

Time Series Analysis

3. Model free methods

Andrew Lesniewski

Baruch College
New York

Fall 2023

Outline

- 1 Time series in frequency domain
- 2 Singular spectrum analysis
- 3 Hurst exponent
- 4 Granger causality
- 5 Entropy methods

Time series in frequency domain

- So far, we have discussed various models within the *parametric approach* to time series analysis.
- The key element of this approach is to specify a time series model with a small (or moderate) number of free parameters which are determined via estimation from a data set.
- While this approach will remain the focus of these lectures, we will now take a brief side trip into the non-parametric (or model free) approach to time series analysis.
- In particular, we will discuss methods of analyzing time series by means of expansion in various basis functions.
- The recurrent neural networks discussed at the end of this course fall into this category.

Time series in frequency domain

- The first approach that we discuss, namely *time series analysis in frequency domain* (in contrast to the *time domain* approach taken so far), is reminiscent of Fourier transform approach in signal processing.
- The idea is to decompose the underlying time series into components, each of which corresponds to evolution *cycles* of different frequencies.
- The appropriate basis functions are the trigonometric functions $\cos(\omega t)$ and $\sin(\omega t)$ or, equivalently, the complex exponential function $e^{i\omega t}$.

Spectral density function

- Let X_t be a covariance stationary time series, such that

$$\sum_{t=-\infty}^{\infty} |\Gamma_t| < \infty. \quad (1)$$

- The *spectral density function* (SDF), or *population spectrum*, of X_t is defined as

$$s_X(\omega) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} \Gamma_t e^{-i\omega t}. \quad (2)$$

It is essentially the Fourier transform of Γ_t .

- From the trigonometric representation of complex numbers, and the fact that $\Gamma_{-t} = \Gamma_t$, we can write this in terms of purely real valued quantities:

$$s_X(\omega) = \frac{1}{2\pi} \left(\Gamma_0 + 2 \sum_{t=1}^{\infty} \Gamma_t \cos(\omega t) \right). \quad (3)$$

Spectral density function for white noise

- The easiest example is that of a white noise, $X_t = \varepsilon_t$. In this case,

$$\Gamma_t = \begin{cases} \sigma^2, & \text{if } t = 0, \\ 0, & \text{otherwise.} \end{cases}$$

- As a consequence, the SDF is constant,

$$s_X(\omega) = \frac{\sigma^2}{2\pi}. \quad (4)$$

Spectral density function for $AR(1)$

- As a next example, let us determine the spectral density function of the $AR(1)$ process.
- From equation (14) in Lecture Notes #1,

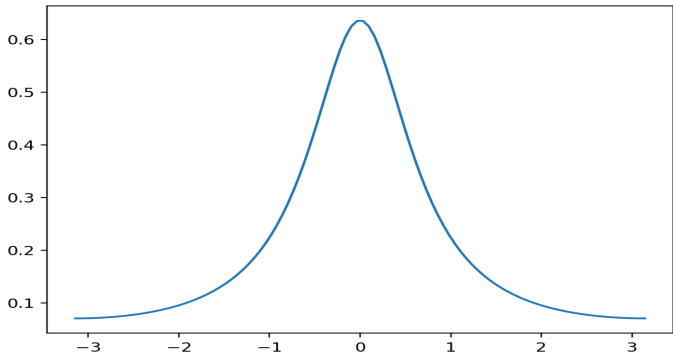
$$\begin{aligned}s_X(\omega) &= \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} \Gamma_0 \beta^{|t|} e^{-i\omega t} \\&= \frac{\Gamma_0}{2\pi} \left(1 + \sum_{t=1}^{\infty} \beta^t e^{i\omega t} + \sum_{t=1}^{\infty} \beta^t e^{-i\omega t} \right) \\&= \frac{\Gamma_0}{2\pi} \left(1 + \frac{\beta e^{i\omega}}{1 - \beta e^{i\omega}} + \frac{\beta e^{-i\omega}}{1 - \beta e^{-i\omega}} \right).\end{aligned}$$

- As a result,

$$s_X(\omega) = \frac{\sigma^2}{2\pi} \frac{1}{1 - 2\beta \cos \omega + \beta^2}. \quad (5)$$

Spectral density function for $AR(1)$

- Below is the plot of (5) with $\beta = 0.5$ and $\sigma = 1$.



Spectral density function for $MA(1)$

- Let us now consider an $MA(1)$ model.
- Using equation (62) in Lecture Notes #1, we see that

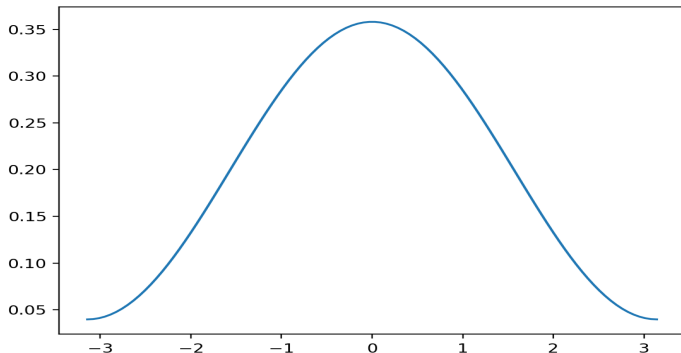
$$s_X(\omega) = \frac{1}{2\pi} ((1 + \theta^2)\sigma^2 + \theta\sigma^2 e^{i\omega} + \theta\sigma^2 e^{-i\omega}).$$

- This implies that the SDF of an $MA(1)$ process is

$$s_X(\omega) = \frac{\sigma^2}{2\pi} (1 + 2\theta \cos \omega + \theta^2). \quad (6)$$

Spectral density function for $MA(1)$

- Below is the plot of (6) with $\theta = 0.5$ and $\sigma = 1$.



