WP-2025-020

MAXIMUM ENTROPY SPECTRAL ANALYSIS

Dilip M.Nachane

Indira Gandhi Institute of Development Research, Mumbai

July 2025

MAXIMUM ENTROPY SPECTRAL ANALYSIS

Dilip M.Nachane

Email(corresponding author): nachane@igidr.ac.in

Abstract

The *maximum entropy principle* is characterized as assuming the least about the unknown parameters in a statistical model. In its applied manifestations, it uses all the available information and makes the fewest possible assumptions regarding the unavailable information. The application of this principle to parametric spectrum estimation leads to an autoregressive transfer function. By appeal to a well known theorem in stochastic processes, a rational transfer function leads to *a factorizable spectrum*. This result combined with a classical theorem of analysis (due to Szegö) forms the basis for two important algorithms for estimating the autoregressive spectrum viz. the Levinson-Durbin and Burg algorithms. The latter leads to estimators which are asymptotically MLEs (maximum likelihood estimators).

Keywords: Entropy—Jaynes' Principle-autoregressive spectrum-spectral factorization-Levinson-Durbin-Burg-MLE

JEL Code: C22, C32

I. Introduction

Let us begin by considering the case of a *discrete random variable* $\{X\}$ which can take on n distinct values $x_j, j = 1 \dots n$ such that $P\{X = x_j\} = p_j$. Assuming the events to be mutually exclusive and exhaustive, we have the side constraints $p_j \geq 0$, $j = 1 \dots n$ and $\sum_{j=1}^{n} p_j = 1$.

Definition 1: The *information* content I_m of the event $\{x_m, m = 1 \dots n\}$ in the above sample is defined as

$I_m = \log_2 \left(\frac{1}{p_m}\right)$ *bits* where the log is to the base 2. (I.1)

If the log is to the base *e*, we define

$I_m = (log_e 2) \log_2 \left(\frac{1}{p_m}\right)$ *nats.*

In future we will only consider information as measured in *bits*.

Definition 2: The expected value of the random variable $I = \{I_m\}_{m=1}^{n}$ is defined as the *entropy* of the sample and is given by

$H(X) = E(I) = -\sum_{j=1}^{n} p_j \, log(p_j)$ (I.2)

The rationale for defining *information* and *entropy* in terms of logs is threefold (se Shannon (1948)). Firstly, it closely corresponds to Boltzmann's (1866) definition of thermodynamic entropy of an *ideal gas.* Secondly, parameters of importance in statistical mechanics and signal processing such as resolution, bandwidth, octaves etc. tend to vary linearly with the logarithm of the number of *microstates* (i.e. *events* as understood above). The third rationale is a bit more complicated. It states that any definition of *information* I_j (of an event *j* with probability p_j) defined above should satisfy the following intuitively reasonable axioms:

(i) I_j is non-negative and *anti-monotonic* i.e. the information $I_j \geq 0$ and I_j increases when p_j decreases.

(ii) I_j is undefined if $p_j = 0$ i.e. the information content of an impossible event cannot be defined.

(iii) I_j=0 if $p_j = 1$ i.e. the information content of a certain event is zero.

(iv) If two events i and j are independent, then their joint information content denoted $I_{i \cap j} = I_i + I_j$ (or if $p_{i \cap j}$ denotes the probability of the joint occurrence of i and j then $p_{i \cap j} = p_i + p_j$

The same conditions are phrased in terms of the corresponding concept of entropy by Shannon (1948). Shannon (op.cit. p. 419-420) then proves that the only functions satisfying the above reasonable conditions on information and entropy respectively are given by (I.1) and (I.2).

There is an intimate connection between the concepts of entropy, information and uncertainty. Consider the discrete probability distribution stated above in which the *random variable* $\{X\}$ takes on n distinct values $x_j, j = 1 \ldots n$ such that $P\{X = x_j\} = p_j$. The most informative distribution would occur when one of the values x_m was known to be true (viz. $P\{X = x_m\} = 1$). In that case, the entropy would be equal to zero. The least informative distribution would occur when there is no reason to favor any one of the propositions over the others and each $p_j = \left(\frac{1}{n}\right)$. In that case, the entropy would be equal to its maximum possible value viz. $nlog(n)$. The entropy can therefore be seen as a numerical measure which describes how uninformative a particular probability distribution is, ranging from zero (completely informative) to $nlog(n)$ (completely uninformative).

Definition 3**:** The concepts of *joint entropy* and *conditional entropy* are straightforward generalizations of Definition 2 above. Thus if $\boldsymbol{X} = \{X_1, X_2 \ldots . X_N\}$ is a collection of N discrete random variables with joint pdf $P(X_1, X_2 \ldots . X_N)$ then the *joint entropy* of this collection is defined as

$$H(X_1, X_2 \ldots . X_N) = -\sum_{X_1} \ldots \sum_{X_N} P(X_1, X_2 \ldots . X_N) Log\ [P(X_1, X_2 \ldots . X_N)] \tag{I.3}$$

while the *conditional entropy* of $(X_1, X_2 \ldots X_{k-1}, X_{k+1} \ldots X_N)$ given $\boldsymbol{X} = x_k$ is given by

$$H(X_1, X_2 \ldots X_{k-1}, X_{k+1} \ldots X_N | \boldsymbol{X} = x_k) = \\ -\sum_{X_1} \ldots \sum_{X_N} P(X_1, X_2 \ldots . X_N) Log\ [P(X_1, X_2 \ldots X_{k-1}, X_{k+1} \ldots X_N | \boldsymbol{X} = x_k)] \tag{I.4}$$

So far we have been concerned with *discrete* random variables. The generalization to *continuous* random variables simply proceeds by replacing the discrete summations in (I.1) to (I.4) above by integrals. We only present the definition for joint entropy, the other concepts can then be easily written.

Definition 4**:** If $\boldsymbol{X} = \{X_1, X_2 \ldots . X_N\}$ is a collection of N *continuous* random variables with joint pdf $f(\boldsymbol{X})$ then the *joint entropy* of this collection is defined as

$$H\big(f(\boldsymbol{X})\big) = -\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f(\boldsymbol{X}) Log\ [f(\boldsymbol{X})] d\boldsymbol{X} = -E\{Log\ [f(\boldsymbol{X})]\} \tag{I.5}$$

We now turn to the important concept of a stochastic process which is defined as follows.

Definition 5**:** A stochastic process $\boldsymbol{X}$ is a collection of random variables $\{X_t, t \in T\} = \{X_t(w), t \in T, w \in \Omega\}$ where T is an index set and (Ω, S, P) is a *probability space* with *probability measure* P. The index set T is usually taken as $T = (-\infty, \infty)$. A *Finite Dimensional Distribution (FDD)* of this stochastic process $\boldsymbol{X}$ simply refers to the *joint distribution function* of any *finite sub-collection* of random variables constituting $\boldsymbol{X}$. We say that a stochastic process $\boldsymbol{X}$ has a well-defined distribution function if the FDDs of any finite sub-collection of $\boldsymbol{X}$ say $\left(X_{t_1} \cdots X_{t_n}\right)$, $t_1 \cdots t_n \in T$ exist for all choices of $t_1 \ \ldots \ t_n \in T$ and $n \geq 1$. The collection of all such FDDs then is said to constitute the $distribution\ function\ of\ X$.

For a stochastic process as defined above, the relevant entropy concept is *entropy rate* as defined below (see Cover and Thomas (2006), p.74-75)

Definition 6 : A stochastic process $\boldsymbol{X}$ is said to be Gaussian if all its FDDs are *multivariate Gaussian* (see Priestley (1981, p.101-104), Cox and Small (1978), Nachane (2006, p. 497) etc.)

