kappa}_n) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{(v_0 \cdot v_1 \cdots v_{n-1})}} \exp\left(-\frac{1}{2}\sum_{t=1}^{n} \frac{(X_t - \widehat{X}_t)^2}{v_{t-1}}\right).$$

Note

We have $v_t = \mathsf{E}\left((X_{t+1} - \widehat{X}_{t+1})^2\right) = \sigma_t^2 = \sigma^2 r_t$ and the likelihood can be rewritten as

$$\frac{1}{(2\pi)^{\frac{n}{2}}}\left(\sigma^2\right)^{-\frac{n}{2}} \prod_{t=1}^{n} r_t^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} S(\boldsymbol{X}_{1:n})\right).$$

where

$$S(\boldsymbol{X}_{1:n}) \equiv S(\boldsymbol{X}_{1:n}, \boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} \frac{(X_t - \widehat{X}_t(\boldsymbol{\phi}, \boldsymbol{\theta}))^2}{r_{t-1}(\boldsymbol{\phi}, \boldsymbol{\theta})}$$

the ML estimates of $(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2)$ are obtained by maximizing this likelihood.

For the non-Gaussian case, we may still use the function

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}, \sigma^2) = \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}} \prod_{t=1}^{n} \{r_{t-1}(\boldsymbol{\phi}, \boldsymbol{\theta})\}^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2} S(\boldsymbol{\phi}, \boldsymbol{\theta})\right)$$

as a mean of estimating the parameters where

$$S(\boldsymbol{\phi}, \boldsymbol{\theta}) = \sum_{t=1}^{n} \frac{(x_t - \widehat{x}_t(\boldsymbol{\phi}, \boldsymbol{\theta}))^2}{r_{t-1}(\boldsymbol{\phi}, \boldsymbol{\theta})}$$

Alternatively, least squares procedure based on

$$(\tilde{\boldsymbol{\phi}}, \tilde{\boldsymbol{\theta}}) = \arg\min S(\boldsymbol{\phi}, \boldsymbol{\theta})$$

with $\tilde{\sigma}^2 = S(\boldsymbol{\phi}, \boldsymbol{\theta})/(n - p - q)$ may also be used. [41] In R, arima allows you to select full ML, or a conditional ML, or least-squares or conditional LS (where the conditional analysis

[41] This is an inefficient procedure if we make the normality assumption. In finite samples, these two methods may be indistinguishable in the finite sample case.

conditions on the original $p+q$ data).[42]

2.2.3. Model selection

In considering model selection, we may prioritize

- fidelity to the observed data ("within-sample" criterion)
- model complexity
- residuals (structure)
- forecasting performance ("out-of-sample") [43]

Akaike Information Criterion (AIC)

The AIC is widely used in statistics; it trades off goodness-of-fit with model complexity. For model M with parameter β and likelihood $\mathcal{L}_M(\beta)$, the AIC is defined as

$$\text{AIC}(M, \boldsymbol{\beta}_M) = -2\ell_M(\widehat{\boldsymbol{\beta}}_M) + 2\dim(\boldsymbol{\beta}_M)$$

for example[44] for ARMA(p, q), we have $p+q+1$ parameters. The model with smallest AIC is selected as most appropriate. This should be used **only** to compare **nested models**; we fit ARMA(p, q) for p, q moderately large, then examine all submodels.

Note

- In small samples, the $2\dim(\boldsymbol{\beta}_M)$ penalty is not sufficiently stringent(*i.e.* it selects overly complex models).
- In large samples, the procedure is **inconsistent**.[45]

Bayesian Information Criterion (BIC)

This criterion adjusts the penalty to relieve both these issues. The BIC takes the form

$$\text{BIC}(M, \boldsymbol{\beta}_M) = -2\ell_M(\widehat{\beta}) + \log(n)\dim(\boldsymbol{\beta}_M)$$

[42] It is rather awkward to use full ML, since we need to condition x_1 on past values, while the conditional ML allows to start at the $p+q+1$ term.

[43] Forecast errors and the innovations algorithm gives a record of how the model is performing at this stage

[44] Note that increasing utility of adding parameters in the model always decreases as we add covariates, while the second term increases linearly. One can plot this as a function of the models, and the one that yields the minimum value is preferred in practice. For the BIC, the intercept will be higher, as $\log(n)$ will be linear in the dimension of the parameter β, for fixed n (the length of the time-series).

[45] Meaning that with probability one, the AIC does not select the model as the sample size becomes infinite.

where n is the length of the series. This has a more stringent penalty, replacing 2 by $\log(n)$ in the penalty term. This adjustment alleviates both the small sample underpenalization **and** the large sample inconsistency of the AIC – the BIC is a consistent model selection criterion.[46] Once the "best model" has been selected, the residuals from the fit should resemble a white noise series (constant variance, uncorrelated).[47]

[46] In this framework, note that consistency is really in terms of correctly choosing the best model from the nested models. This is also interesting in the context of model misspecification in the context of inference (when fitting a ML in an incorrect model, for example selecting from a set of models, which does not include the DGM.

[47] There are other suggestions in the Brockwell and Davis book, but they do not address large sample size problematics. The arima function in R has a slot with the residuals.

Chapter 3
Non-Stationary and Seasonal Models

Section 3.1: ARIMA models

If d is a non-negative integer, we call $\{X_t\}$ an ARIMA model of order (p, d, q) if the process $\{Y_t\}$ defined by

$$Y_t = (1 - B)^d X_t$$

is a causal ARMA(p, q) process, i.e. $\{X_t\}$ satisfies

$$\Phi(B)(1 - B)^d X_t = \Theta(B)Z_t \qquad Z_t \sim \text{WN}(0, \sigma^2). \tag{3.10}$$

$\{X_t\}$ is stationary if and only if $d = 0$. In all other cases ($d \geq 1$), we can add polynomial trend of order $d - 1$ and the resulting process still satisfies the ARIMA equation (3.10).

However, as $\Phi(Z) = 0$ has all reciprocal roots inside the unit circle, we merely need to difference $\{X_t\}$ d times and perform inference on the resulting differenced series $y_t = (1 - B)^d x_t$.

In practice, d is not known, so we may try $d = 1, 2, \ldots$ in turn and assess the stationarity of the resulting series. However, in practice, it can be difficult to distinguish a factor $(1 - B)$ from $(1 - \xi B)$ with $|\xi| \to 1$ in the ARMA polynomial.

Example 3.1

Consider the following processes, the AR(2) process defined by $(1-0.6B)(1-0.99B)X_t = Z_t$ versus the ARIMA$(1, 1, 0)$ $(1 - 0.6B)(1 - B)X_t = Z_t$. These processes look very similar in a simulation. The solution to the second equation is non-stationary. This is clear by the drift for larger periods looking at the graphs (see handout). Differencing when we shouldn't yields a more complicated process. We thus want a statistical way to distinguish these two models.

Section 3.2: Unit roots

The presence of a unit root ($|\xi| = 1$) in the AR polynomial fundamentally changes properties of the process, estimators, tests, etc.