Spectral density for $ARMA(p, q)$

- The calculations above can be generalized to produce an expression for the $ARMA(p, q)$ model:

$$\psi(L)X_t = \alpha + \varphi(L)\varepsilon_t, \quad (7)$$

where our notation follows Lecture Notes #1.

- Namely, as you will show in Homework Assignment #3, the SDF is then given by

$$s_X(\omega) = \frac{\sigma^2}{2\pi} \left| \frac{\varphi(e^{i\omega})}{\psi(e^{i\omega})} \right|^2. \quad (8)$$

Spectral density for $ARMA(p, q)$

- We factorize the polynomials $\psi(z)$ and $\varphi(z)$,

$$\begin{aligned}\psi(z) &= (1 - \lambda_1 z) \dots (1 - \lambda_p z), \\ \varphi(z) &= (1 - \mu_1 z) \dots (1 - \mu_q z),\end{aligned}$$

where $1/\lambda_1, \dots, 1/\lambda_p$, and $1/\mu_1, \dots, 1/\mu_q$ are the roots of $\psi(z)$ and $\phi(z)$, respectively.

- Then, explicitly,

$$s_X(\omega) = \frac{\sigma^2}{2\pi} \frac{(1 - 2\mu_1 \cos \omega + \mu_1^2) \dots (1 - 2\mu_q \cos \omega + \mu_q^2)}{(1 - 2\lambda_1 \cos \omega + \lambda_1^2) \dots (1 - 2\lambda_p \cos \omega + \lambda_p^2)}. \quad (9)$$

Spectral density function

- In general, the spectral density function $s_X(\omega)$ has the following properties:
 - (i) It is non-negative.
 - (ii) It is a periodic function of ω with period 2π (assuming $h = 1$).
 - (iii) It is continuous in ω .
- The autocovariance can be calculated from the population spectrum by means of

$$\Gamma_t = \int_{-\pi}^{\pi} s_X(\omega) e^{i\omega t} d\omega. \quad (10)$$

- This is an immediate consequence of the fact that

$$\int_{-\pi}^{\pi} e^{i\omega(t-s)} d\omega = \begin{cases} 2\pi, & \text{if } t = s' \\ 0, & \text{otherwise.} \end{cases} \quad (11)$$

Spectral density function

- Alternatively, we can do it without complex numbers:

$$\Gamma_t = \int_{-\pi}^{\pi} s_X(\omega) \cos(\omega t) d\omega. \quad (12)$$

- In particular,

$$\Gamma_0 = \int_{-\pi}^{\pi} s_X(\omega) d\omega, \quad (13)$$

i.e. the variance of X_t is equal to the area under the population spectrum between $-\pi$ and π .

- This also leads to the interpretation of $s_X(\omega)$ as the fraction of the variance that is attributable to cycles of frequency ω .
- There is a general result that states that any covariance-stationary time series process can be expressed in terms of its spectral data.

Spectral representation theorem

- Namely, there exists a unique complex valued stochastic function $z_X(\omega)$, such that

$$X_t = \mu + \int_{-\pi}^{\pi} e^{i\omega t} z_X(\omega) d\omega, \quad (14)$$

where $\mu = E(X_t)$.

- Since X_t is real valued, the random function $z_X(\omega)$ must have the following symmetry property:

$$\overline{z_X(\omega)} = z_X(-\omega). \quad (15)$$

- Furthermore, $z_X(\omega)$ has the following properties:

(i) For all ω ,

$$E(z_X(\omega)) = 0. \quad (16)$$

(ii) For all ω, ω' ,

$$E(z_X(\omega) \overline{z_X(\omega')}) = s_X(\omega) \delta(\omega - \omega'), \quad (17)$$

where $\delta(\omega - \omega')$ denotes Dirac's delta function.

Spectral representation theorem

- This result is known as the *spectral representation theorem* or *Cramer's theorem*.
- The spectral representation theorem can also be written in terms of real quantities only.
- Namely, we define

$$\begin{aligned}a_X(\omega) &= \operatorname{Re} z_X(\omega), \\ b_X(\omega) &= -\operatorname{Im} z_X(\omega)\end{aligned}\tag{18}$$

(the negative sign is just for convenience).

Spectral representation theorem

- Note that the random functions $a_X(\omega)$ and $b_X(\omega)$ have the following properties:
 - (i) As a consequence of (15):

$$\begin{aligned}a_X(-\omega) &= a_X(\omega), \\ b_X(-\omega) &= -b_X(\omega).\end{aligned}\tag{19}$$

- (ii) This implies the following identity:

$$a_X(\omega)^2 + b_X(\omega)^2 = |z_X(\omega)|^2.\tag{20}$$

- As a result, we can write

$$X_t = \mu + \int_{-\pi}^{\pi} (\cos(\omega t) a_X(\omega) + \sin(\omega t) b_X(\omega)) d\omega.\tag{21}$$

Sample periodogram

- A complete proof of the spectral representation theorem is a bit technical, and can be found in specialized mathematical literature.
- Instead, we will interpret it in terms sample data.
- Let x_1, \dots, x_T be observations of X_t , and let $\hat{\Gamma}_t$ denote the estimated autocovariance as defined by equation (5) in Lecture Notes #1.
- For any ω , the estimated sample spectral density function,

$$\hat{s}_X(\omega) = \frac{1}{2\pi} \sum_{t=-(T-1)}^{T-1} \hat{\Gamma}_t e^{-i\omega t}. \quad (22)$$

is called the *sample periodogram*.

