ON THE CONNECTION OF NON-LINEAR SYSTEM IDENTIFICATION AND MEASURES OF DEPENDENCE

By

I. Joó and P. Várlaki

Department of Transport Economics, Technical University, Budapest

Received August 30, 1979

Presented by Prof. Dr. E. Borotvás

1. Introduction

By the identification of systems we mean the construction of the model (optimal in a certain sense) of an unknown system[1] on the basis of its input and output processes observed. The "classical" theory on filtration, extrapolation and interpolation of stationary time series [1, 2] developed by Wiener and Kolmogorov independently from each other during the World War, II can be regarded the basic work of system identification. In the appendix to Wiener's fundamental work [1] Levinson introduces Wiener's epoch — making results in "identification" formulation.

The essence of the identification of linear systems is: the actual output is estimated by the linear functional of the input signal realization and the arbitrary weighting function in the estimation is determined on the basis of the principle of "minimal mean squares error" from the Wiener—Hopf integral equation. The Wiener—Hopf integral equation is of the Fredholm type [5], and in it there are auto and cross correlation coefficients. Thus the (linear) system identification is based on the application of these measures of dependence. These measures of dependence were introduced by K. Pearson and F. Galton and were generalized by Wiener and Kolmogorov for stochastic processes.

In the investigation of linear systems it is especially favourable if a so-called "white noise" generator [4] is used at the input.[2] In this case, the so-called weighting function of the system is obtained as the cross-correlation coefficient (function) of the input and output signal.

In this paper this "active identification" method will be generalized for nonlinear systems of Hammerstein type [3].

[1] The system can be the subject of telecommunication, biological economical etc. investigations [3, 4].

[2] Good methods of creating of "white noise", which cannot be developed physically, are at our disposal, see e.g. [4].

The identification of non-linear dynamic systems, that began with Wiener's fundamental MIT report [6, 7] took another direction. It concentrated less on the application of statistical measures of dependence[2] in the construction of optimal estimations. The canonical methods developed this way (the Wiener non-linear method, the Volterra model, etc.) are often problematic in practice when applying them in order to acquire the necessary basic information and their treatment demands significant expenditure. Neither is the identification of the non-linear dynamic systems by making use of the linear correlation theory expedient in general, because in a significant number of practical cases the linear model does not ensure the adequate description of the real non-linear systems. It must be noted that on the basis of the Wiener correlation theory (linear model) the solution of the identification problem is relatively simple and "can be comprehended", furthermore this theory is "aesthetic". Its significant advantage (compared with the non-linear methods used up till now) is that it works only with the characteristic of the first two moments (with expected values and correlation functions). The application of higher moments would make the solution of the identification problem very difficult.

On the information available there is a natural demand for the development of a typical statistical method that is similar to the linear (correlation) theory, and generalizes it, furthermore it is adequate for the identification of quite wide classes of non-linear systems without the use of higher moments.

In this paper there is an attempt to satisfy for the class of Hammerstein non-linear systems, with the aid of conditional expected value and correlation ratio. The given procedure of course contains the Wiener's linear theory as a special case. It will be proved that our procedure is the most effective in the class of Hammerstein systems, in case of "white" noise input signals.

[2] The fundamental measures of dependence of the random variables (not entirely in general) were first defined and applied in the course of his biometrical researches by the eminent English statisticians, K. Pearson and F. Galton in the turn of the century. These concepts are: *correlation coefficient, correlation ration* and *contingency*. (The conditional average concept that is in the definition of the correlation ratio and is called at present *conditional expectation*, was introduced by Galton in 1885 in course of his biometrical investigations.) These fundamental concepts were later supplemented with two important theoretical measures of dependence, with the *maximal correlation* (Gebelein, 1941) and with the *paired dependence-modulus of the series of random variables* (Rényi, 1959). All of these can be found in Rényi's [8] book with reference. The correlation coefficient, the correlation ratio and the mean square contingency for arbitrary random variables were defined by Kolmogorov (1933) and Rényi in 1959. The correlation coefficient for processes was interpreted by Wiener and Kolmogorov, the correlation ratio for stationary stochastic processes by Rajbman in 1963, under the name *dispersion function* [9].