Dickey-Fuller test

Suppose $\{X_t\} \sim$ AR(1) where

$$X_t - \mu = \phi(X_{t-1} - \mu) + Z_t$$

with $Z_t \sim \text{WN}(0, \sigma^2)$, if $|\phi| < 1$, then by standard theory for method of moments estimators, we have that

$$\sqrt{n}(\widehat{\phi} - \phi) \rightsquigarrow \mathcal{N}(0, 1 - \phi^2).$$

However, if $|\phi| = 1$, this asymptotic result doesn't hold. Suppose $\phi = 1$; then

$$\begin{aligned}(1 - B)X_t &= \mu(1 - \phi) + (\phi - 1)X_t + Z_t \\ \Leftrightarrow X_t &= \mu + \phi(X_{t-1} - \mu) + Z_t.\end{aligned}$$

Therefore,

$$(1 - B)X_t = \phi_0^* + \phi_1^* X_{t-1} + Z_t$$

where $\phi_0^* = \mu(1-\phi)$ and $\phi_1^* = \phi - 1$. As $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, we can estimate ϕ_0^*, ϕ_1^* via OLS: we regress the series $(1 - B)x_t = x_t - x_{t-1}$ on x_{t-1} – the forms of $\widehat{\phi}_0^*$ and $\widehat{\phi}_1^*$ are identical to the forms of the intercept and slope estimators in the OLS. The Wald-type test statistic derived by Dickey and Fuller is then

$$t_{\text{DF}} = \frac{\widehat{\phi}_1^*}{\widehat{\text{se}}(\widehat{\phi}_1^*)}$$

used to test the null hypothesis $H_0 : \widehat{\phi}_1^* = 0$ i.e. that $H_0 : \phi = 1$ and we are in the presence of a unit root.[48] Here

$$\widehat{\text{se}}(\widehat{\phi}_1^*) = \frac{\sum_{t=2}^n (x_t - x_{t-1} - \widehat{\phi}_0^* - \widehat{\phi}_2 x_{t-1})}{(n-3)\sqrt{\sum_{t=2}^n (x_{t-1} - \bar{x})^2}}$$

The null distribution of t_{DF} is non-standard (even asymptotically) – but for finite n, can be approximated using Monte-Carlo. The test of H_0 is carried out against the one-sided alternative $H_1 : \phi_1^* < 0$ (equivalently $H_1 :, \phi < 1$). See handout: the first three panels (with parameters close to 1, of 0.9,0.99 and 0.999 show asymptotic normality, where

$$\frac{\sqrt{n}(\widehat{\phi} - \phi)}{\sqrt{1 - \phi^2}} \sim \mathcal{N}(0, 2)$$

while the final panel shows the DF statistic distribution. Note that this test is available in R through the adf.test function.

[48] One still needs to write down the properties of the distribution of the test statistic, in terms of the Brownian motion.

For the AR(p), we have

$$X_t - \mu = \sum_{j=1}^{p} \phi_j (X_{t-j} - \mu) + Z_t$$

which imply

$$(1-B)X_t = \phi_0^* + \phi_1^* X_{t-1} + \sum_{j=2}^{p} \phi_j^* (1-B) X_{t-j+1} + Z_t$$

In this formulation we have

$$\begin{aligned}
\phi_0^* &= \mu(1 - \phi_1 - \cdots - \phi_p) \\
\phi_1^* &= \sum_{j=1}^{p} \phi_j - 1 \\
\phi_j^* &= -\sum_{k=j}^{p} \phi_k
\end{aligned}$$

for $k = 2, \ldots, p$.

If $\{X_t\} \sim \text{AR}(p)$ with one unit root, then $(1-B)X_t \sim \text{AR}(p-1)$ is stationary. Therefore, we can test for a unit root by testing

$$H_0 : \phi_1^* = 0$$

using $t_{\text{ADF}} = \phi_1^* / \widehat{\text{se}}(\phi_1^*)$ where the numerator and denominator are obtained from regressing $(1-B)x_t$ on x_{t-1} and $(x_{t-1} - x_{t-2}), (x_{t-2} - x_{t-3})$, etc. However, the null distribution is different from the AR(1) case. The ADF test function uses a look-up table to calibrate the test statistic. This is thus the **Augmented Dickey-Fuller** test.

Note

Further extensions of this test exist for models with trends.

In the case of multiple unit roots, sequential testing, applying the Dickey-Fuller procedure in each case, doing first differencing multiple times.

Note

Consider the model with $\phi(B) = (1 - \xi B)(1 - \bar{\xi} B)$, roots that appear in conjugate pair, *i.e.* the AR(2) model with reciprocal roots $\xi = \frac{1}{r} e^{-i\omega} = \frac{1}{r}\cos(\omega) - i\frac{1}{r}\sin(\omega)$ and $\bar{\xi} = \frac{1}{r} e^{i\omega}$

where $r > 1$. Then

$$\phi(B) = (1 - 2\Re(\xi B) + |\xi|^2 B^2) = \left(1 - \frac{2}{r}\cos(\omega)B + \left|\frac{1}{r}\right|^2 B^2\right)$$

This is an AR(2) model with ACF

$$\rho(h) = \frac{1}{r^h}\frac{\sin(h\omega + \lambda)}{\sin(\lambda)}$$

for $h = 0, \pm 1, \pm 2$ where

$$\lambda = \arctan\left[\frac{r^2+1}{r^2-1}\tan(\omega)\right]$$

that is $\rho(h)$ is a decaying sinusoidal function. If $r \to 1$ from above, then $\rho(h) \to \cos(\omega h)$. For $r > 1$, we see the following picture

Section 3.3: Seasonal ARIMA models (SARIMA)

Recall the lag-difference operator ∇_s

$$\begin{aligned}\nabla_s X_t &= X_t - X_{t-s} \\ &= X_t - B^s X_t \\ &= (1 - B^s)X_t.\end{aligned}$$

This operator allows us to construct seasonally non-stationary models; whereas the ARIMA model considers $(1-B)^d\Phi(B)X_t = \Theta(B)Z_t$, the seasonal ARIMA (SARIMA) allows $(1 - B^s)^d\Phi(B)X_t = \Theta(B)Z_t$. This is a form of non-stationary model, where $\Phi(B)X_t = \Theta(B)Z_t$ defines a stationary (causal) ARMA(p, q) process.

The general form of the SARIMA is as follows: For d, D non-negative integers and seasonality s, $\{X_t\}$ is a seasonal ARIMA process

$$\{X_t\} \sim \text{SARIMA}(p, d, q)(P, D, Q)_s$$

if $\{Y_t\}$ defined by

$$Y_t = (1 - B)^d(1 - B^s)^D X_t$$

upon commuting the operators, is a causal ARMA process determined by

$$\Phi(B)\breve{\Phi}(B^s)Y_t = \Theta(B)\breve{\Theta}(B^s)Z_t$$

where $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$

$$\Phi(z) = 1 - \phi_1 z - \cdots - \phi_p z^p$$
$$\breve{\Phi}(z) = 1 - \Phi_1 - \cdots - \Phi_P z^P$$
$$\Theta(z) = 1 - \theta_1 z - \cdots - \theta_q z^q$$
$$\breve{\Theta}(z) = 1 - \Theta_1 - \cdots - \Theta_Q z^Q$$

of order respectively p, P, q, Q that is $\{Y_t\} \sim \mathrm{ARMA}(p + sP, q + sQ)$. This is introduced to allow stochastic seasonality to be introduced. We can apply the inverse of the differencing (cumulate) to get the forecast for Y_t. Thus for inference and forecasting for seasonal ARIMA models proceeds by preprocessing $\{X_t\}$ to $\{Y_t\}$, then performing inference forecasting for $\{Y_t\}$, and then undoing the differencing and seasonal differencing. The introduction of differencing of different order changes radically the dataset, therefore comparing AIC and BIC, with different length dataset, should not be done in practice.