Sample periodogram

- We can then verify that

$$\hat{\Gamma}_0 = \int_{-\pi}^{\pi} \hat{s}_X(\omega) d\omega, \quad (23)$$

i.e. the area under the periodogram is equal to the sample variance.

- In order to formulate the sample version of the spectral representation theorem, we assume that T is odd, and denote $\omega_j = 2\pi j/T$, for $j = -M, -M+1, \dots, M$, where $M = (T-1)/2$.

Sample periodogram

- For each j , we define

$$\hat{z}_X(\omega_j) = \frac{1}{T} \sum_{t=1}^T e^{-i\omega_j t} x_t - \hat{\mu}. \quad (24)$$

- Notice that

$$\hat{z}_X(\omega_0) = 0. \quad (25)$$

- Then

$$x_t = \hat{\mu} + \sum_{j=-M}^M e^{i\omega_j t} \hat{z}_X(\omega_j). \quad (26)$$

Sample periodogram

- To see this, we multiply both sides of (24) by $e^{i\omega_j s}$ and sum over $j = 1, \dots, M$, and notice that

$$\sum_{j=-M}^M e^{i\omega_j(s-t)} = \begin{cases} T, & \text{if } s = t, \\ 0, & \text{otherwise.} \end{cases}$$

- Finally, notice that

$$\sum_{j=1}^T (x_t - \hat{\mu})^2 = \sum_{j=-M}^M |\hat{z}_X(\omega_j)|^2. \quad (27)$$

Singular spectrum analysis

- *Singular spectrum analysis* (SSA) is a model free feature extraction methodology, which may be thought of as a variant of the principal component analysis (PCA).
- Its extension to multivariate time series (not discussed here) is referred to as *multi channel singular spectrum analysis* (M-SSA).
- We consider a sample from a time series X_1, \dots, X_T , and let $1 < l < T$ be the length of the rolling window. Then $k = T - l + 1$ is the number of lagged vectors.
- The basic algorithm of SSA consists of two stages:
 - (i) embedding,
 - (ii) reconstruction.

Singular spectrum analysis

- Embedding is carried out in two steps. First, we form the *trajectory matrix*:

$$\mathcal{X} = \begin{pmatrix} X_1 & X_2 & \dots & X_k \\ X_2 & X_3 & \dots & X_{k+1} \\ \vdots & \vdots & \dots & \vdots \\ X_l & X_{l+1} & \dots & X_T \end{pmatrix}. \quad (28)$$

- Note that $\mathcal{X}_{ij} = X_{i+j-1}$; matrices of this form are called *Hankel matrices*.
- The columns in the trajectory matrix correspond to the observations of the time series as the length l observation window slides forward.

Singular spectrum analysis

- Then, we perform the singular value decomposition (SVD) of the trajectory matrix \mathcal{X} :

- (i) Let $\mathcal{S} = \mathcal{X}\mathcal{X}^\top$. Then \mathcal{S} is positive definite; we denote its eigenvalues by $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_l \geq 0$, and the corresponding orthonormal system of eigenvectors by U_1, U_2, \dots, U_l . The numbers $\sqrt{\lambda_i}$ are called the *singular values* of \mathcal{X} .
- (ii) Let $r = \text{rank}(\mathcal{X})$ (typically, $r = l$), and set $V_i = \frac{1}{\sqrt{\lambda_i}} \mathcal{X}^\top U_i$, for $i = 1, \dots, l$.
- (iii) Then

$$\mathcal{X} = \mathcal{X}_1 + \mathcal{X}_2 + \dots + \mathcal{X}_r, \quad (29)$$

where $\mathcal{X}_i = \sqrt{\lambda_i} U_i V_i^\top$ are rank 1 matrices, called *elementary matrices*. The triple $(\sqrt{\lambda_i}, U_i, V_i)$ is called an *eigentriple* (ET) of the SVD and the vectors $\sqrt{\lambda_i} V_i = \mathcal{X}^\top U_i$ are the *principal components*.

- (iv) The numpy implementation of SVD is called `numpy.linalg.svd`.

Singular spectrum analysis

- The reconstruction stage is performed in two steps. First, we partition the set of indices $I = \{1, \dots, r\}$ into m disjoint subsets $I = I_1 \cup \dots \cup I_m$. For each subset I_k , form the sum

$$\mathcal{X}_{I_k} = \sum_{i \in I_k} \mathcal{X}_i. \quad (30)$$

Clearly, this defines a decomposition of the trajectory matrix into components:

$$\mathcal{X} = \mathcal{X}_{I_1} + \dots + \mathcal{X}_{I_m}. \quad (31)$$

- The final step is *diagonal averaging*. Each matrix \mathcal{X}_{I_k} in the decomposition (31) is transformed into a new *reconstructed time series* $(\tilde{X}_1^{(k)}, \tilde{X}_2^{(k)}, \dots, \tilde{X}_T^{(k)})$ by means of the following procedure.