Definition 7 : The *entropy rate* of a stochastic process $\boldsymbol{X}$ (as in Definition 5) is defined as

$$\mathcal{E} = \lim_{n\to\infty} \left(\frac{1}{n}\right) H\left(X_{t_1} \cdots X_{t_n}\right) \tag{I.6}$$

if this limit exists. Here $H\left(X_{t_1} \cdots X_{t_n}\right)$ is the *joint entropy* of $\left(X_{t_1} \cdots X_{t_n}\right)$ as defined in (I.5).

II. Spectral Factorization

As a preliminary to the subsequent discussion, we need to invoke the basic theory of *spectral factorization.*

Definition 8 : The *one-sided z-transform* of a discrete process[1]

$\boldsymbol{X}(t) = \{X(0), X(1), X(2) \ldots\}$ is defined as

$$\boldsymbol{X}(z) = \mathcal{Z}[\boldsymbol{X}(t)] = \sum_{k=0}^{\infty} X(k)\, z^{-k},\ z\, \epsilon\, \mathcal{C} \tag{II.1}$$

A special class of z-transforms are the *rational* z-transforms i.e. those which can be expressed as the ratio of two finite polynomials viz.

$$\boldsymbol{X}(z) = \frac{\boldsymbol{B}(z)}{\boldsymbol{A}(z)} = \frac{\sum_{k=0}^{M_b} b_k z^{-k}}{\sum_{p=0}^{M_a} a_p z^{-p}},\quad M_b \leq M_a \tag{II.2}$$

The z-transform possesses several interesting properties (see Johnson (2012), p.101-107). Of these, the *time-shift* property is of special relevance. It states that

$$\mathcal{Z}[\boldsymbol{X}(t-k)] = z^{-k}\mathcal{Z}[\boldsymbol{X}(t)] = z^{-k}\boldsymbol{X}(z) \tag{II.3}$$

Property (II.3) enables us to write the following ARMA (N_a, N_b) model

$$\sum_{j=0}^{N_a} a_j X(t-j) = \sum_{j=0}^{N_a} b_j \epsilon(t-j) \text{ with } a_0, b_0 = 1 \tag{II.4}$$

as

$$\left(\sum_{j=0}^{N_a} a_j z^{-j}\right) \boldsymbol{X}(z) = \left(\sum_{j=0}^{N_b} b_j z^{-j}\right) \boldsymbol{E}(z) \tag{II.5}$$

where $\boldsymbol{E}(t) = \{\epsilon(0), \epsilon(1), \epsilon(2) \ldots\}$ is the error process.

Definition 9 :The *transfer function* of the ARMA model (II.5) is defined as the rational function

[1] We can also define a *two-sided* z-transform (see Candy (1988), p.16).

$$\boldsymbol{H}(z) = \frac{\boldsymbol{B}(z)}{\boldsymbol{A}(z)} = \frac{\sum_{k=0}^{N_b} b_k z^{-k}}{\sum_{k=0}^{N_a} a_k z^{-k}}, \quad N_b \leq N_a \tag{II.6}$$

It is easily seen that the *transfer function* of the AR model $(b_k = 0, k = 1 \ldots N_b)$ and of the MA model $(a_k = 0, k = 1 \ldots N_b)$ can be written respectively as

$$\boldsymbol{H}(z) = \frac{\boldsymbol{B}(z)}{\boldsymbol{A}(z)} = \frac{1}{\sum_{k=0}^{N_a} a_k z^{-k}} \tag{II.7}$$

$$\boldsymbol{H}(z) = \frac{\boldsymbol{B}(z)}{\boldsymbol{A}(z)} = \sum_{k=0}^{N_b} b_k z^{-k} \tag{II.8}$$

Definition 10 : The *power spectral density* (PSD) $s_X^{ARMA}(z)$ of an ARMA process $\boldsymbol{X}(t)$ such as (II.6) is defined as

$$s_X^{ARMA}(z) = \boldsymbol{H}(z)\boldsymbol{H}^*(z) = |\boldsymbol{H}(z)|^2 S_\epsilon(z) \tag{II.9}$$

where the superscript (*) denotes complex conjugate and $S_\epsilon(z)$ is the spectrum of the process $\boldsymbol{E}(t)$. If (as is usually the case) $\boldsymbol{E}(t)$ is a white noise process with variance σ_ϵ^2, then $S_\epsilon(z) = \sigma_\epsilon^2$ and we can write (II.9) as

$$s_X^{ARMA}(z) = |\boldsymbol{H}(z)|^2 \sigma_\epsilon^2 = \left|\frac{\boldsymbol{B}(z)}{\boldsymbol{A}(z)}\right|^2 \sigma_\epsilon^2 \tag{II.10}$$

The spectrum of the ARMA model can be written in terms of the z-transform (as in (II.10)) or more commonly in terms of the *angular velocity* ω defined by $z = e^{i\omega}$ i.e.

$$s_X^{ARMA}(\omega) = \left|\frac{\sum_{k=0}^{N_b} b_k e^{-ik\omega}}{\sum_{k=0}^{N_a} a_k e^{-ik\omega}}\right|^2 \sigma_\epsilon^2 \tag{II.11}$$

The spectrums of the AR and MA models are then respectively

$$s_X^{AR}(\omega) = \left|\frac{1}{\sum_{k=0}^{N_a} a_k e^{-ik\omega}}\right|^2 \sigma_\epsilon^2 \tag{II.12}$$

$$s_X^{MA}(\omega) = \left|\sum_{k=0}^{N_b} b_k e^{-ik\omega}\right|^2 \sigma_\epsilon^2 \tag{II.13}$$

We are now in a position to state the Spectral Factorization theorem (Sayed and Kailath (2001))

Theorem 1: Let $S_X(z)$ denote the *power spectral density* (PSD) of a stochastic process $\boldsymbol{X}(t)$ which is stationary and has a *rational transfer function* $H(z)$ (see Definitions 8 and 9) . Additionally it satisfies the following conditions : (i) $S_X(z) > 0$ for $|z| = 1$ or in terms of ω(angular velocity)

$S_X\left(e^{i\omega}\right) > 0$ for $\omega \in (-\pi, \pi)$ and (ii) $\left(\frac{1}{2\pi}\right)\int_{-\pi}^{\pi} \ln\left[S_X\left(e^{i\omega}\right)\right] d\omega > -\infty$ (Paley-Wiener condition).

Then $S_X(z)$ admits the following factorization

$$S_X(z) = L(z)\beta_X L^*(z^{-1}) \tag{II.14}$$

where $L(z)$ is a finite degree polynomial with *minimum phase* i.e. it has all its *zeros* and *poles* strictly inside the unit circle and $\lim_{|z|\to\infty} L(z) = 1$. Further β_X is a real positive scalar.

We can also state the result (II.14) in terms of the angular frequency ω as

$$S_X(e^{i\omega}) = L(e^{i\omega})\beta_X L^*(e^{-i\omega}) \qquad \text{(II.15)}$$

III. Maximum Entropy Principle

The maximum entropy principle is usually attributed to Brillouin (1956, p.159-161) and Jaynes (1963, 1968). The principle can be characterized in several ways , Jaynes (1968, p.229) defines the principle as the method which "assumes the least" about the unknown parameters. Ulrych and Bishop (1975, p.184) reformulate the principle as the method which uses all the available information (being the expected value of the random information variable I_m −−see (I. 2)above) and is "maximally noncommittal with regard to the unavailable information" (a point which becomes important when we later discuss applications of the principle to spectral analysis)

A. *Entropy of the Multivariate Gaussian Stochastic Process*

Burg (1967, 1968) who is primarily credited with the development of the *Maximum Entropy Spectral Method (MESM)* proceeds by first developing the MESM for the multivariate Gaussian Stochastic process. Consider a typical FDD of this process defined over the vector