To implement this in R

- inference: use the arima function
- forecasting: forecast.Arima

from the "forecast" library in R.

Chapter 4
State-space models

By coupling two (stationary) processes, we can construct even more complicated models; for example, we can take the first process to be **latent**, thus describing hidden time-varying structure, whereas the second process is constructed conditional on process 1 and is used to represent the observed data.

We consider the **marginal** structure implied by this joint model for the observed data. In general, we will use vector-valued time series processes: $\{\boldsymbol{X}_t\} = (X_{t1}, \ldots, X_{tK})^\top$ is a $K \times 1$ vector-valued process that potentially exhibits

- autocorrelation (between elements)
- cross-correlation (across elements).

We have

$$\gamma_{ij}(t, s) = \mathsf{Cor}(X_{ti}, X_{sj})$$

describing the auto and cross-correlation. For stationary processes, $\boldsymbol{\mu} = \mathsf{E}(\boldsymbol{X}_t)$ is a $K \times 1$ vector and $\boldsymbol{\Gamma}(h) = [\gamma_{ij}(h)]_{i=1,j=1}^{K}$ a $K \times K$ matrix, where $\gamma_{ij}(h)$ is the autocovariance for the i^{th} component and $\gamma_{ij}(h) = \gamma_{ji}(-h)$. The extension to vector-valued time series is not very complicated. The extension of the white-noise process is the following: $\{\boldsymbol{W}_t\} \sim \text{WN}(\boldsymbol{0}, \boldsymbol{\Sigma}_t)$ where $\boldsymbol{\Sigma}_t$ is a $K \times K$ matrix such that $\mathsf{E}(\boldsymbol{W}_t) = \boldsymbol{0}_K$ and $\mathsf{E}\left(\boldsymbol{W}_t \boldsymbol{W}_s^\top\right) = \boldsymbol{\Sigma}_t$ if $s = t$ and zero otherwise.

Section 4.1: State-Space Formulation

The linear state space model for $\{\boldsymbol{Y}_t\}$ of dimension $K \times 1$ is specified by two relations that together form a **dynamic linear model**.

We have the **observation equation**

$$\boldsymbol{Y}_t = \boldsymbol{G}_t \boldsymbol{X}_t + \boldsymbol{W}_t, \qquad t = 0, \pm 1, \pm 2, \ldots$$

where $\{\boldsymbol{G}_t\}$ is a sequence of $K \times L$ deterministic matrices, and where $\{\boldsymbol{W}_t\} \sim \text{WN}(\boldsymbol{0}, \boldsymbol{\Sigma}_t)$ that is $\boldsymbol{Y}_t$ is a linear combination of elements of $\boldsymbol{X}_t$ and the **state equation**

$$\boldsymbol{X}_{t+1} = \boldsymbol{F}_t \boldsymbol{X}_t + \boldsymbol{V}_t, \qquad t = 0, \pm 1, \pm 2, \ldots$$

where $\{\boldsymbol{F}_t\}$ is a sequence of $L \times L$ deterministic matrices and where $\boldsymbol{V}_t \sim \text{WN}(\boldsymbol{0}, \boldsymbol{\Omega}_t)$. [49]

To complete the model, it is normal to consider $t \geq 1$ and setting $\boldsymbol{X}_1$ as a random vector, uncorrelated with the residuals $\{\boldsymbol{W}_t\}, \{\boldsymbol{V}_t\}$.

Note

1. Typically, the white-noise series are uncorrelated: $\mathsf{E}\left(\boldsymbol{W}_t \boldsymbol{V}_s^\top\right) = 0 \; \forall \; s, t$.
2. If we require correlation between $\{\boldsymbol{W}_t\}$ and $\{\boldsymbol{V}_s\}$, a amended formulation can be proposed, for example with cross-correlation and drift terms, we might write the observation and state equation respectively as

$$\begin{aligned} \boldsymbol{Y}_t &= \boldsymbol{G}_t \boldsymbol{X}_t + \boldsymbol{d}_t + \boldsymbol{H}_{1t} \boldsymbol{Z}_t \\ \boldsymbol{X}_{t+1} &= \boldsymbol{F}_t \boldsymbol{X}_t + \boldsymbol{e}_t + \boldsymbol{H}_{2t} \boldsymbol{Z}_t \end{aligned}$$

   where $\{\boldsymbol{Z}_t\}$ is WN($\boldsymbol{0}$ of dimension $(K + L) \times 1, \boldsymbol{\Sigma}_t$) process, $\boldsymbol{H}_{1t}$ is of dimension $(K \times (K + L))$, $\boldsymbol{H}_{2t}$ is of dimension $(L \times (K + L))$ and $\boldsymbol{d}_t, \boldsymbol{e}_t$ are deterministic "drift" terms.
3. With careful definition of $\{\boldsymbol{X}_t\}$, one can construct correlated process with more general correlation structure that is implied by the AR(1)-type form.
4. The extension of this formulation to non-linear state space models is possible, but more challenging in terms of inference.

For the simple model, we have that

$$\begin{aligned} \boldsymbol{X}_t &= \boldsymbol{F}_t \boldsymbol{X}_{t-1} + \boldsymbol{V}_{t-1} \\ &= \boldsymbol{F}_t (\boldsymbol{F}_{t-1} \boldsymbol{X}_{t-2} + \boldsymbol{V}_{t-2}) + \boldsymbol{V}_{t-1} \\ &= \boldsymbol{F}_t \boldsymbol{F}_{t-1} \boldsymbol{X}_{t-2} + \boldsymbol{F}_t \boldsymbol{V}_{t-2} + \boldsymbol{V}_{t-1} \\ &= \boldsymbol{F}_t \boldsymbol{F}_{t-1} (\boldsymbol{F}_{t-2} \boldsymbol{X}_{t-3} + \boldsymbol{V}_{t-3}) + \boldsymbol{F}_t \boldsymbol{V}_{t-2} + \boldsymbol{V}_{t-1} \\ &\dots \\ &= f_t(\boldsymbol{X}_1, \boldsymbol{V}_1, \dots, \boldsymbol{V}_{t-1}) \end{aligned}$$

and similarly, $\boldsymbol{Y}_t = g_t(\boldsymbol{X}_1, \boldsymbol{V}_1, \dots, \boldsymbol{V}_{t-1}, \boldsymbol{W}_t)$.

Note

For $t > s$,

$$\mathsf{E}\left(\boldsymbol{V}_t \boldsymbol{X}_s^\top\right) = \mathsf{E}\left(\boldsymbol{V}_t \boldsymbol{Y}_s^\top\right) = 0$$

[49] We will be able to marginalize over the parameters for $\boldsymbol{X}_t$. We usually take $\boldsymbol{G}_t, \boldsymbol{F}_t$ as constant over time, otherwise inference may be impossible due to the limited amount of data points.

and similarly

$$\mathsf{E}\left(\boldsymbol{W}_t\boldsymbol{X}_s^\top\right) = \mathsf{E}\left(\boldsymbol{W}_t\boldsymbol{Y}_s^\top\right) = 0$$

as for the ARMA models.