Singular spectrum analysis

- Let A be an $l \times k$ -matrix, and let $T = l + k - 1$. We denote

$$A_{ij}^* = \begin{cases} A_{ij}, & \text{if } l < k, \\ A_{ji}, & \text{otherwise.} \end{cases} \quad (32)$$

Diagonal averaging transforms the matrix A into a time series $\tilde{A}_1, \dots, \tilde{A}_T$ as follows:

$$\tilde{A}_j = \begin{cases} \frac{1}{j} \sum_{m=1}^k A_{m,j-m+1}^*, & \text{for } 1 \leq j < l \wedge k, \\ \frac{1}{l \wedge k} \sum_{m=1}^{l \wedge k} A_{m,j-m+1}^*, & \text{for } l \wedge k \leq j \leq l \vee k, \\ \frac{1}{N-j+1} \sum_{m=k-l \vee k+1}^{T-l \vee k+1} A_{m,j-m+1}^*, & \text{for } l \vee k \leq j \leq T. \end{cases} \quad (33)$$

- As a result, the original time series is represented as a sun of m reconstructed series;

$$X_t = \sum_{i=1}^m \tilde{X}_t^{(i)}. \quad (34)$$

Singular spectrum analysis

- The choice of the rolling window length l is an important matter.
- It should be sufficiently large so that each lagged time series incorporates the essential features of the original series X_1, \dots, X_N .
- It is a good idea to perform SSA with different choices of l .

SSA of a simulated $I(1)$ process

- The figure below shows the results of SSA of the simulated $I(1)$ process given by the following specification:

$$X_t = 1.1 + X_{t-1} + 5.0\varepsilon_t, \quad (35)$$

where $\varepsilon_t \sim N(0, 1)$.

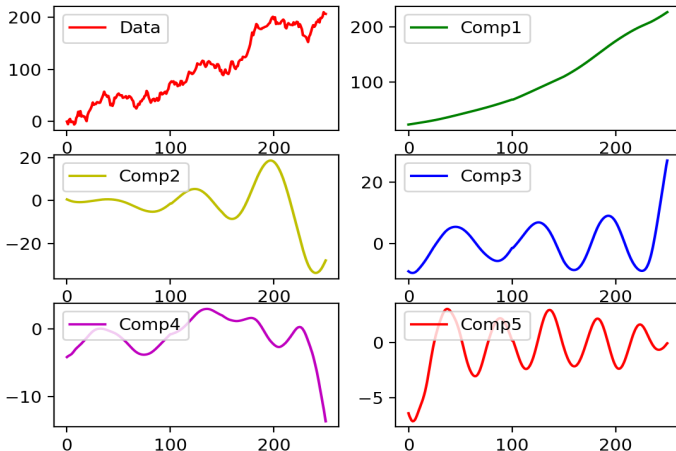
- The upper left plot shows the actual time series, while the remaining ones show the first five SSA components.
- The cumulative weights, defined as

$$CW_j = \frac{\lambda_1 + \dots + \lambda_j}{\lambda_1 + \dots + \lambda_l}, \quad (36)$$

of the plotted components are:

$$\begin{aligned} CW_1 &= 0.595, \\ CW_2 &= 0.653, \\ CW_3 &= 0.698, \\ CW_4 &= 0.720, \\ CW_5 &= 0.737. \end{aligned} \quad (37)$$

SSA of a simulated $AR(1)$ process



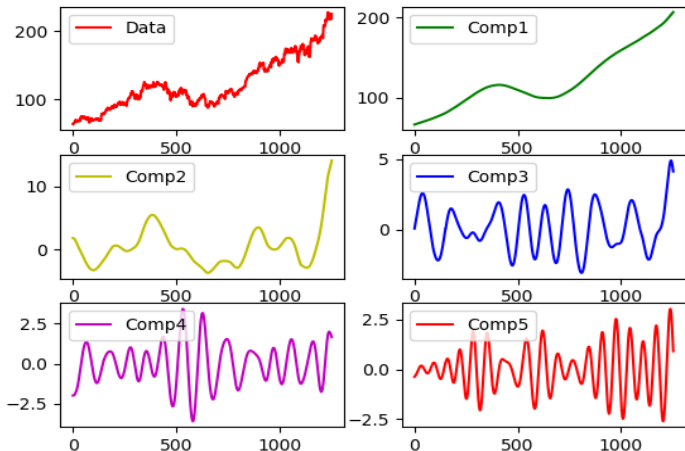
SSA of the AAPL share price

- The next figure below shows the results of SSA of the share price of Apple (AAPL) during the five-year period ending on September 28, 2018.
- As before, the upper left plot shows the actual time series, while the remaining ones show the first five SSA components.
- The weights of the plotted components are:

$$\begin{aligned}W_1 &= 0.769, \\W_2 &= 0.033, \\W_3 &= 0.016, \\W_4 &= 0.013, \\W_5 &= 0.011.\end{aligned}\tag{38}$$

- Notice that the first component (trend) is responsible for 76.9% of the dynamics.