$\boldsymbol{X} = \{X_1, X_2 \ldots . X_N\}$. By Definition 6, $\boldsymbol{X}$ is an N-dimensional Gaussian variate $\mathbb{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ whose joint pdf is given by

$$f(\boldsymbol{X}) = (2\pi)^{-N/2}|\boldsymbol{\Sigma}|^{-1/2} exp\left[-\frac{(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})}{2}\right] \qquad \text{(III.1)}$$

(where $\boldsymbol{\mu} = E(\boldsymbol{X})$ and $\boldsymbol{\Sigma}$ is the variance-covariance matrix with typical entry $\boldsymbol{\Sigma_{ij}} = Cov(X_i, X_j)$)

and the entropy is given by (I.5) as

$$H_N(\boldsymbol{X}) = -E\{Log\,[f(\boldsymbol{X})]\}$$

$$= \frac{N}{2}Log(2\pi) + \frac{1}{2}Log|\boldsymbol{\Sigma}| + \frac{1}{2}E[(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})] \qquad \text{(III.2)}$$

But

$$E[(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})] = E[tr(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})] \qquad \text{(III.3)}$$

(the quantity in brackets being a scalar)

But

$$E[tr(\boldsymbol{X}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})] = E[tr\{\boldsymbol{\Sigma}^{-1}(\boldsymbol{X}-\boldsymbol{\mu})'(\boldsymbol{X}-\boldsymbol{\mu})\}] \qquad \text{(III.4)}$$

(see e.g. Hohn (1964, p.16))

The r.h.s. of (III.4) can be written as

$$tr\{\boldsymbol{\Sigma}^{-1}E[(\boldsymbol{X}-\boldsymbol{\mu})'(\boldsymbol{X}-\boldsymbol{\mu})]\}= tr\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\} = tr(\boldsymbol{I}_N) = N \quad \text{(III.5)}$$

From (III.1) to (III.4) we get

$$H_N(\boldsymbol{X}) = \frac{N}{2}[Log(2\pi)+1] + \frac{1}{2}Log|\boldsymbol{\Sigma}| \quad \text{(III.6)}$$

The entropy rate for a Gaussian stochastic process is then

$$\mathcal{E} = \lim_{N\to\infty}\left(\frac{1}{N}\right)H_N(\boldsymbol{X}) = \frac{1}{2}[Log(2\pi)+1] + \lim_{N\to\infty} Log|\boldsymbol{\Sigma}|^{\frac{1}{2N}} \quad \text{(III.7)}$$

The first term on the r.h.s. is constant and will not play any role in the maximization process, and so we drop it and take the entropy rate as

$$\mathcal{E} = \lim_{N\to\infty} Log|\boldsymbol{\Sigma}|^{\frac{1}{2N}} \quad \text{(III.8)}$$

As the matrix $\boldsymbol{\Sigma}$ is both Hermitian and non-negative definite, its *N* eigenvalues

$\lambda_k \quad (k = 1 \dots N)$ are real and non-negative and further

$$|\boldsymbol{\Sigma}| = \prod_{k=1}^{N}\lambda_k \quad \text{(III.9)}$$

($|\boldsymbol{\Sigma}|$ denotes $(\det(\boldsymbol{\Sigma}))$)

From (III.8) and (III.9), we get

$$\mathcal{E} = \lim_{N\to\infty}\left(\frac{1}{2N}\right)\sum_{k=1}^{N} Log\,(\lambda_k) \quad \text{(III.10)}$$

B. Szegö's Theorem and Maximum Entropy of a Gaussian Process

We now invoke a classical theorem due to Szegö (1915) (see Gohberg and Kupnik (1969), Parter (1986) and Widom (1989) for expositions and extensions of the theorem)

Theorem 2 (Szegö): If $\lambda_k \quad (k = 1 \dots N)$ are real eigenvalues of a Toeplitz matrix[2] $T_N(f)$ associated with a bounded real-valued function f with entries on the unit circle then for any Riemann integrable function G, we have

$$\lim_{N\to\infty}\left(\frac{1}{2N}\right)\sum_{k=1}^{N} G\,(\lambda_k) = \left(\frac{1}{4\pi}\right)\int_{-\pi}^{\pi} G[f(\omega)]d\omega \quad \text{(III.11)}$$

Now let us take $G = Log(.)$ and let the function f be defined on the entries of Σ by

[2] A $(p \times q)$ matrix $\boldsymbol{A}$ is said to be Toeplitz if all the entries along each diagonal are equal i.e. A_{ij} depends only on $(i-j)$ (see Stoica and Moses (2015), p.362).

$$f(\omega) = \sum_{j=-\infty}^{\infty} R(j) e^{-i\omega j} = S(\omega) \quad \text{(III.12)}$$

where $S(\omega)$ is the power spectrum of $\boldsymbol{X}$ and ω is the angular frequency and

$$R(k) = Cov(X_j, X_{j+k}),$$

Combining (III.10) to (III.12), we get

$$\mathcal{E} = \left(\frac{1}{4\pi}\right) \int_{-\pi}^{\pi} Log\,[S(\omega)] d\omega \quad \text{(III.13)}$$

Now, we are already given $R(k), |k| \leq N$ and by the Maximum Entropy Principle we do not make any assumptions about $R(k), |k| > N$. Thus the entropy can be maximized only w.r.t. $R(k), |k| > N$

Thus, application of the Maximum Entropy Principle leads to

$$\frac{\partial \mathcal{E}}{\partial R(k)} = \left(\frac{1}{4\pi}\right) \int_{-\pi}^{\pi} \left(\frac{1}{S(\omega)}\right) \left(\frac{\partial S(\omega)}{\partial R(k)}\right) d\omega = \left(\frac{1}{4\pi}\right) \int_{-\pi}^{\pi} \left(\frac{1}{S(\omega)}\right) e^{-i\omega k} d\omega = 0\,,\ |k| > N \quad \text{(III.14)}$$

Putting

$$C(k) = \left(\frac{1}{2\pi}\right) \int_{-\pi}^{\pi} \left(\frac{1}{S(\omega)}\right) e^{-i\omega k} d\omega = \left(\frac{1}{2\pi}\right) \int_{-\pi}^{\pi} \left(\frac{1}{S(\omega)}\right) e^{i\omega k} d\omega, \quad \text{(III.15)}$$

we see that $C(k)$ are the Fourier Coefficients of the function $(1/(S(\omega))) = S^{-1}(\omega)$. Further, by (III.14), these Fourier coefficients, are non-zero only for $|k| \leq N$. The Fourier series expansion of $S^{-1}(\omega)$ is thus (by the Wiener-Khintchine theorem))

$$S^{-1}(\omega) = \sum_{k=-N}^{N} C(k) e^{-i\omega k} \quad \text{(III.16)}$$

which shows that $S^{-1}(\omega)$ is the power spectrum of a process with covariances $C(k)$. Additionally, as $S(\omega)$ is a power spectrum with a rational transfer function, $S^{-1}(\omega)$ will also have a rational transfer function and will additionally satisfy the premises of Theorem 1 above. Thus $S^{-1}(\omega)$ will admit a spectral factorization

$$S^{-1}(\omega) = M(z) \gamma M^*(z^{-1}) = \gamma |M(z)|^2 \quad \text{(III.17)}$$

Thus

$$S(\omega) = \gamma^{-1} \frac{1}{|M(z)|^2} \quad \text{(III.18)}$$

which from ((II.12)) implies that $S(\omega)$ is the spectrum of an AR process with the variance of the error term $\sigma_\epsilon^2 = \gamma^{-1}$

It is important to remember that the result (III.18) is crucially dependent on the Gaussianity assumption made in the previous section. However this not as restrictive as it sounds because of the following result due to Papoulis (1991, p. 575).