2. The connection of the identification error and the measures of dependence

Let $f(t)$ and $g(t)$ be stationary time series ([1]). On the cross correlation ratio of g referring to f the quantity

$$\eta(s) = \frac{(\mathbf{M}\{\mathbf{M}^2[g(t)|f(t-s)]\} - \mathbf{M}^2[g(t)])^{1/2}}{\mathbf{D}[g(t)]} = \frac{\mathbf{D}\{\mathbf{M}[g(t)|f(t-s)]\}}{\mathbf{D}[g(t)]} \tag{1}$$

is meant, assuming that expectations, $\mathbf{M}[f(t)]$, $\mathbf{M}[g(t)]$, and variances $\mathbf{D}^2[f(t)]$, $\mathbf{D}^2[g(t)]$ are finite.

It is evident: $0 \leq \eta(s) \leq 1$ ([8]).

The main result of this paper is the following.

Theorem:

Let $f(t) = f(T^t\omega)$ be arbitrary stationary with unit intensity white noise (see [11, 14] for definition), furthermore $g(t) = g(\tilde{T}^t\omega)$ arbitrary normalized stationary time series (stochastic process) where $\{T^t\}$, $\{\tilde{T}^t\}$ are measure preserving, metrically transitive transformation groups. Then the difference of the minimal mean square errors of the Wiener's linear and of the Hammerstein non-linear models can be determined from the following:

$$\inf_{K \in L^2(0,\infty)} \lim_{T\to\infty} \frac{1}{T} \int_0^T \left| g(t) - \int_0^\infty f(t-s)\,\mathbf{K}(s)\,ds \right|^2 dt - \tag{2}$$

$$- \inf_{\tilde{K} \in L^2(0,\infty)} \inf_{G \in \mathfrak{G}_f} \lim_{T\to\infty} \frac{1}{T} \int_0^T \left| g(t) - \int_0^\infty \mathbf{G}(f(t-s))\,\tilde{\mathbf{K}}(s)\,ds \right|^2 dt = \int_0^\infty (\eta^2(t) - \rho^2(t))\,dt.$$

where $\varrho(t)$ is the cross correlation coefficient (see Appendix), furthermore $\mathbf{G}_f \overset{\text{def}}{=} \{\mathbf{G} \in L^2(-\infty,\infty) : \mathbf{G}(f) \in L^2(-\infty,\infty)$ is normalized, $\mathbf{M}[\mathbf{G}(f(t))] = 0$, $D^2[G(f(t))] = 1\}$.

Note

The theorem is true also for the practically important stationary stochastic processes in the "dispersion sense" [9], without assuming their ergodicity.

Thus, if the $x(t)$ input process is a stationary white noise with unit intensity and with an expected value of zero, and $y(t)$ (normalized) stochastic output processes are stationary dependents in dispersion sense, and also $\mathbf{G}$ is a Borel measurable normalized function, then the following expression gives the difference between the minimal mean square error of the linear and Hammerstein models:

$$\inf_{h} \mathbf{M}\Big(y(t) - \int_0^\infty h(\tau)\, x(t-\tau)\, d\tau\Big)^2 - \inf_{\hat{G}} \inf_{h^*} \mathbf{M}\Big(y(t) - \int_0^\infty h^*(\tau)\, \hat{\mathbf{G}}\,[x(t-\tau)]\, d\tau\Big)^2 = \\ = \int_0^\infty \big(\eta_{yx}^2(s) - \varrho_{yx}^2(s)\big)\, ds \tag{3}$$

where $\eta_{yx}(t)$ and $\rho_{yx}(t)$ are the cross-correlation ratio and coefficient between $y(t)$ and $x(t)$.