SSA of the AAPL share price

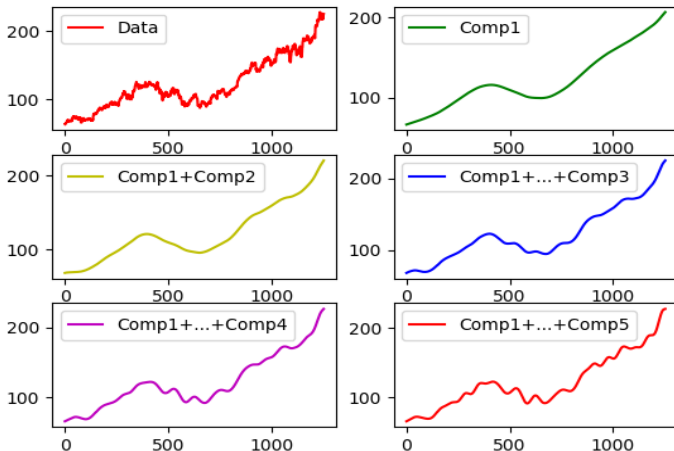


SSA of the AAPL share price

- Finally, the next figure shows the cumulative components of the dynamics of AAPL.
- The cumulative weights of the plotted components are:

$$\begin{aligned}CW_1 &= 0.769, \\CW_2 &= 0.802, \\CW_3 &= 0.818, \\CW_4 &= 0.831, \\CW_5 &= 0.852.\end{aligned}\tag{39}$$

SSA of the AAPL share price



Measuring long term memory in time series

- An alternative way (to the ACF) of measuring long range dependence (“memory”) in a time series is the *Hurst exponent*.
- The hydrologist Harold Edwin Hurst introduced this concept in 1951, when he was investigating the question of how to regularize the flow of the Nile River.
- Since ancient times, the Nile River has been known for its characteristic long-term behavior: long periods of drought were followed by long periods of recurring floods. Floods fertilized the soil so that in flood years the yield of crop was abundant. The account of this phenomenon in the Bible was: “Seven years of great abundance are coming throughout the land of Egypt, but seven years of famine will follow them”.
- The Nile River exhibits long periods with high maximal water level and long periods of low levels. The time series of levels does not show a global trend and appears stationary, even though, over short time periods, there seem to be cycles and local trends.
- Quantitatively, the phenomenon of long term dependence can be formulated in terms of *rescaled range (R/S) analysis*. Consider a time series of data X_t .

Measuring long term memory in time series

- Let us fix t and consider a time series of length k , $X = X_{t+1}, \dots, X_{t+k}$. The *rescaled range* (or the *R/S statistics*) is defined as follows:

- (i) Calculate the mean $m_{t,k} = \frac{1}{k} \sum_{i=1}^k X_{t+i}$.
- (ii) Define the de-meanned series: $Y_{i,t,k} = X_{t+i} - m_{t,k}$, for $i = 1, \dots, k$.
- (iii) Form the cumulative series of deviations: $Z_{i,t,k} = \sum_{j=1}^i Y_{j,t,k}$ for $i = 1, \dots, k$.
- (iv) Compute the range R of the cumulative deviations:

$$R(t, k) = \max(Z_{1,t,k}, Z_{2,t,k}, \dots, Z_{k,t,k}) - \min(Z_{1,t,k}, Z_{2,t,k}, \dots, Z_{k,t,k}).$$

- (v) Compute the standard deviation S :

$$S(t, k) = \sqrt{\frac{1}{k} \sum_{i=t+1}^{t+k} (X_i - m_{t,k})^2}.$$

- (vi) The rescaled range is defined as $Q(t, k) = R(t, k)/S(t, k)$.

Measuring long term memory in time series

- If, for $k \rightarrow \infty$, Q is scattered around a straight line,

$$\log Q(t, k) \approx a + H \log k, \quad (40)$$

we call the coefficient H the Hurst exponent.

- In probabilistic terms (assuming ergodicity) this means that

$$\log E(Q(t, k)) \approx a + H \log k. \quad (41)$$

- We say that the time series exhibits long term memory, if $H > 1/2$.
- This is to be contrasted with a random walk, for which $H = 1/2$.
- Hurst's discovery that for the Nile River data R/S behaves like a constant times k^H , with $H > 1/2$, is known as the *Hurst effect*.

Measuring long term memory in time series

- In finance, evaluating the Hurst exponents of asset prices may help to formulate trend based strategies:
 - (i) $H < 1/2$ (short memory) is typical of mean reverting assets,
 - (ii) $H = 1/2$ is a random walk,
 - (iii) $H > 1/2$ (long memory) is characteristic of trending assets.
- Like any other indicator, the Hurst exponent may work sometimes, but it may not work at all some other times.
- Also, a strong pattern could (and eventually will) reverse sharply, without prior warning.
- A good general reference on long term memory processes is [1].

Granger causality

- Time series analysis allows for a concept of causality that help to determine whether one time series is useful in forecasting another one.
- This concept, the *Granger causality*, is a statistical test formulated as follows.
- Consider two time series X_t and Y_t , which are assumed to follow the model:

$$\begin{aligned}X_t &= \alpha_X + \beta_{XX}X_{t-1} + \beta_{XY}Y_{t-1} + \varepsilon_t, \\Y_t &= \alpha_Y + \beta_{YX}X_{t-1} + \beta_{YY}Y_{t-1} + \eta_t,\end{aligned}\tag{42}$$

where ε_t and η_t are (possibly correlated) white noises.