Theorem 3: Suppose $\boldsymbol{X}$ is an N-dimensional zero-mean variate $\mathbf{X} = \{\mathrm{X}_1, \mathrm{X}_2 \ldots . \mathrm{X}_\mathrm{N}\}$ with unknown joint pdf $f(\boldsymbol{X})$, but we are given its Var-Cov matrix $\boldsymbol{\Sigma}$ with typical entry

$\boldsymbol{\Sigma_{ij}} = Cov(X_i, X_j)$. Additionally $\boldsymbol{\Sigma}$ is positive definite. Then the pdf which maximizes the entropy (as defined by (I.5)) is given by $f(\boldsymbol{X}) = \mathbb{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

From our previous theorems we get the important result that

Theorem 4 : The result (III.18) is important for it shows that if $S_X(z)$ is the *power spectral density* (PSD) of a stochastic process $\boldsymbol{X}(t)$ which is stationary and has a *rational transfer function* $H(z)$ (see Definitions 8 and 9) then a model which maximizes the Entropy of $\boldsymbol{X}(t)$ is an AR process.

IV Estimation of the AR Coefficients : Levinson-Durbin Algorithm

This recursive algorithm was first suggested by Levinson (1947) and later extended by Durbin (1960). We assume a given sample $\{x_1, x_2 \ldots . x_N\}$ in mean –deviation form.

Consider an AR(L) process $(L \ll N)$

$$x_t = \sum_{j=1}^{L} \alpha_j x_{t-j} + \epsilon_t \qquad \text{(IV.1)}$$

where $\{\epsilon_t\}$ is a white noise process with mean 0 and variance σ_ϵ^2. We will use the notation P_M to denote the value of σ_ϵ^2 (prediction error) for an M-th order autoregression. By taking the expectations $E(x_{t-k}\epsilon_t)$ in (IV.1), we find that for $k > 0, E(x_{t-k}\epsilon_t) = 0$. Using this fact we get the well-known Yule-Walker equations

The Yule-Walker equations of order M are $(M \leq L)$

$$\begin{cases} \hat{\rho}_1 - \hat{\alpha}_1^M \hat{\rho}_0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . - \hat{\alpha}_M^M \hat{\rho}_{M-1} = 0 \\ \hat{\rho}_2 - \hat{\alpha}_1^M \hat{\rho}_1 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . - \hat{\alpha}_M^M \hat{\rho}_{M-2} = 0 \\ \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . \\ \hat{\rho}_M - \hat{\alpha}_1^M \hat{\rho}_{M-1} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . - \hat{\alpha}_M^M \hat{\rho}_0 = 0 \end{cases} \qquad \text{(IV.2)}$$

where the $\hat{\rho}_k, k = 1,2..$ are the autocorrelations at $lag\ k$.

Or in matrix form

$C(M)\widehat{\boldsymbol{\alpha}}^{\boldsymbol{M}} = \widehat{\boldsymbol{\rho}}^{\boldsymbol{M}}$ where

$$\widehat{\boldsymbol{\alpha}}^{\boldsymbol{M}} = \begin{bmatrix} \hat{\alpha}_1^M \\ \hat{\alpha}_2^M \\ \ldots . \\ \hat{\alpha}_M^M \end{bmatrix} \quad \text{and } \widehat{\boldsymbol{\rho}}^{\boldsymbol{M}} = \begin{bmatrix} \hat{\rho}_1 \\ \hat{\rho}_2 \\ \ldots . \\ \hat{\rho}_M \end{bmatrix}$$

$$C(M) = \begin{bmatrix} \hat{\rho}_0 & \hat{\rho}_1 & \cdots & \hat{\rho}_{M-1} \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ \hat{\rho}_{M-1} & \cdots & \cdots & \hat{\rho}_0 \end{bmatrix}$$

By Yule- Walker equations of order $(M+1)$

$$C(M+1)\begin{bmatrix} \hat{\alpha}_1^{M+1} \\ \hat{\alpha}_2^{M+1} \\ \cdots \\ \hat{\alpha}_M^{M+1} \\ \hat{\alpha}_{M+1}^{M+1} \end{bmatrix} = \begin{bmatrix} \hat{\rho}_1 \\ \hat{\rho}_2 \\ \cdots \\ \hat{\rho}_M \\ \hat{\rho}_{M+1} \end{bmatrix} \qquad \text{(IV.3)}$$

where

$$C(M+1) = \begin{bmatrix} \hat{\rho}_0 & \hat{\rho}_1 & \cdots & \cdots & \hat{\rho}_M \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \hat{\rho}_{M-1} & \cdots & \cdots & \hat{\rho}_0 & \hat{\rho}_1 \\ \hat{\rho}_M & \cdots & \cdots & \cdots & \hat{\rho}_0 \end{bmatrix}$$

Suppose in (IV.3) we take only the first M rows. Then the l.h.s. of (IV.3) becomes

$$C(M)\begin{bmatrix} \hat{\alpha}_1^{M+1} \\ \hat{\alpha}_2^{M+1} \\ \cdots \\ \hat{\alpha}_M^{M+1} \end{bmatrix} + \hat{\alpha}_{M+1}^{M+1}\begin{bmatrix} \hat{\rho}_M \\ \hat{\rho}_{M-1} \\ \cdots \\ \hat{\rho}_1 \end{bmatrix} \qquad \text{(IV.4)}$$

Hence from (IV.3) and (IV.4) we can write

$$C(M)\begin{bmatrix} \hat{\alpha}_1^{M+1} \\ \hat{\alpha}_2^{M+1} \\ \cdots \\ \hat{\alpha}_M^{M+1} \end{bmatrix} = \begin{bmatrix} \hat{\rho}_1 \\ \hat{\rho}_2 \\ \cdots \\ \cdots \\ \hat{\rho}_M \end{bmatrix} - \hat{\alpha}_{M+1}^{M+1}\begin{bmatrix} \hat{\rho}_M \\ \hat{\rho}_{M-1} \\ \cdots \\ \hat{\rho}_1 \end{bmatrix} \qquad \text{(IV.5)}$$

Or

$$\begin{bmatrix} \hat{\alpha}_1^{M+1} \\ \hat{\alpha}_2^{M+1} \\ \cdots \\ \hat{\alpha}_M^{M+1} \end{bmatrix} = C^{-1}(M)\begin{bmatrix} \hat{\rho}_1 \\ \hat{\rho}_2 \\ \cdots \\ \cdots \\ \hat{\rho}_M \end{bmatrix} - \hat{\alpha}_{M+1}^{M+1}C^{-1}(M)\begin{bmatrix} \hat{\rho}_{(M)} \\ \cdots \\ \cdots \\ \hat{\rho}_1 \end{bmatrix} \qquad \text{(IV.6)}$$

But

$$C^{-1}(M)\begin{bmatrix}\hat{\rho}_1\\ \hat{\rho}_2\\ \dots.\\ \dots.\\ \hat{\rho}_M\end{bmatrix}=\begin{bmatrix}\hat{\alpha}_1^M\\ \hat{\alpha}_2^M\\ \dots\\ \hat{\alpha}_M^M\end{bmatrix} \quad \text{and } C^{-1}(M)\begin{bmatrix}\hat{\rho}_M\\ \dots.\\ \dots.\\ \dots.\\ \hat{\rho}_1\end{bmatrix}=\begin{bmatrix}\hat{\alpha}_M^M\\ \dots\\ \dots\\ \hat{\alpha}_1^M\end{bmatrix} \tag{IV.7}$$

Combining (IV.6) and (IV.7), we get

$$\begin{bmatrix}\hat{\alpha}_1^{M+1}\\ \hat{\alpha}_2^{M+1}\\ \dots\\ \hat{\alpha}_M^{M+1}\end{bmatrix}=\begin{bmatrix}\hat{\alpha}_1^M\\ \hat{\alpha}_2^M\\ \dots\\ \hat{\alpha}_M^M\end{bmatrix}-\hat{\alpha}_{M+1}^{M+1}\begin{bmatrix}\hat{\alpha}_M^M\\ \dots\\ \dots\\ \hat{\alpha}_1^M\end{bmatrix} \tag{IV.8}$$

(IV.8) is an important step in the recursion because it expresses the first M coefficients at the $(M+1)-th$ iteration in terms of the *known* coefficients of the $M-th$ iteration and the last *unknown* coefficient of the $(M+1)-th$ iteration. Thus all the coefficients at the

$(M+1)-\ th$ iteration will be determined once we know $\hat{\alpha}_{M+1}^{M+1}$. To this task we now address ourselves.