Example 4.1

Let $\{Y_t\} \sim \text{AR}(1)$ with $Y_t = \phi Y_{t-1} + Z_t$ and $Z_t \sim \text{WN}(0, \sigma^2), |\phi| < 1$. Let $F_t \equiv [\phi]\ \forall\ t$, the 1×1 matrix. and let

$$X_{t+1} = \phi X_t + V_t$$
$$X_1 = Y_1 = \sum_{j=0}^{\infty} \phi^j Z_{1-j}$$

In this example, the observation equation is

$$Y_t = G_t X_t + W_t, \qquad G_t = [1]\ \forall\ t, \qquad W_t \equiv 0\ \forall\ t$$

and the state equation is given by

$$X_{t+1} = F_t X_t + V_t, \qquad F_t \equiv [\phi]\ \forall\ t, \qquad V_t = Z_{t+1}$$

Example 4.2 (State-space formulation of $\{Y_t\} \sim \text{AR}(p)$)

Let $\{Y_t\}$ be a causal AR(p) time series process with respect to $\{Z_t\} \sim \text{WN}(0, \sigma^2)$, that is

$$Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p} = Z_t$$

and let $\boldsymbol{X}_t = (Y_{t-p+1}, Y_{t-p+2}, \cdots, Y_t)^\top$ a $(p \times 1)$ vector. The observation equation is given by

$$Y_t = \boldsymbol{G}_t \boldsymbol{X}_t + W_t, \qquad \boldsymbol{G}_t = (0, 0, \ldots, 0, 1)\ \forall\ t, \qquad W_t \equiv 0\ \forall\ t$$

and the state equation by

$$X_{t+1} = \boldsymbol{F}_t X_t + V_t$$

which can be written as

$$\begin{bmatrix} Y_{t-p+2} \\ Y_{t-p+3} \\ \vdots \\ Y_{t+2} \\ Y_{t+1} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \phi_p & \phi_{p-1} & \phi_{p-2} & \cdots & \phi_1 \end{bmatrix} \begin{bmatrix} Y_{t-p+1} \\ Y_{t-p+2} \\ \vdots \\ Y_{t+1} \\ Y_t \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} Z_{t+1}$$

Example 4.3

Let $\{Y_t\} \sim$ ARMA$(1,1)$ process, with

$$(1 - \phi B)Y_t = (1 + \theta B)Z_t$$

where $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ and $|\phi|, |\theta| < 1$. We could rewrite this as $Y_t = \phi Y_{t-1} + Z_t + \theta Z_{t-1}$. Let now $X_{t+1} = \phi X_t + Z_{t+1}$ (an AR(1) process), so that

$$\begin{bmatrix} X_t \\ X_{t+1} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & \phi \end{bmatrix} \begin{bmatrix} X_{t-1} \\ X_t \end{bmatrix} + \begin{bmatrix} 0 \\ Z_{t+1} \end{bmatrix}$$

i.e. $\boldsymbol{X}_{t+1} = \boldsymbol{F}_t \boldsymbol{X}_t + \boldsymbol{V}_t$. Then the observation equation is

$$Y_t = \boldsymbol{G}_t \boldsymbol{X}_t + \boldsymbol{W}_t, \qquad Y_t = \begin{bmatrix} \theta & 1 \end{bmatrix} \begin{bmatrix} X_{t-1} \\ X_t \end{bmatrix}$$

namely $\boldsymbol{G}_t = [\theta\ 1], \boldsymbol{W}_t = 0,\ \forall\, t$ and the state equation is given by

$$\boldsymbol{X}_{t+1} = \boldsymbol{F}_t \boldsymbol{X}_t + \boldsymbol{V}_t, \qquad \boldsymbol{V}_t = [0, 1]^\top Z_{t+1}$$

We can verify this

$$\begin{aligned} \boldsymbol{G}_t \boldsymbol{X}_t &= \theta X_{t-1} + X_t \\ &= \theta X_{t-1} + \phi X_{t-1} + Z_t \end{aligned}$$

$$\begin{aligned} Y_t &= \theta X_{t-1} + X_t \\ &= \theta(\phi X_{t-2} + Z_{t-1}) + (\phi X_{t-1} + Z_t) \\ &= \phi(X_{t-1} + \theta X_{t-2}) + Z_{t-1} + Z_t \\ &= \phi Y_{t-1} + \theta Z_{t-1} + Z_t \end{aligned}$$

Example 4.4

Let $\{Y_t\} \sim \text{MA}(1)$ and $Y_t = Z_t + \theta Z_{t-1}, \{Z_t\} \sim \text{WN}(0, \sigma^2)$ and let

$$\boldsymbol{X}_t = \begin{bmatrix} Y_t \\ \theta Z_t \end{bmatrix}$$

with the observation equation

$$Y_t = \boldsymbol{G}_t \boldsymbol{X}_t + \boldsymbol{W}_t, \qquad \boldsymbol{G}_t = [1 \;\; 0], \qquad \boldsymbol{W}_t \equiv 0 \; \forall \, t$$

and the state equation

$$\boldsymbol{X}_{t+1} = \boldsymbol{F}_t \boldsymbol{X}_t + \boldsymbol{V}_t, \qquad \boldsymbol{F}_t = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad \boldsymbol{V}_t = \begin{bmatrix} 1 \\ \theta \end{bmatrix} Z_{t+1}$$

The representation is not unique; alternatively, we could have set $X_t = Z_{t-1}$ and then in the observation equation set $\boldsymbol{G}_t = [\theta], \boldsymbol{W}_t = Z_t$ and for the state equations $F_t = [0], V_t = Z_t$. [50]

Low dimension or high dimensional representation can be used, in the next case the higher dimension is more transparent.

Example 4.5 (ARMA(p, q) process)

Suppose $\{Y_t\} \sim \text{ARMA}(p, q)$ with respect to $\{Z_t\} \sim \text{WN}(0, \sigma^2)$ is causal, $\Phi(B) Y_t = \Theta(B) Z_t$. Let $m = \max\{p, q+1\}$, $\phi_j = 0 \; \forall \; j > p, \theta_j = 0 \; \forall \; j > q$ and $\theta_0 = 1$

Let $\{X_t\} \sim \text{AR}(p)$ with $\Phi(B) X_t = Z_t$ and $\boldsymbol{X}_t = [X_{t-M+1}, X_{t-m+2}, \ldots, X_{t-1}, X_t]^\top$, an $(m \times 1)$ vector. Then the observation equation

$$\boldsymbol{Y}_t = \boldsymbol{G}_t \boldsymbol{X}_t + \boldsymbol{W}_t, \qquad \boldsymbol{G}_t = [\theta_{n-1}, \theta_{n-2}, \ldots, \theta_1, \theta_0], \qquad \boldsymbol{W}_t \equiv 0$$

and the state equation

$$\boldsymbol{X}_{t+1} = \boldsymbol{F}_t \boldsymbol{X}_t + \boldsymbol{V}_t, \qquad \boldsymbol{F}_t = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ \phi_m & \phi_{m-1} & \phi_{m-2} & \cdots & \phi_1 \end{bmatrix}_{(m \times m),} \qquad \boldsymbol{V}_t = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ Z_{t+1} \end{bmatrix}_{(m \times 1)}$$