- If $\beta_{XY} = 0$, there is a directional asymmetry between the two time series: X_t serves as an input variable, while Y_t is an output variable.
- In other words, Y_t does not influence X_t , while X_t influences Y_t .

Granger causality

- Consider now two k -period forecasts of X_t in the model above:

$$X_{t+k|f}^* = E(X_{t+k} | X_{1:t}, Y_{1:t}),$$

$$X_{t+k|p}^* = E(X_{t+k} | X_{1:t}).$$

- In other words, $X_{t+k|f}^*$ is the forecast based on full information up time t , while $X_{t+k|p}^*$ is based only on the (partial) information generated by the first series.

Granger causality

- Let $\varepsilon_{k|f}$ and $\varepsilon_{k|p}$ denote the corresponding forecasting errors.
- Then, we say that Y_t Granger-causes X_t , if

$$\text{Var}(\varepsilon_{k|f}) < \text{Var}(\varepsilon_{k|p}). \quad (43)$$

- In other words, Y_t Granger-causes X_t , if the quality of the forecast of X_t is improved by including the information about Y_t , namely, the variance of the residuals declines.

Granger causality

- The null hypothesis in the Granger test is *no Granger causality*, i.e.
 $H_0 : \beta_{XY} = 0$ in (42).
- This is essentially an F -test on the regression coefficient (having no explanatory power).
- One can generalize the discussion above to specifications with larger number of lags $p \geq 2$.
- In this case, the null hypothesis cannot be rejected, if no lagged values of Y_t are retained in the regression.

Granger causality

- Note that Granger causality is not necessarily identical with physical causality. Granger causality is observed only indirectly, through a time series.
- If both component processes X_t and Y_t are driven by common (but not explicitly specified) additional variables (processes), one might still be likely not to reject the alternative hypothesis of Granger causality.
- Such additional variables are referred to as *confounding variables*.
- The Granger test is designed to handle two variables, and may result in incorrect conclusions when the underlying relationship involves confounding variables that causally affect both variables X_t and Y_t .

Entropy

- The concept of Granger causality can be reformulated in terms of information transfer between two time series, using the concept of *transfer entropy*.
- Transfer entropy is defined in an essentially model free manner, lending itself to time series models beyond the *ARIMA* family of time series.
- The price for the model freeness is a bit of formalism required.
- This formalism, the entropy methods, have actually very important applications in finance (and broad data science), and we will take this as an opportunity to review them in some detail.

Entropy

- In order to lighten up on the math, we will sometimes be assuming that, for each t , X_t can take on only one of finitely many state values in $A = \{x_1, \dots, x_K\}$.
- The probability of each of the states is denoted by p_i , $p_i = P(X_t = x_i)$. Clearly,

$$\sum_{i=1}^K p_i = 1.$$

- More generally, consider first a discrete random variable X , and let $p = (p_1, \dots, p_K)$, $p_i = P(X = x_i)$, denote its probability distribution.
- The *Shannon entropy* of the random variable X is defined by:

$$H(X) = - \sum_{i=1}^K p_i \log p_i. \quad (44)$$

Entropy

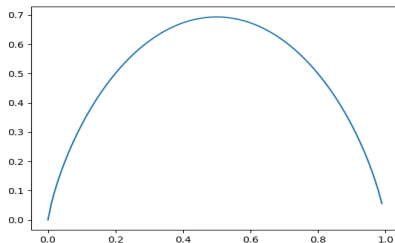
- The quantity $\log \frac{1}{p_i}$ is called the *Shannon information* (expressed in *nats*) about the k -th outcome of the random variable X .
- Shannon entropy is the average information and can be interpreted as an overall measure of information contained in the probability distribution of X .
- The lower the entropy, the higher the information content of the probability distribution.
- The Shannon entropy has the following properties:
 - (i) It is always nonnegative.
 - (ii) Its value is 0, if one of the p_i 's is 1.
 - (iii) It reaches its maximum value $\log K$, if the distribution is uniform, $p_i = 1/K$, for all $i = 1, \dots, K$.

Entropy

- As an example, consider the case of a binary random variable, $K = 2$.
- Then $p = (w, 1 - w)$ and its entropy is given by the function:

$$h(w) = -w \log(w) - (1 - w) \log(1 - w). \quad (45)$$

- Its graph is given below:



Entropy

- Is $X \in \mathbb{R}^n$ is a random variable with a continuous probability distribution $p(x)$, its Shannon entropy is defined by

$$H(X) = - \int p(x) \log p(x) d^n x. \quad (46)$$

- Its properties are similar to the properties of the entropy of a discrete random variable, except that it is **not** necessarily non-negative.
- For this reason, it is often referred to as the *differential entropy*.

Entropy

- For example, if $X \sim N(\mu, \sigma^2)$ is a normal random variable, then

$$H(X) = \frac{1}{2} \log(2\pi e\sigma^2). \quad (47)$$

- In general, if $X \sim N(\mu, \Sigma)$ is a multivariate Gaussian random variable, then

$$H(X) = \frac{1}{2} \log \det(2\pi e\Sigma). \quad (48)$$

Joint and conditional entropy

- Assume now that we have a joint (discrete) probability distribution

$$p_{i,j} = P(X = x_i, Y = y_j), \text{ for } i = 1, \dots, K_1, \text{ and } j = 1, \dots, K_2,$$

of two random variables X and Y .