We now augment the Yule-Walker equations by the following equation obtained by taking $E(x_{t-k}\epsilon_t)\ for\ \ k=0$, which yields

$$\hat{\rho}_0-\hat{\alpha}_1^M\hat{\rho}_1 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.-\hat{\alpha}_M^M\hat{\rho}_M=P_M \tag{IV.9}$$

(recall that P_M is called the prediction error and denotes the value of σ_ϵ^2 for an M-th order autoregression)

Adding (IV.9) to the system (IV.2) we get

$$\begin{cases}\hat{\rho}_0-\hat{\alpha}_1^M\hat{\rho}_1 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.-\hat{\alpha}_M^M\hat{\rho}_M=P_M\\ \hat{\rho}_1-\hat{\alpha}_1^M\hat{\rho}_0 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.-\hat{\alpha}_M^M\hat{\rho}_{M-1}=0\\ \hat{\rho}_2-\hat{\alpha}_1^M\hat{\rho}_1 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.-\hat{\alpha}_M^M\hat{\rho}_{M-2}=0\\ \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.\\ \hat{\rho}_M-\hat{\alpha}_1^M\hat{\rho}_{M-1} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.-\hat{\alpha}_M^M\hat{\rho}_0=0\end{cases} \tag{IV.10}$$

(IV.10) can be written in matrix form as

$$\begin{bmatrix}\hat{\rho}_0 & \hat{\rho}_1 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.\hat{\rho}_M\\ \hat{\rho}_1 & \hat{\rho}_0 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.\hat{\rho}_{M-1}\\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\ \hat{\rho}_M & \hat{\rho}_{M-1} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.\hat{\rho}_0\end{bmatrix}\begin{bmatrix}1\\ -\hat{\alpha}_1^M\\ \dots.\\ \dots\dots\\ -\hat{\alpha}_M^M\end{bmatrix}=\begin{bmatrix}P_M\\ 0\\ \dots.\\ \dots\dots\\ 0\end{bmatrix} \tag{IV.11}$$

Similarly the Yule-Walker equations of order (M+1) can be written as

$$\begin{bmatrix} \hat{\rho}_0 & \hat{\rho}_1 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \hat{\rho}_{M+1} \\ \hat{\rho}_1 & \hat{\rho}_0 \quad \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \hat{\rho}_M \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \\ \hat{\rho}_{M+1} & \hat{\rho}_M \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots \hat{\rho}_0 \end{bmatrix} \begin{bmatrix} 1 \\ -\hat{\alpha}_1^{M+1} \\ \dots\dots \\ \dots\dots \\ -\hat{\alpha}_{M+1}^{M+1} \end{bmatrix} = \begin{bmatrix} P_{M+1} \\ 0 \\ \dots\dots \\ \dots\dots \\ 0 \end{bmatrix} \quad \text{(IV.12)}$$

Let us now confine our attention to the first row of the matrix on the l.h.s. of (IV.12)

$$\hat{\rho}_0 - \hat{\alpha}_1^{M+1}\hat{\rho}_1 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots - \hat{\alpha}_{M+1}^{M+1}\hat{\rho}_{M+1} = P_{M+1} \quad \text{(IV.13)}$$

Using (IV.8) we write (IV.13) as

$$\hat{\rho}_0 - \hat{\alpha}_{M+1}^{M+1}\hat{\rho}_{M+1} - P_{M+1} =$$

$$[\hat{\rho}_1 \ \hat{\rho}_2 \dots\dots \hat{\rho}_M] \begin{bmatrix} \hat{\alpha}_1^{M+1} \\ \hat{\alpha}_2^{M+1} \\ \dots \\ \hat{\alpha}_M^{M+1} \end{bmatrix} = [\hat{\rho}_1 \ \hat{\rho}_2 \dots\dots \hat{\rho}_M] \begin{bmatrix} \hat{\alpha}_1^{M} \\ \hat{\alpha}_2^{M} \\ \dots \\ \hat{\alpha}_M^{M} \end{bmatrix} - \hat{\alpha}_{M+1}^{M+1} [\hat{\rho}_1 \ \hat{\rho}_2 \dots\dots \hat{\rho}_M] \begin{bmatrix} \hat{\alpha}_M^{M} \\ \dots \\ \dots \\ \hat{\alpha}_1^{M} \end{bmatrix}$$

$$=(\hat{\rho}_0 - P_M) - \hat{\alpha}_{M+1}^{M+1}(\hat{\rho}_1\hat{\alpha}_M^M + \hat{\rho}_2\hat{\alpha}_{M-1}^M \dots\dots + \hat{\rho}_M\hat{\alpha}_1^M) \quad \text{(IV.14)}$$

Simplifying (IV.14) we get

$$P_{M+1} = P_M - \hat{\alpha}_{M+1}^{M+1}[\hat{\rho}_{M+1} - \hat{\rho}_1\hat{\alpha}_M^M - \hat{\rho}_2\hat{\alpha}_{M-1}^M \dots\dots - \hat{\rho}_M\hat{\alpha}_1^M] \quad \text{(IV.15)}$$

Putting

$$\Delta_M = [\hat{\rho}_{M+1} - \hat{\rho}_1\hat{\alpha}_M^M - \hat{\rho}_2\hat{\alpha}_{M-1}^M \dots\dots - \hat{\rho}_M\hat{\alpha}_1^M] \quad \text{(IV.16)}$$

And we can write (IV.15) as

$$P_{M+1} = P_M - \hat{\alpha}_{M+1}^{M+1}\Delta_M \quad \text{(IV.17)}$$

$$P_{M+1} = P_M - \hat{\alpha}_{M+1}^{M+1}[\hat{\rho}_{M+1} - \hat{\rho}_1\hat{\alpha}_M^M - \hat{\rho}_2\hat{\alpha}_{M-1}^M \dots\dots - \hat{\rho}_M\hat{\alpha}_1^M]$$

Consider similarly the last row of (IV.12)

to get

$$\hat{\rho}_{M+1} = \hat{\rho}_M\hat{\alpha}_1^{M+1} + \hat{\rho}_{M-1}\hat{\alpha}_2^{M+1} - \dots\dots\dots\dots\dots\dots\dots\dots\dots + \hat{\rho}_0\hat{\alpha}_{M+1}^{M+1} \quad \text{(IV.18)}$$

Rearranging l.h.s. of (IV.18) gives

$$\hat{\rho}_{M+1} = \hat{\rho}_M(\hat{\alpha}_1^M - \hat{\alpha}_{M+1}^{M+1}\hat{\alpha}_M^M) - \hat{\rho}_{M-1}(\hat{\alpha}_2^M - \hat{\alpha}_{M+1}^{M+1}\hat{\alpha}_{M-1}^M) \dots\dots - \hat{\rho}_1(\hat{\alpha}_M^M - \hat{\alpha}_{M+1}^{M+1}\hat{\alpha}_1^M) - \hat{\rho}_0\hat{\alpha}_{M+1}^{M+1}$$

or

$$0 = [\hat{\rho}_{M+1} - \hat{\rho}_1\hat{\alpha}_M^M - \hat{\rho}_2\hat{\alpha}_{M-1}^M \dots\dots - \hat{\rho}_M\hat{\alpha}_1^M] - \hat{\alpha}_{M+1}^{M+1}[\hat{\rho}_0 - \hat{\rho}_1\hat{\alpha}_1^M \dots\dots\dots\dots\dots\dots\dots - \hat{\rho}_M\hat{\alpha}_M^M]$$

$$0 = \Delta_M - \hat{\alpha}_{M+1}^{M+1}P_M \quad \text{(IV.19)}$$

(on using (IV.9) and (IV.16)

From (IV.19) we get

$$\hat{\alpha}_{M+1}^{M+1} = \left(\frac{\Delta_M}{P_M}\right) \qquad \text{(IV.20)}$$

Note that the algorithm assumes that the autocorrelations $\widehat{\boldsymbol{\rho}}_{\mathbf{0}}, \widehat{\boldsymbol{\rho}}_{\mathbf{1}}$..etc. are all known i.e. calculated from the data $\{\boldsymbol{x_1}, \boldsymbol{x_2} \ldots . \boldsymbol{x_N}\}$ before hand. Hence $\Delta_{\boldsymbol{M}}$ is fully known as it depends on the known autocorrelations and the known autoregression coefficients of the $\boldsymbol{M-th}$ iteration. Further, by (IV.9), $\boldsymbol{P_M}$ is completely known from the $\boldsymbol{M-th}$ iteration.