Alternatively, let $m = \max\{p, q\}, \phi_j = 0 \; \forall \; j > p$. We use the linear process representation

[50] Independence is not in the formulation. We can also reduce the dimension of the state. It is rather simple in the MA(1) case.

to get another state-space form. Now $\Phi(B)Y_t = \Theta(B)Z_t$, which in terms imply that

$$\begin{aligned} Y_t &= \{\Phi(B)\}^{-1}\Theta(B)Z_t \\ &= \Psi(B)Z_t \\ &= \sum_{j=0}^{\infty} \Psi_j Z_{t-j} \end{aligned}$$

where $\{\psi_j\}$ are the coefficients in the expansion of $\Psi(Z) = \frac{\Theta(Z)}{\Phi(Z)}$ as a power series in Z. Let $\{X_t\}$ be defined by the AR(1)equation,

$$\boldsymbol{X}_{t+1} = \boldsymbol{F}_t \boldsymbol{X}_t + \boldsymbol{V}_t,$$

an $(m \times 1)$ system (VAR(1)) where $\boldsymbol{F}_t$ is as before, and $V_t = [\psi_1, \psi_2, \ldots, \psi_m]^\top Z_{t+1} = \boldsymbol{H} Z_{t+1}$ say. The observation equation is

$$\boldsymbol{Y}_t = \boldsymbol{G}_y \boldsymbol{X}_t + \boldsymbol{W}_t$$

with $\boldsymbol{G}_t = [1, 0, \ldots, 0], \boldsymbol{W}_t = Z_t$

and to justify this, we will differ this until we discuss multivariate processes. [51]

$\boldsymbol{F}_t, \boldsymbol{G}_t$ can be allowed to vary over time periods in a deterministic fashion, so that the process is not stationary, but there is some stability.

Stationarity and Stability

Recall the state equation

$$\boldsymbol{X}_{t+1} = \boldsymbol{F}_t \boldsymbol{X}_t + \boldsymbol{V}_t, \quad t = 0, \pm 1, \pm 2$$

[51] The lowest dimension one can get is $\max\{p, q\}$. For arbitrary state space representation, there is no unique formulation, can add a term in the observation equation and remove it in the state equation.

and suppose $\boldsymbol{F}_t = \boldsymbol{F}\ \forall\ t$. Then

$$\begin{aligned}\boldsymbol{X}_{t+1} &= \boldsymbol{F}\boldsymbol{X}_t + \boldsymbol{V}_t \\ &= \boldsymbol{F}(\boldsymbol{F}\boldsymbol{X}_{t-1} + \boldsymbol{V}_{t-1}) + \boldsymbol{V}_t \\ &= \boldsymbol{F}^2\boldsymbol{X}_{t-1} + \boldsymbol{F}\boldsymbol{V}_{t-1} + \boldsymbol{V}_t \\ &\vdots \\ &= \boldsymbol{F}^{n+1}\boldsymbol{X}_{t-n+1} + \sum_{j=0}^{n} \boldsymbol{F}^j \boldsymbol{V}_{t-j}\end{aligned}$$

that is for $\boldsymbol{X}_{t+1}$ to be bounded in probability (for all t), we must have that $\boldsymbol{F}^n$ stays bounded for all n. Write

$$\boldsymbol{F} = \boldsymbol{E}\boldsymbol{D}\boldsymbol{E}^{-1}$$

as the eigendecomposition for $\boldsymbol{F}$, then $\boldsymbol{D}$ is the matrix of eigenvalues of $\boldsymbol{F}$, and $\boldsymbol{F}^n = \boldsymbol{E}\boldsymbol{D}^n\boldsymbol{E}^{-1}$, *i.e.* for $\boldsymbol{F}^n$ to stay bounded (elementwise), we require that

$$\boldsymbol{D}^n = [\text{diag}(\lambda_1, \ldots, \lambda_L)]^n$$

stays bounded. But

$$\boldsymbol{D}^n = \text{diag}(\lambda_1^n, \ldots, \lambda_L^n)$$

and therefore, we need $|\lambda_j| \leq 1$ (*i.e.* $|\lambda_1| < 1$ under the usual convention.); thus we need solutions of

$$\det(\boldsymbol{F} - \boldsymbol{I}_Z) = 0$$

to be within the unit circle, or equivalently in Brockwell and Davis notation,

$$\det(\boldsymbol{I} - \boldsymbol{F}_Z) \neq 0\ \forall\ z \in \mathbb{C}, |z| \leq 1.$$

Section 4.2: Basic Structural Models

Local level model

The observation equation can be written simply as

$$Y_t = M_t + W_t, \qquad \{W_t\} \sim \text{WN}(0, \sigma_W^2)$$

and the state equation

$$M_{t+1} = M_t + V_t, \qquad \{V_t\} \sim \text{WN}(0, \sigma_V^2)$$

where V_t, W_t are uncorrelated processes. Thus the mean is evolving as a random walk and Y_t varies according to some noise. Here $K = L = 1$, so a 1 dimensional state model. The state equation here is clearly a random walk. Let

$$Y_t^* = \nabla Y_t = Y_t - Y_{t-1} = V_{t-1} + (W_t - W_{t-1})$$

and clearly, $\mathsf{E}(Y_t) = 0$ and

$$\mathsf{E}\left(Y_t^* Y_{t+h}^*\right) = \mathsf{E}\left((V_{t-1} + (W_t - W_{t-1})(V_{t+h-1} + (W_{t+h} - W_{t+h-1}))\right)$$

and setting h, we get

$$\mathsf{E}\left(Y_t^* Y_{t+h}^*\right) = \begin{cases} \mathsf{E}\left(Y_t^{*2}\right) = \sigma_V^2 + 2\sigma_W^2 & \text{if } h = 0 \\ \mathsf{E}\left(Y_t^* Y_{t+1}^*\right) = -\sigma_W^2 & \text{if } h = 1 \\ \mathsf{E}\left(Y_t^* Y_{t+h}^2\right) = 0 & \text{if } h \geq 2 \end{cases}$$

which corresponds to an ARIMA(0,1,1) model.

Local slope model

The observation equation is $Y_t = M_t + W_t$, the state equation $M_t = M_{t-1} + B_{t-1} + V_{t-1}^{(0)}$ where $B_t = B_{t-1} + U_{t-1}$ where $\{W_t\} \sim \text{WN}(0, \sigma_W^2)$, where $\{V_t^{(0)}\} \sim \text{WN}(0, \sigma_V^2)$ and $\{U_t\} \sim \text{WN}(0, \sigma_U^2)$. let

$$X_t = \begin{bmatrix} M_t \\ B_t \end{bmatrix}, \quad V_t = \begin{bmatrix} V_t^{(0)} \\ U_t \end{bmatrix}$$

Then

$$X_{t+1} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} X_t + V_t;$$

with $Y_t = [1 \quad 0] X_t + W_t$. Here $V_t \sim \text{WN}(0, \boldsymbol{Q})$ where $\boldsymbol{Q} = \begin{pmatrix} \sigma_V^2 & 0 \\ 0 & \sigma_U^2 \end{pmatrix}$ and the dimensions parameter of the state space model are $L = 2, K = 1$.