- The *joint entropy* of X and Y is defined as

$$H(X, Y) = - \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} p_{i,j} \log p_{i,j}. \quad (49)$$

Joint and conditional entropy

- Let $p_{i|j} = P(X = x_i | Y = y_j)$ denote the conditional probability distribution of X given Y .
- The *conditional entropy* of X given Y is defined as

$$H(X|Y) = - \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} p_{i,j} \log p_{i|j}. \quad (50)$$

- The conditional entropy measures the information content in the probability distribution of X given the knowledge of Y .
- If X and Y are independent, then $H(X|Y) = H(X)$.

Joint and conditional entropy

- The joint and conditional entropies are related as follows:

$$H(X, Y) = H(Y) + H(X|Y). \quad (51)$$

- Proof:*

$$\begin{aligned} H(X, Y) &= - \sum_i \sum_j p_{i,j} \log p_{i,j} \\ &= - \sum_i \sum_j p_{i,j} \log p_{i|j} p_j \\ &= - \sum_i \sum_j p_{i,j} \log p_{i|j} - \sum_i \sum_j p_{i,j} \log p_j \\ &= - \sum_i \sum_j p_{i,j} \log p_{i|j} - \sum_j p_j \log p_j \\ &= H(X|Y) + H(Y). \end{aligned}$$

Kullback-Leibler divergence

- Suppose now $q = (q_1, \dots, q_K)$, $j = 1, \dots, K$, is another probability distribution of the random variable X .
- This could possibly be an *a priori* guess of p or a parametric model of p .
- A measure of distance (or “divergence”) between the distributions p and q , is given by the *Kullback-Leibler divergence*, a.k.a. *relative entropy*:

$$\text{KL}(p\|q) = \sum_{i=1}^K p_i \log \frac{p_i}{q_i}. \quad (52)$$

- It is widely used in machine learning and data science in model training / calibration.

Kullback-Leibler divergence

- For example, in the binary case, $p = (w, 1 - w)$, $q = (v, 1 - v)$,

$$\text{KL}(p\|q) = w \log \frac{w}{v} + (1 - w) \log \frac{1 - w}{1 - v} .$$

- For continuous probability distributions $p(x)$ and $q(x)$, their Kullback-Leibler divergence is defined by the integral

$$\text{KL}(p\|q) = \int p(x) \log \frac{p(x)}{q(x)} dx . \quad (53)$$

Kullback-Leibler divergence

- The Kullback-Leibler divergence has the following properties (valid in both discrete and continuous cases):
 - (i) $KL(p||q) \geq 0$.
 - (ii) $KL(p||q) = 0$ if and only if $p = q$.
- The proof is not hard, but it uses some properties of convex functions, which I will cover in detail in the Optimization Techniques in Finance course.
- Note that, unlike the conventional measure of distance, the Kullback-Leibler divergence is not symmetric in its arguments: $KL(p||q) \neq KL(q||p)$.

Mutual information

- The mutual information between two random variables is defined by

$$I(X; Y) = H(X) + H(Y) - H(X, Y). \quad (54)$$

- $I(X; Y)$ has the following properties:

(i)

$$I(X; Y) = I(Y; X). \quad (55)$$

(ii)

$$I(X; Y) = H(X) - H(X|Y). \quad (56)$$

(iii)

$$I(X; Y) = H(Y) - H(Y|X). \quad (57)$$

(iv)

$$I(X; X) = H(X). \quad (58)$$

- *Proof:* Relations (56) and (57) are consequences of (51). The other relations are obvious.

Mutual information

- Mutual information measures the increase of information about X due to the available information about a random variable Y .
- If X and Y are independent, then $H(X|Y) = H(X)$ and $I(X; Y) = 0$ (nothing learned about X from Y).
- The mutual information of X and Y can explicitly be expressed in the form:

$$I(X; Y) = \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} p_{i,j} \log \frac{p_{i,j}}{p_i p_j}, \quad (59)$$

which is the same as the Kullback-Leibler divergence between the joint distribution $p_{X,Y}$ and the product distribution $p_X p_Y$,

$$I(X; Y) = \text{KL}(p_{X,Y}, p_X p_Y). \quad (60)$$

Mutual information

- In other words, the mutual information of two random variables is a measure of distance between their joint distribution and the product of their respective marginals.
- In particular, mutual information is non-negative,

$$I(X; Y) \geq 0. \quad (61)$$

- Finally, the *conditional mutual information* of X, Y given Z is defined by

$$I(X, Y|Z) = H(X|Z) - H(X|Y, Z). \quad (62)$$

Measuring time series entropy

- Measuring (estimating) various types of time series entropies poses challenges and is still subject of ongoing research.
- A number of algorithms have been developed for stationary time series.
- The MLE method leads to the following *plug-in* estimator.
- Let $\hat{p}_i = \frac{n_i}{N}$ denote the observed frequency of the i -th outcome x_i of the variable X_t , calculated from N observations.
- The plug-in estimator of the Shannon entropy is given by

$$\hat{H}(X) = - \sum_{i=1}^K \hat{p}_i \log \hat{p}_i. \quad (63)$$

- The plug-in estimator is a biased estimator, and tends to underestimate the true entropy:

$$E(\hat{H}(X)) \leq H(X). \quad (64)$$

Measuring time series entropy

- A better estimator for the continuous (differential) Shannon entropy is the *Kozachenko-Leonenko (KL) estimator*.
- We order the observations x_i of X so that $x_1 < x_2 < \dots < x_N$, and set

$$\hat{H}(X) = \psi(N) - \psi(1) + \frac{1}{N-1} \sum_{i=1}^{N-1} \log(x_{i+1} - x_i). \quad (65)$$

- Here, $\psi(x)$ denotes the *digamma function*,

$$\psi(x) = \frac{d}{dx} \log \Gamma(x), \quad (66)$$

where $\Gamma(x)$ denotes Euler's gamma function

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt. \quad (67)$$

- The digamma function is neatly implemented in `scipy.special`.