(IV.8) and (IV.20) together complete our recursion because as said earlier (IV.8) expresses the first M coefficients at the $(M+1)-th$ iteration in terms of the *known* coefficients of the $M-th$ iteration and the last *unknown* coefficient of the $(M+1)-th$ iteration. But (IV.20) determines this *(M+1)-th* coefficient $\hat{\alpha}_{M+1}^{M+1}$ in terms of Δ_M and P_M which are both known coefficients at the $M-th$ iteration itself. Thus all the coefficients at the $(M+1)-th$ iteration are fully determined and the recursion is complete. To start the recursion we use the Yule-Walker equations (IV.2) for *M=1* to get

$$\hat{\alpha}_1^1 = \frac{\hat{\rho}_1}{\hat{\rho}_0} \qquad \text{(IV.21)}$$

We can also develop a recursion for the prediction errors P_M. We start this recursion with

$$P_0 = Var\,(\sigma_\epsilon^2) = \left(\frac{\sum_{t=1}^{N} x_t^2}{N}\right)$$

Further, we use the first equation of the system (IV.10) to obtain

$$P_1 = \hat{\rho}_0 - \hat{\alpha}_1^1 \hat{\rho}_1 \qquad \text{(IV.22)}$$

From (IV.20) we know that

$$\Delta_M = \hat{\alpha}_{M+1}^{M+1} P_M \qquad \text{(IV.23)}$$

Finally combining (IV.23) and (IV.17) we get the following recursion for

$$P_{M+1} = [1 - (\hat{\alpha}_{M+1}^{M+1})^2] P_M \qquad \text{(IV.24)}$$

We have now fully described how all the estimates that we need for our model can be determined recursively.

This completes the details of the Levinson-Durbin algorithm. Further discussions of this algorithm may be found in Ulrych and Bishop (1975), Smylie et al (1973), Shen e al (2011) etc.

V. Estimation of the AR Coefficients : Burg Algorithm

The Burg algorithm is also a recursive algorithm and coincides with the Levinson-Durbin algorithm right upto Step (IV.8) but then instead of using (IV.20) to determine $\hat{\alpha}_{M+1}^{M+1}$ it uses a *maximization approach* as follows.

As in the Levinson-Durbin algorithm, we assume a given sample $\{x_1, x_2 \ldots . x_N\}$ and consider an AR(L) process $(L \ll N)$

$$x_t = \sum_{j=1}^{L} \alpha_j x_{t-j} + \epsilon_t \qquad \text{(V.1)}$$

Considering the stage M of the algorithm we introduce two types of prediction errors a *feedback prediction error* $P_{b,t}^{(M)}$ and a *feedforward prediction error* $P_{f,t}^{(M)}$ at time *t* are defined as follows

$$P_{b,t}^{(M)} = = \left[x_{t+M} - \sum_{j=1}^{M} \alpha_j^{(M)} x_{t-j}\right] \qquad \text{(V.2)}$$

$$P_{f,t}^{(M)} = \left[x_t - \sum_{j=1}^{M} \alpha_j^{(M)} x_{t+j}\right] \qquad \text{(V.3)}$$

The total *feedback and feedforward prediction errors at stage M* are defined as

$$P_b^{(M)} = \sum_{t=1}^{N-M} P_{b,t}^{(M)}$$

and

$$P_f^{(M)} = \sum_{t=1}^{N-M} P_{f,t}^{(M)}$$

The coefficient $\hat{\alpha}_{M+1}^{M+1}$ at the *(M+1)-th* stage is determined by maximizing the average of the two squared errors $\Pi^{(M+1)}$ defined as

$$\Pi^{(M+1)} = \left(\frac{1}{2(N-M-1)}\right) \sum_{t=1}^{N-M-1} \left\{\left(P_{b,t}^{(M+1)}\right)^2 + \left(P_{f,t}^{(M+1)}\right)^2\right\} \qquad \text{(V.4)}$$

But

$$P_{b,t}^{(M+1)} = x_{t+M+1} - \hat{\alpha}_1^{(M+1)} x_{t+M} - \hat{\alpha}_2^{(M+1)} x_{t+M-1} - \cdots - \hat{\alpha}_M^{(M+1)} x_{t+1} - \hat{\alpha}_{M+1}^{(M+1)} x_t$$

$$= x_{t+M+1} - \left[\hat{\alpha}_1^{(M)} - k\hat{\alpha}_M^{(M)}\right] x_{t+M} - \left[\hat{\alpha}_2^{(M)} - k\hat{\alpha}_{M-1}^{(M)}\right] x_{t+M-1} - \cdots \left[\hat{\alpha}_M^{(M)} - k\hat{\alpha}_1^{(M)}\right] x_{t+1} - k x_t$$

(where to simplify the notation we have put $k = \hat{\alpha}_{M+1}^{(M+1)}$)

Thus

$$P_{b,t}^{(M+1)} = x_{t+M+1} - \hat{\alpha}_1^{(M)} x_{t+M} - \hat{\alpha}_2^{(M)} x_{t+M-1} \ldots . \hat{\alpha}_M^{(M)} x_{t+1} - k\Big[-\hat{\alpha}_M^{(M)} x_{t+M} - \hat{\alpha}_{M-1}^{(M)} x_{t+M-1} \ldots . - \hat{\alpha}_1^{(M)} x_{t+1} + x_t\Big] \qquad \text{(V.5)}$$

Put

$$A_t^{(M+1)} = x_{t+M+1} - \hat{\alpha}_1^{(M)} x_{t+M} - \hat{\alpha}_2^{(M)} x_{t+M-1} \ldots . - \hat{\alpha}_M^{(M)} x_{t+1} \quad \text{(V.6)}$$

$$B_t^{(M+1)} = -\hat{\alpha}_M^{(M)} x_{t+M} - \hat{\alpha}_{M-1}^{(M)} x_{t+M-1} \ldots . - \hat{\alpha}_1^{(M)} x_{t+1} + x_t \quad \text{(V.7)}$$

Then

$$P_{b,t}^{(M+1)} = A_t^{(M+1)} - kB_t^{(M+1)} \quad \text{(V.8)}$$

Similarly,

$$P_{f,t}^{(M+1)} = x_t - \hat{\alpha}_1^{(M+1)} x_{t+1} - \hat{\alpha}_2^{(M+1)} x_{t+2} - \cdots \hat{\alpha}_M^{(M+1)} x_{t+M} - kx_{t+M+1}$$

$$= x_t - \left[\hat{\alpha}_1^{(M)} - k\hat{\alpha}_M^{(M)}\right] x_{t+1} - \left[\hat{\alpha}_2^{(M)} - k\hat{\alpha}_{M-1}^{(M)}\right] x_{t+2} - \cdots \left[\hat{\alpha}_M^{(M)} - k\hat{\alpha}_1^{(M)}\right] x_{t+M} - kx_{t+M+1}$$

$$= x_t - \hat{\alpha}_1^{(M)} x_{t+1} - \hat{\alpha}_2^{(M)} x_{t+2} - \cdots \hat{\alpha}_M^{(M)} x_{t+M} - k\left[x_{t+M+1} - \hat{\alpha}_1^{(M)} x_{t+M} - \cdots - \hat{\alpha}_{M-1}^{(M)} x_{t+2} - \hat{\alpha}_M^{(M)} x_{t+1}\right] \quad \text{(V.9)}$$