Seasonal model

Suppose we wish to construct a stochastic seasonal component with period d say. Recall that $\{S_t\}$ is satisfied $S_{t+d} = S_t$ all $t > d$ and $\sum_{t=1}^{d} S_t = 0$.

Write

$$Y_{t+1} = -Y_t - Y_{t-1} - \cdots - Y_{t-d+2} + S_t$$

where $\{S_t\}$ is a zero mean random variable. We can write this in a state-space formulation as follows: Let

$$\boldsymbol{X}_t = \begin{bmatrix} Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-d+2} \end{bmatrix}_{((d-1)\times 1)}$$

so that

$$Y_t = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix} \boldsymbol{X}_t + W_t.$$

Now the state space equation is

$$\boldsymbol{X}_{t+1} = \boldsymbol{F}\boldsymbol{X}_t + \boldsymbol{V}_t$$

with

$$\boldsymbol{F} = \begin{bmatrix} -1 & -1 & -1 & \cdots & -1 \\ 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 \cdots & \vdots & \\ 0 & 0 & 1 \ddots & \vdots & \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \quad , \boldsymbol{V}_t = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} S_t$$

and where $\{S_t\} \sim \text{WN}(0, \sigma_S^2)$

Combining the three components altogether: Let $\boldsymbol{X}_t = [M_t, B_t, Y_t, Y_{t-1}, \ldots, Y_{t-d+2}]^\top$. Then

$$\texttt{Obs} \quad Y_t = \boldsymbol{G}_t\boldsymbol{X}_t + \boldsymbol{W}_t, \qquad \boldsymbol{G} = [1, 0, 1, 0, \ldots 0], \qquad \boldsymbol{W}_t \sim \text{WN}(0, \sigma_W^2)$$

and

$$\texttt{State} \quad \boldsymbol{X}_{t+1} = \boldsymbol{F}_t\boldsymbol{X}_t + \boldsymbol{V}_t$$

where $\boldsymbol{F}$ comprises the block diagonal from the previous models.

$$\boldsymbol{F}_t \equiv \boldsymbol{F} = \begin{bmatrix} 1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & -1 & -1 & \cdots & -1 \\ \vdots & \vdots & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & 0 & \ddots & \cdots & \ddots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix}$$

[Reconstructed states from the plots in air data.] σ_1^2 is the variance from $\{M_t\}$, σ_2^2 the variance from $\{B_t\}$ and similarly σ_3^2 for $\{S_t\}$ and

$$\boldsymbol{V}_t \sim \text{WN}(0, \boldsymbol{Q}) \text{ where } \boldsymbol{Q} = \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$$

In this code, we have B_{t+1} is constant, so this has been adjusted in the code accordingly.

$$\begin{aligned} Y_t &= M_t + (-Y_{t-1} - \cdots - Y_{t-d+1}) + W_t \\ M_{t+1} &= M_t + B_t + V_t^{(0)} \\ B_{t+1} &= V_t + U_t \\ Y_{t+1} &= (-Y_t - \cdots - Y_{t-d+2}) + S_t \end{aligned}$$

This is a 17 parameters maximum likelihood, we get a plot with a slope around 2. Amending the code to maximize over the logged data (log-scale transformation), we have a much more linear slope, with less leak from the seasonal part. In this form (not-logged), it is capturing the heteroskedasticity of the model. The seasonal component and the forecast are better.

Section 4.3: Filtering and Smoothing: the Kalman Filter

See handouts from now on– The key ideas are the following: imagine for simplicity a Normal dataset; in the Gaussian case,

$$\begin{aligned} \texttt{Obs} \quad & Y_t|\boldsymbol{X}_t \sim \mathcal{N}_K(\boldsymbol{G}_t\boldsymbol{X}_t, \boldsymbol{\Sigma}_t) \\ \texttt{State} \quad & \boldsymbol{X}_{t+1}|\boldsymbol{X}_t \sim \mathcal{N}_L(\boldsymbol{F}_t\boldsymbol{X}_t, \boldsymbol{\Omega}_t) \end{aligned}$$

Then the **prediction** is

$$p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t}) = \int p(\boldsymbol{x}_{t+1}|\boldsymbol{x}_t)p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})d\boldsymbol{x}_t$$

and $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ is not yet available. This comes back to the **filtering**, where

$$p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t}) = \frac{p(y_t|\boldsymbol{x}_t)p(\boldsymbol{x}_t|\boldsymbol{y}_{1:(t-1)})}{p(y_t|\boldsymbol{y}_{1:(t-1)})}$$

and the latter term is again of the form found in the prediction. This suggests that we compute (recursively) $p(\boldsymbol{x}_1), p(\boldsymbol{x}_1|y_1), p(\boldsymbol{x}_2|y_1), p(\boldsymbol{x}_2|y_2), \ldots$. For time point t, in the Gaussian case, if $p(\boldsymbol{x}_{1:n}, \boldsymbol{y}_{1:n})$ is jointly Gaussian, so therefore $p(\boldsymbol{x}_t|\boldsymbol{y}_{1:t})$ is also Gaussian, as is $p(\boldsymbol{x}_{t+1}|\boldsymbol{y}_{1:t})$ (any conditional distribution or marginal is also normal). Hence all we need to do is track **moments** of the prediction and filtering distribution. For example,

$$\begin{aligned}
\boldsymbol{X}_t|\boldsymbol{X}_1, \boldsymbol{Y}_{1:t} &\sim \mathcal{N}_L(\boldsymbol{a}_{t|t}, \boldsymbol{P}_{t|t}) \\
\boldsymbol{X}_{t+1}|\boldsymbol{X}_1, \boldsymbol{Y}_{1:t} &\sim \mathcal{N}_K(\boldsymbol{a}_{t+1|t}, \boldsymbol{P}_{t+1}|t).
\end{aligned}$$

By the law of iterated expectation,

$$\begin{aligned}
\mathsf{E}\left(\boldsymbol{X}_{t+1}|\boldsymbol{x}_1, \boldsymbol{y}_{1:t}\right) &= \boldsymbol{F}_t\boldsymbol{a}_{t|t} = \boldsymbol{a}_{t+1|t} \\
\mathsf{Var}\left(\boldsymbol{X}_{t+1}|\boldsymbol{x}_1, \boldsymbol{y}_{1:t}\right) &= \boldsymbol{F}_t\boldsymbol{P}_{t|t}\boldsymbol{F}_t^\top + \boldsymbol{\Omega}_t = \boldsymbol{P}_{t+1|t} \\
\mathsf{E}\left(Y_{t+1}|\boldsymbol{x}_1, \boldsymbol{y}_{1:t}\right) &= \boldsymbol{G}_{t+1}\boldsymbol{a}_{t+1|t} \\
\mathsf{Var}\left(Y_{t+1}|\boldsymbol{x}_1, \boldsymbol{y}_{1:t}\right) &= \boldsymbol{G}_{t+1}\boldsymbol{P}_{t+1|t}\boldsymbol{G}_{t+1}^\top + \boldsymbol{\Sigma}_{t+1}
\end{aligned}$$

Thus for likelihood-based estimation,

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{y}_{1:n}) = f(y_1|\boldsymbol{\theta}) \prod_{t=2}^{n} f(y_t|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{\theta})$$

where

$$f(y_t|\boldsymbol{y}_{1:(t-1)}, \boldsymbol{\theta}) \sim \mathcal{N}_k(\boldsymbol{G}_{t+1}\boldsymbol{a}_{t+1|t}, \boldsymbol{G}_{t+1}\boldsymbol{P}_{t+1|t}\boldsymbol{G}_{t+1}^\top + \boldsymbol{\Sigma}_t)$$

and we can compute the likelihood using terms computed during the Kalman filter. This however heavily depends on the Gaussianity assumption. Conjugate prior distribution and discrete distribution are examples where we can compute numerically using the Kalman-Filter. The recursive calculations and the algorithm are outlined in the R code provided.