Measuring time series entropy

- The standard (and highly effective) algorithm for estimating the mutual information of a (continuous) distribution is the *Kraskov, Stögbauer and Grassberger (KSG) algorithm* [6].
- It is essentially an extension of the ideas behind the KL algorithm, and I refer you the original paper or a summary in [2].
- Care has to be exercised when attempting to estimate the entropy of a non-stationary time series.

Transfer entropy

- Assume now that we are analyzing a time series model X_t .
- The concepts developed above can be applied to various random variables related to X_t , lagged values of X_t , etc.
- For example:
 - (i) The Shannon entropy of X_t is $H(X_t)$.
 - (ii) The mutual information of two time series X_t and Y_t is $I(X_t; Y_t)$.
 - (iii) An entropy measure capturing the dynamics of the time series over the period of j lags is given by $H(X_t | X_{t-j:t-1})$.

Transfer entropy

- The *transfer entropy* from the process Y_t to X_t is defined as the mutual information of X_t and $Y_{t-j:t-1}$ conditioned on $X_{t-j:t-1}$:

$$T(Y \rightarrow X) = I(X_t, Y_{t-j:t-1} | X_{t-j:t-1}). \quad (68)$$

- In other words, transfer entropy from Y_t to X_t measures the increase of information of X_t due to the inclusion of lagged information about Y_t , given lagged information about X_t .
- This also can be rewritten as

$$T(Y \rightarrow X) = H(X_t | X_{t-j:t-1}) - H(X_t | X_{t-j:t-1}, Y_{t-j:t-1}). \quad (69)$$

- In the simplest case, if the two processes X_t and Y_t are independent, then, for any number of lags j ,

$$p(x_t | x_{t-j:t-1}, y_{t-j:t-1}) = p(x_t | x_{t-j:t-1}), \quad (70)$$

and $T(Y \rightarrow X) = 0$.

Transfer entropy

- In the case of a discrete valued process,

$$T(Y \rightarrow X) = \sum_{x_{t-j:t}} \sum_{y_{t-j:t-1}} p(x_{t-j:t}, y_{t-j:t-1}) \log \frac{p(x_t | x_{t-j:t-1}, y_{t-j:t-1})}{p(x_t | x_{t-j:t-1})}. \quad (71)$$

- Transfer entropy is a very elegant and economic concept of causal dependence among time series.
- It applies to time series models that are not necessarily linear, or whose residuals are necessarily normally distributed.
- In case of autoregressive models with normally distributed disturbances, transfer entropy is essentially identical with the statistics used to test Granger causality.
- Namely, consider the single lag model (42) discussed above.

Transfer entropy and Granger causality

- Transfer entropy with $j = 1$ is given by

$$\begin{aligned} T(Y \rightarrow X) &= I(X_t, Y_{t-1} | X_{t-1}) \\ &= H(X_t | X_{t-1}) - H(X_t | X_{t-1}, Y_{t-1}) \end{aligned} \quad (72)$$

- The terms on the right hand side can be evaluated by an explicit calculation. The result turns out to be [2]

$$T(Y \rightarrow X) = \frac{1}{2} \log \frac{\text{Var}(\varepsilon_{1|f})}{\text{Var}(\varepsilon_{1|p})}. \quad (73)$$








Up to the constant $\frac{1}{2}$, this is the statistics used in the Granger causality test.

- The same concepts can be extended to multivariate time series (which we will study later), with the corresponding increase in the notational complexity.

Transfer entropy and Granger causality

- Estimation of transfer entropy from observed data is a bit of a challenge, as reliable estimates require large sample sets.
- Unlike the Granger test, which is a test on a linear regression coefficient, estimating transfer entropy requires information on the probability distributions of the processes.
- For algorithms and references I refer you to the book [2].

References

-  [1] Beran, J.: *Statistics for Long-Memory Processes*, Chapman & Hall (1994).
-  [2] Bossomaier, T., Barnett, L., Harre, M., and Lizier J. T.: *An Introduction to Transfer Entropy*, Springer (2016).
-  [3] Cover, T. M., and Thomas, J. A.: *Elements of Information Theory*, Wiley (2006).
-  [4] Golyandina, N., and Zhiglyavsky, A.: *Singular Spectrum Analysis for Time Series*, Springer (2013).
-  [5] Hamilton, J. D.: *Time Series Analysis*, Princeton University Press (1994).
-  [6] Kraskov, A., Stögbauer, H., and Grassberger, P.: Estimating mutual information, *Phys. Rev.*, **E69**, 066138 - 066153 (2004).
-  [7] Schreiber, T.: Measuring Information transfer, *Phys. Rev. Lett.*, **85**, 461 - 464 (2000).