Hence

$$P_{f,t}^{(M+1)} = B_t^{(M+1)} - kA_t^{(M+1)} \quad \text{(V.10)}$$

Note that the expressions $A_t^{(M+1)}$ and $B_t^{(M+1)}$ are independent of k

We now minimize expression (V.10) w.r.t. k

$$\frac{\partial \Pi^{(M+1)}}{\partial k} = \left(\frac{1}{2(N-M-1)}\right) \sum_{t=1}^{N-M-1} \left\{ 2\left(P_{b,t}^{(M+1)}\right) \frac{\partial P_{b,t}^{(M+1)}}{\partial k} + 2\left(P_{f,t}^{(M+1)}\right) \frac{\partial P_{f,t}^{(M+1)}}{\partial k} \right\}$$

$$= \left(\frac{-1}{(N-M-1)}\right) \sum_{t=1}^{N-M-1} \left\{ -\left(P_{b,t}^{(M+1)}\right) B_t^{(M+1)} - \left(P_{f,t}^{(M+1)}\right) A_t^{(M+1)} \right\}$$

$$= \left(\frac{1}{(N-M-1)}\right) \sum_{t=1}^{N-M-1} \left\{ \left(A_t^{(M+1)} - kB_t^{(M+1)}\right) B_t^{(M+1)} + \left(B_t^{(M+1)} - kA_t^{(M+1)}\right) A_t^{(M+1)} \right\}$$

$$= \left(\frac{1}{(N-M-1)}\right) \sum_{t=1}^{N-M-1} \left\{ \left[A_t^{(M+1)} B_t^{(M+1)} - k\left(B_t^{(M+1)}\right)^2\right] + \left[A_t^{(M+1)} B_t^{(M+1)} - k\left(A_t^{(M+1)}\right)^2\right] \right\} \quad \text{(V.11)}$$

Putting $\frac{\partial \Pi^{(M+1)}}{\partial k} = 0$ yields the value of k as

$$k = \left(\frac{2}{(N-M-1)}\right) \left[\frac{\sum_{t=1}^{N-M-1} A_t^{(M+1)} B_t^{(M+1)}}{\sum_{t=1}^{N-M-1} \left\{ \left(A_t^{(M+1)}\right)^2 + \left(B_t^{(M+1)}\right)^2 \right\}} \right] \quad \text{(V.12)}$$

Since we are at the $(M+1) - th$ recursion all the quantities upto the $M - th$ iteration are known. Hence $A_t^{(M+1)}$ and $B_t^{(M+1)}$ are fully known and thus the quantity $k = \hat{\alpha}_{M+1}^{(M+1)}$ is also fully now fully known and (exactly as in the Levinson-Durbin algorithm above), by (IV.8) all the coefficients at the $(M+1) - th$ iteration will be determined once we know $\hat{\alpha}_{M+1}^{M+1}$.

One key advantage of the Burg method over the Levinson-Durbin method is the following. The Burg estimation of $\hat{\alpha}_{M+1}^{M+1}$ does not involve $\hat{\rho}_{M+1}$, unlike the Levinson-Durbin method where $\hat{\rho}_{M+1}$ figures in the estimation of $\hat{\alpha}_{M+1}^{M+1}$ via (IV.16) and (IV.20). The autocorrelations can be estimated recursively (if needed) via (IV.18) once $\hat{\alpha}_{M+1}^{M+1}$ has been estimated. The initial value for the recursion can be taken as

$$\hat{\rho}_0 = \left(\frac{1}{N}\right)\sum_{t=1}^{N} x_t^2 \qquad \text{(V.12)}$$

This completes the discussion of the Burg algorithm. Further details may be found in Burg (1972, 1975), Andersen (1974), Herring (1977), Candy (1988, p.350-353) etc.

VI. Statistical Properties of Burg Algorithm

An interesting result noted by Ulrych and Bishop (1975, Appendix 2) is that the Burg recursive algorithm is equivalent to an *appropriate* maximum likelihood method. This of course means that the Burg estimates are *asymptotically unbiased* if certain *regularity conditions* (see Wald (1949), Haldane & Smith (1956) etc.) are satisfied.

Similarly they are *consistent* under a set of minor restrictions and are *asymptotically normal.* Their prediction error variance viz. $Var\left(P_b^{(M)}\right) = \left(\frac{2}{\upsilon}\right)V$ where V is the true variance of the process and υ (the degrees of freedom) is related to the number of data points N and the order of the autoregression M by $\upsilon = \left(\frac{N}{M}\right)$. This is comparable to the non-parametric window estimates where the degrees of freedom are given by $\upsilon = 2bN$ where N is the number of data points and b is the spectral window bandwidth (see Ulrych and Bishop (1975, p.192)).

VII.Estimation of the Autoregression Order

The crucial step in the estimation of the AR spectrum is the determination of the correct order of the model. An inaccuracy here would affect the final spectrum considerably. For example the variance of the Burg estimates increases very significantly when the order of the autoregression is overestimated. Fortunately, this problem has been virtually thrashed bare in the three decades 1970-2000, and a number of model order estimates have been suggested. A brief list would minimally include the following –Akaike's FPE (Final Prediction Error), Akaike's AIC (Information Criterion), Parzen's CAT1-3 (Criterion Autoregressive Transfer Function), Hannan-Quinn, Mallows'C_p (Conditional Prediction Criterion), Schwarz's BIC (Bayesian Information Criterion) etc. Reviews and comparisons of the various properties of these criteria have been done in Amemiya (1980), Lutkepohl (1991), Nachane (1991) etc. As this literature is very familiar now, we do not include a discussion here. However, it is important to note that all these criteria are based on evaluation of the model parameters by any *consistent* method such

as the *maximum likelihood* (which coincides with *least squares* on the assumption of Gaussianity of disturbances) *prior* to the order determination. They can also lead to misspecification errors via over-fitting or under-fitting.

An approach which is quite different from the above approaches and relatively unfamiliar to applied economists is that based on *entropy* considerations and thus fits well with the general tenor of this paper. The approach has been proposed by Chan et al (1974) and by Ishii et al (1978) and further discussed by Jategaonkar et al (1982).

We assume a given sample $\{x_1, x_2 \ldots . x_N\}$ from a stochastic process $\{X_t, t\epsilon I\ \}$ with $E(X_t) = 0$. If we desire to fit an appropriate AR model for this sample using the Burg estimates (which we know to be consistent). Just as in the case of the FPE, AIC and other criteria we decide $a\ prìorı$ on a maximum order for the model say $L \ll N$. We next define

$$\Theta_M = \begin{bmatrix} 1 & \hat{\rho}_1 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \hat{\rho}_{M-1} \\ \hat{\rho}_1 & 1 & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \hat{\rho}_{M-2} \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ \hat{\rho}_{M-1} & \hat{\rho}_{M-2} & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & 1 \end{bmatrix} \qquad \text{(VII.1)}$$

where the $\hat{\rho}_k, k = 1,2..$ is the autocorrelation at $lag\ k$ exactly as above.