The Kalman recursions are given by the following equations, where $\boldsymbol{a}_{t|t}$ are the best linear predictors and $\boldsymbol{P}_{t|t}$ the corresponding mean-square error matrix. The quantities $\boldsymbol{v_t}$ and $\boldsymbol{M}_t$ denote the one-step-ahead error in forecasting y_t conditional on the information set at time

$t-1$ and its MSE, respectively.

$$\boldsymbol{v}_t = y_t - \boldsymbol{G}_t \boldsymbol{a}_{t|t-1} \tag{4.11a}$$
$$\boldsymbol{M}_t = \boldsymbol{G}_t \boldsymbol{P}_{t|t-1} \boldsymbol{G}_t^\top + \boldsymbol{\Sigma}_t \tag{4.11b}$$
$$\boldsymbol{a}_{t|t} = \boldsymbol{a}_{t|t-1} + \boldsymbol{P}_{t|t-1} \boldsymbol{G}_t^\top \boldsymbol{M}_t^{-1} \boldsymbol{v}_t \tag{4.11c}$$
$$\boldsymbol{P}_{t|t} = \boldsymbol{P}_{t|t-1} - \boldsymbol{P}_{t|t-1} \boldsymbol{G}_t^\top \boldsymbol{M}_t^{-1} \boldsymbol{G}_t \boldsymbol{P}_{t|t-1} \tag{4.11d}$$
$$\boldsymbol{a}_{t+1|t} = \boldsymbol{F}_t \boldsymbol{a}_{t|t} \tag{4.11e}$$
$$\boldsymbol{P}_{t+1|t} = \boldsymbol{F}_t \boldsymbol{P}_{t|t} \boldsymbol{F}_t^\top + \boldsymbol{\Omega}_t \tag{4.11f}$$

which can be combined into further recursions, used to get another form for the prediction and simplify the former forms

$$\boldsymbol{a}_{t+1|t} = \boldsymbol{F}_t \boldsymbol{a}_{t|t-1} + \boldsymbol{K}_t \boldsymbol{v}_t \tag{4.12a}$$
$$\boldsymbol{K}_t = \boldsymbol{F}_t \boldsymbol{P}_{t|t-1} \boldsymbol{G}_t^\top \boldsymbol{M}_t^{-1} \tag{4.12b}$$
$$\boldsymbol{P}_{t+1|t} = \boldsymbol{F}_t \boldsymbol{P}_{t|t-1} \boldsymbol{L}_t^\top + \boldsymbol{\Omega}_t \tag{4.12c}$$
$$\boldsymbol{L}_t = \boldsymbol{F}_t + \boldsymbol{K}_t \boldsymbol{G}_t \tag{4.12d}$$

and a smoothing algorithm can be applied to a state-space model given a fixed set of data; estimates of the state vector are computed at each t using all available information. Denote by $\boldsymbol{a}_{t|n}$ the smoothed linear estimates for $t \in \{0, \ldots, n-1\}$ given all data until point n, that is $\boldsymbol{a}_{t|n} = \mathsf{E}(\boldsymbol{X}_t | \boldsymbol{y}_{1:n})$ along with $\boldsymbol{P}_{t|n}$ via backward recursions.

$$\boldsymbol{P}_t^* = \boldsymbol{P}_t \boldsymbol{F}_t^\top \boldsymbol{P}_{t+1|t} \tag{4.13a}$$
$$\boldsymbol{a}_{t|n} = \boldsymbol{a}_{t|t} + \boldsymbol{P}_t^* (\boldsymbol{a}_{t+1|n} - \boldsymbol{a}_{t+1|t}) \tag{4.13b}$$
$$\boldsymbol{P}_{t|n} = \boldsymbol{P}_{t|t} + \boldsymbol{P}_t^* (\boldsymbol{P}_{t+1|n} - \boldsymbol{P}_{t+1|t}) \boldsymbol{P}_t^* \tag{4.13c}$$

The Gaussian linear state space model is of the form

$$Y_t = X_t + W_t$$
$$X_{t+1} = X_t + V_t$$

where we assume that $W_t \sim \mathcal{N}(0, \sigma_W^2)$ and $V_t \sim \mathcal{N}(0, \sigma_V^2)$ are independent white noise series.

The joint likelihood for observations and states up to n is of the form

$$p(\boldsymbol{x}_{1:n}, y_{1:n}) = p(x_1) p(y_1 | x_1) \prod_{i=2}^{n} p(x_i | x_{i-1}) p(y_i | x_i)$$

which is a two parameter model in the Gaussian case; alternatively, we could also set x_1 to be unknown; it could be regarded as having some density, or even a state which corresponds to a degenerate distribution. The joint likelihood explicited above is follows a multinormal distribution $\mathcal{N}_{2n}(\cdot,\cdot)$, and from standard results all marginal distributions – but also conditional distributions, are also Gaussian. We could thus compute the first two moments to get a complete characterization of the model.

(1) Filtering.

As earlier mentioned, this calculations is a Bayes theorem calculation, where we abstract for now of the terms involving y_t in the denominator. We could regard the proportionality, having

$$\begin{aligned} p(x_t|\boldsymbol{y}_{1:t}) &\propto p(y_t|x_t)p(x_t|\boldsymbol{y}_{1:t-1}) \\ &\propto \exp\left(-\frac{1}{2\sigma_W^2}(y_t - x_t)^2\right) \times \exp\left(-\frac{1}{2S_{t|t-1}^2}(x_t - m_{t|t-1})^2\right) \end{aligned}$$

as from the observation and state equations, we have

$$Y_t|X_t = x_t \sim \mathcal{N}(x_t, \sigma_W^2) \qquad X_{t+1}|X_t = x_t \sim \mathcal{N}(x_t, \sigma_V^2)$$

and the recursion assumption is that $X_t|Y_{1:t-1} \sim \mathcal{N}(m_{t|t-1}, S_{t|t-1}^2)$, which depends only on low dimensional summary statistics.