We now choose the order of the model as $M\ (M \leq L)$ for which the following quantity is minimized

$$V_M = \left[log\left(\frac{N-M}{N-2M-1}\right) + log|\Theta_{M+1}| - log|\Theta_M|\right] \qquad \text{(VII.2)}$$

Note : It is shown by Ishii et al (1978) that the Akaike FPE can be written as

$$V^*_M = \left[log\left(\frac{N+M}{N-M}\right) + log|\Theta_{M+1}| - log|\Theta_M|\right] \qquad \text{(VII.3)}$$

so that the entropy based criteria are not totally distinct from the earlier criteria.

Note that the entropy based lag selection method requires computation of the determinant of the matrix Θ_M for successive values of *M*. There are now various methods available for fast computation of determinants of square matrices. The three most prominent methods are

(i) The LU decomposition in which the given matrix A is decomposed as he product of 2 matrices –an *upper triangular* matrix U and a *lower triangular* matrix L.

(ii) The QR decomposition in which the given matrix A is decomposed as the product of an *orthonormal matrix* Q and an *upper triangular matrix* R

(iii) The Cholesky decomposition applies only to *Hermitian positive definite matrices*. If the matrix H is *Hermitian positive definite* then H can be written as the product of a *lower triangular matrix* L and its *conjugate transpose* L^{*T}.

These decompositions simplify the determinant computations considerably and the algorithms based on them are discussed in Bareiss (1966), Bunch and Hopcroft (1974), Golub and Van Loan (1996), Camarero (2018), Strang (2019) etc.

Note : This paper confines itself to the theoretical aspects of the Maximum Entropy Spectral Methods. Empirical applications are planned to follow a little later.

REFERENCES

Amemiya, T. (1980) : "Selection of regressors", *International Economic Review,* vol. 21, p. 331-354

Andersen, M. (1974) : "On the calculation of filter coefficients for maximum entropy analysis", *Geophysics,* vol. 39, p. 69-72

Bareiss, E.H. (1966) : *Multi-Step Integer Preserving Gaussian Elimination,* Argonne National Laboratory Report, ANL-7213, May

Boltzmann, L. (1866) :"*Ü*ber die Mechanische Bedeutung des Zweiten Haupstatzesder W*ä*rmetheorie", *Weiner Berichte*, vol. 53, p. 195-220

Brillouin, L. (1956) : *Science and Information Theory",* Academic Press, New York, p. 159-161

Bunch, J. and J. Hopcroft (1974) : "Triangular factorization and inversion by fast matrix multiplication", *Mathematics of Computation,* vol. 28, p.231-236

Burg, J.P. (1972) : "The relationship between Maximum entropy spectra and maximum likelihood spectra", *Geophysics,* vol. 37, p. 375-376

Burg, J.P. (1975) : *Maximum Entropy Spectral Analysis*, Ph.D dissertation, Stanford University

Camarero, C. (2018) : " Simple, fast and practicable algorithms for Cholesky, LU and QR decompositions using fast rectangular matrix multiplication" arXiv (Preprint) 1812.02056

Candy, J.V. (1988) : *Signal Processing : The Modern Approach* (International Edition), McGraw Hill Book Co., Singapore

Chan, C.W., C.J.Harris and P.E.Wellstead (1974) : " An order testing criterion for a mixed autoregressive moving average process", *International Journal of Control,* vol. 20 (No.5), p.817-834

Cover, T.M. and J.A. Thomas (2006: *Elements of Information Theory*)(2nd edition), John Wiley & Sons, Hoboken, NJ,USA

Cox, D.R. and N.J.H. Small (1978) : "Testing multivariate normality", *Biometrika,* vol. 65(2),

p. 263-272

Durbin, J. (1960) : “The fitting of time series models”, *Review of International Institute of Statistics,* vol. 28, p.233-244

Gohbeg, I,C. and N.Ya. Krupnik (1969) : “On the algebra generated by Toeplitz matrices”, *Functional Analysis and Applications,* vol.3(2), p. 119-127

Golub, G.H. and C.F. Van Loan (1996) (3rd edition) : *Matrix Computations,* John Hopkins University Press, Baltimore

Haldane, J.B.S. and M.S. Smith (1956) : “The sampling distribution of a maximum likelihood estimate”, *Biometrika,* vol. 43, (1-2), p. 96-103

Herring, R.W. (1977) : *A Review of Maximum Entropy Spectral Analysis*, CRC Technical Note No. 685, Communications Research Centre, Department of Communications, Canada

Hohn, F.E. (1964) : *Elementary Matrix Algebra,* Macmillan Co. New York

Ishii, N., A. Iwata and N.Suzumura (1978) : “ Evaluation of an autoregressive process by information measure”, *International Journal of Systems Science,* Vol. 9 (No.7), p.743-751

Jategaonkar, R.V., J.R.Roal and S. Balakrishna (1982) : “Determination of model order for dynamical sytems”, *IEEE Transctions on Systems, Man and Cybernetics,* vol. SMC-12 (No. 1), p.56-62

Jaynes, E.T. (1963) : “New engineering applications of information theory”, *Proceedings of the First Symposium on Engineering Applications of Random Function Theory and* Probability (p. 163-203) *,* edited by J.L.Bogdanoff and F.Kozin, John Wiley, New York

Jaynes, E.T. (1968) : “Prior probabilities”, *IEEE Transactions on Systems Science and Cybernetics,* SSC-4, P.227-241

Johnson, J. R. (2012) : *Introduction to Digital Signal Processing,* PHI Learning Pvt. Ltd, Delhi

Levinson, H. (1947) : “The Wiener RMS (root mean square) error criterion in filter design and prediction”, *Journal of Mathematical Physics,* vol. 25, p.261-278

Lütkepohl, H. (1991) : *Introduction to Multiple Time Series,* Springer, Berlin

Nachane, D.M. (1991) : “"Practical aspects of causal inference : An Indian application" *Sankhya Ser.B* vol. 53 (Dec), p.384-402

Nachane, D.M. (2006) : *Econometrics : Theoretical Foundations and Empirical Perspectives,* Oxford University Press, Delhi, India

Papoulis, A. (1991)(3rd edition) : *Probability, Random Variables and Stochastic Processes,* McGraw-Hill Inc., Singapore

Parter, S.(1986) : “On the distribution of singular values of Toeplitz matrices”, *Linear Algebra and Applications,* vol. 80, p.115-130

Priestley, M.B. (1981) : *Spectral Analysis and Time Series,* Academic Press, London

Sayed, A.H. and T.Kailath (2001) : “A survey of spectral factorization methods” , *Numerical Linear Algebra with Applications,* vol. 8, p. 467-496

Shannon, C.E. (1948) : “A mathematical theory of communication*”, The Bell System Technical Journal,* vol. 27, (No.3), P.379-423

Shen, J., T.Tang and L.L.Wang (2011) : *Spectral Methods : Algorithms, Analysis and Applications,* Springer, Berlin

Smylie, D.E., G.K.C.Clarke and T.J.Ulrych (1973) : “Analysis of irregularities in the earth’s rotation”, *Methods in Computational Physics,* vol. 13, p. 391-430

Stoica, P. and R. Moses (2015) : *Spectral Analysis of Signals,* PHI Learning Pvt. Ltd., Delhi, India

Strang, G. (2019) : *Linear Algebra and Learning from Data,* Cambridge University Press, Cambridge, UK

G. Szego(1915) : “Ein Greenzwertsatz *ü*ber die Toeplitzschen einer reellen positive funktion” *Mathematische Annalen*, vol. 76 (4), p.490-503

Ulrych, T.J. and T.N.Bishop (1975) : “Maximum entropy spectral analysis and autoregressive decomposition”, *Review of Geophysics and Space Physics”,* vol. 13 (No.1), p. 183-200

Wald, A. (1949) : “Notes on the consistency of the maximum likelihood estimate”, *Annals of Mathematical Statistics,* vol. 20 (4), p. 595-601

Widon, H.(1989) : “On the singular values of Toeplitz matrices”, *Zeitschrift für Analysis und thre Anwendungen,* Bd. (3), p. 221-229