$$p(x_t|\boldsymbol{y}_{1:t}) \propto \exp\left(-\frac{1}{2}\left[\frac{1}{\sigma_W^2}(x_t - y_t)^2 + \frac{1}{S_{t|t-1}^2}(x_t - m_{t|t-1})^2\right]\right)$$

as this is an univariate distribution in x_t; we want to get this in terms of

$$\exp\left(-\frac{1}{2S_{t|t}^2}(x_t - m_{t|t})^2\right)$$

such that $X_t|Y_{1:t} \sim \mathcal{N}(m_{t|t}, S_{t|t}^2)$. But the following is a simple "complete the square" calculation, using the fact that

$$A(x-a)^2 + B(x-b)^2 = M(x-m)^2 + \text{constant} = (A+B)\left(x - \frac{Aa+Bb}{A+B}\right)^2 + \frac{AB}{A+B}(a-b)^2$$

where $m = (Aa+Bb)/(A+B)$ and $M = A+B$. We thus get

$$\exp\left(-\frac{1}{2}\left[\left(\frac{1}{\sigma_W^2} + \frac{1}{S_{t|t-1}^2}\right)\left(x_t - \frac{y_t/\sigma_W^2 + m_{t|t-1}/S_{t|t-1}^2}{1/\sigma_W^2 + 1/S_{t|t-1}^2}\right)^2 + \text{constant}\right]\right)$$

which imply that

$$m_{t|t} = \frac{\frac{y_t}{\sigma_W^2} + \frac{m_{t|t-1}}{S_{t|t-1}^2}}{\frac{1}{\sigma_W^2} + \frac{1}{S_{t|t-1}^2}} = \frac{y_t S_{t|t-1}^2 + \sigma_W^2 m_{t|t-1}}{S_{t|t-1}^2 + \sigma_W^2} \quad \text{and} \quad S_{t|t}^2 = \left(\frac{1}{\sigma_W^2} + \frac{1}{S_{t|t-1}^2}\right)^{-1}$$

and so the parameters of the Normal distributions are known in terms of the previous observations.

(2) Predictions This is a posterior predictive-type calculation. We have

$$\begin{aligned}
p(x_{t+1}|\boldsymbol{y}_{1:t}) &= \int p(x_{t+1}|x_t, \boldsymbol{y}_{1:t}) p(x_t|\boldsymbol{y}_{1:t}) dx_t \\
&\equiv \int p(x_{t+1}|x_t) p(x_t|\boldsymbol{y}_{1:t}) dx_t \\
&\propto \int \exp\left(-\frac{1}{2}\frac{1}{\sigma_V^2}\left(x_{t+1} - x_t\right)^2\right) \exp\left(-\frac{1}{2S_{t|t}^2}(x_t - m_{t|t})^2\right) dx_t \\
&= \exp\left(-\frac{1}{2\left(S_{t|t}^2 + \sigma_V^2\right)}\left(x_{t+1} - m_{t|t}\right)^2\right) \int \exp\left(-\frac{1}{2}\frac{S_{t|t}^2 + \sigma_V^2}{\sigma_V^2 S_{t|t}^2}(x - m^*)^2\right) \\
&\propto \exp\left(-\frac{1}{2}\frac{1}{\sigma_V^2 + S_{t|t}^2}(x_{t+1} - m_{t|t})^2\right)
\end{aligned}$$

using the fact that $x_{t+1}|x_t \perp \boldsymbol{y}_{1:t}$ by assumption. We can therefore conclude that $p(x_{t+1}|\boldsymbol{y}_{1:t}) \sim \mathcal{N}(m_{t+1|t}, S_{t+1|t}^2)$. For the prediction, we could also use a shortcut using conditional independence and do an iterated expectation and iterated variance calculation. We will again derive (in this case verify) that $m_{t+1|t} = m_{t|t}$ and $S_{t+1|t}^2 = \sigma_V^2 +^2 S_{t|t}^2$.

$$m_{t+1|t} = \mathsf{E}\left(X_{t+1}|\boldsymbol{y}_{1:t}\right) = \mathsf{E}_{X_t|\boldsymbol{Y}_{1:t}}\left(\mathsf{E}_{X_{t+1}|X_t,\boldsymbol{Y}_{1:t}}(X_t|X_t, \boldsymbol{y}_{1:t})\right) = \mathsf{E}_{X_t|\boldsymbol{Y}_{1:t}}(X_t) = m_{t|t}$$

and similarly for the variance calculation,

$$\begin{aligned}
&\mathsf{Var}\left(X_{t+1}|\boldsymbol{y}_{1:t}\right) \\
&= \mathsf{Var}_{X_t|\boldsymbol{Y}_{1:t}}\left(\mathsf{E}_{X_{t+1}|X_t,\boldsymbol{Y}_{1:t}}(X_{t+1}|X_t, \boldsymbol{y}_{1:t})\right) + \mathsf{E}_{X_t|\boldsymbol{Y}_{1:t}}\left(\mathsf{Var}_{X_{t+1}|X_t,\boldsymbol{Y}_{1:t}}(X_{t+1}|X_t, \boldsymbol{y}_{1:t})\right) \\
&= S_{t|t}^2 + \sigma_V^2.
\end{aligned}$$

It remains to calculate $p(\boldsymbol{y}_{1:n}) = p(y_1)p(y_2|y_1 p(y_3|y_1, y_2) \cdots p(y_n|\boldsymbol{y}_{1:n-1})$. This is again a

posterior-type calculation, which can be made using

$$p(y_t|\boldsymbol{y}_{1:t-1}) = \int p(y_t|x_t)p(x_t|\boldsymbol{y}_{1:t-1})dx_t$$

both Gaussian in X_t or

$$p(x_t|\boldsymbol{y}_{1:t}) = \frac{p(y_t|x_t)p(x_t|\boldsymbol{y}_{1:t-1})}{p(y_t|\boldsymbol{y}_{1:t-1})}$$

but the simplest calculation is via iterated moment calculation. Indeed,

$$\begin{aligned}\mathsf{E}\left(Y_t|\boldsymbol{Y}_{1:t-1}\right) &= \mathsf{E}_{X_t|Y_{1:t-1}}\left(\mathsf{E}_{Y_t|X_t}(Y_t|X_t)\right) \\ &= \mathsf{E}_{X_t|Y_{1:t-1}}(X_t) = m_{t|t-1}\end{aligned}$$

and for the variance,

$$\begin{aligned}\mathsf{Var}\left(Y_t|\boldsymbol{Y}_{1:t-1}\right) &= \mathsf{Var}_{X_t|Y_{1:t-1}}\left(\mathsf{E}_{Y_t|X_t}(Y_t|X_t)\right) + \mathsf{E}_{X_t|Y_{1:t-1}}\left(\mathsf{Var}_{Y_t|X_t}(Y_t|X_t)\right) \\ &= \mathsf{Var}_{X_t|Y_{1:t-1}}\left(X_t|Y_{1:t-1}\right) + \mathsf{E}_{X_t|Y_{1:t-1}}(\sigma_W^2) \\ &= S_{t|t-1}^2 + \sigma_W^2\end{aligned}$$

Chapter 5
Financial time series models

To model

- asset/bond/option prices
- interest rates
- exchange rates

simple stationary/non-stationary models are not sufficiently sophisticated (complex) to capture observed dynamics. Such series often involve non-stationary components (time-varying mean, unit roots) and also temporal heteroskedasticity, (*i.e.* time-varying variance).

License