Proof

The infimum referring to $\widetilde{\mathbf{K}}$ function, where $\mathbf{G}$ is fixed is posited for the $\widetilde{\mathbf{K}}^*$ that satisfies the following Fredholm integral equation (obtained through the use of the Birkhoff's ergodic theorem):

$$\mathbf{M}\,[g(t)\, \mathbf{G}(f(t-u))] = \int_0^\infty \widetilde{\mathbf{K}}^*(s)\, [\mathbf{M}\{\mathbf{G}\,[f(t-u)]\ \mathbf{G}[f(t-s)]\}\, ds,\ u \geq 0 \tag{4}$$

On the other hand, on the basis of the very definition of white noise, where $t \neq s$, $f(t)$ and $f(s)$ are independent, according to the well known theorem of probability theory it follows that $\mathbf{G}(f(t))$ and $\mathbf{G}(f(s))$ are independent and we get by a complicated calculation (First of all it is necessary to prove this for discrete[1] stationary time series, then to realize appropriate limiting transitions on the basis of Riesz construction of the Lebesque integral [5]. For the sake of simplicity we will not deal with the proof; an exact proof will be given in the next paper) from (4)

$$\widetilde{\mathbf{K}}^*(s) = \mathbf{M}\,(g(t)\, \mathbf{G}\,[f(t-s)]). \tag{5}$$

On the basis of Wiener [1]

$$\inf_{\widetilde{K} \in L^2(0,\infty)} \lim_{T\to\infty} \frac{1}{T} \int_0^T \Big| g(t) - \int_0^\infty \mathbf{G}(f(t-s))\, \widetilde{\mathbf{K}}(s)\, ds \Big|^2 dt = \\ = 1 - \int_0^\infty \widetilde{\mathbf{K}}^*(s)\, \mathbf{M}\,[g(t)\, \mathbf{G}(f(t-s))]\, ds \tag{6}$$

where $\widetilde{\mathbf{K}}^*(s)$ means the unambiguous solution of the integral equation (4).

From (5) and (6) is derived

$$\inf_{\widetilde{K} \in L^2(0,\infty)} \inf_{G \in \mathcal{G}_f} \lim_{T\to\infty} \frac{1}{T} \int_0^T \Big| g(t) - \int_0^\infty \mathbf{G}\,[f(t-s)]\, \widetilde{\mathbf{K}}(s)\, ds \Big|^2 dt = \\ = \inf_{G \in \mathcal{G}_f} \Big(1 - \int_0^\infty \mathbf{M}^2\,[g(t)\, \mathbf{G}(f(t-s))]\, ds\Big). \tag{7}$$

[1] Theorem is new also for the discrete case the presented proof is correct and complete for it.

and

$$\inf_{G \in \mathcal{G}_f} \left(1 - \int_0^\infty \mathbf{M}^2 [g(t)\, \mathbf{G}(f(t-s))]\, ds\right) = 1 - \sup_{G \in \mathcal{G}_f} \int_0^\infty \mathbf{M}^2 [g(t)\, \mathbf{G}(f(t-s))]\, ds =$$

$$= 1 - \int_0^\infty \sup_{G \in \mathcal{G}_f} \mathbf{M}^2 [g(t)\, \mathbf{G}(f(t-s))]\, ds. \tag{8}$$

It can be proved that in this case the supremum can be put behind the integral. The basis of the proof (on the basis of generalized Beppo—Levy's theorem) is that

$$\sup_{G \in \mathcal{G}_f} \mathbf{M}^2 [g(t)\, \mathbf{G}(f(t-s))]$$

$\mathbf{G}_0 \in \mathcal{G}_f$ function can be reached with only one function on a big set. A theorem of Rényi gives the connection of the correlation coefficient and the correlation ratio ([8]. p. 228):

$$\sup_{G \in \mathcal{G}_f} \mathbf{M}^2 [g(t)\, \mathbf{G}[f(t-s)]] = \mathbf{M}\{\mathbf{M}^2 [g(t) \mid f(t-s)]\} = \eta^2(s). \tag{9}$$

From (7), (8) and (9) we get

$$\inf_{K \in L^2(0,\infty)} \inf_{G \in \mathcal{G}_f} \lim_{T \to \infty} \frac{1}{T} \int_0^T \left| g(t) - \int_0^\infty \mathbf{G}[f(t-s)]\, \widetilde{\mathbf{K}}(s)\, ds \right|^2 dt = 1 - \int_0^\infty \eta^2(s)\, ds. \tag{10}$$

Referring to linear estimation

$$\rho(u) = \int_0^\infty \mathbf{K}(s)\, \varphi(u-s)\, ds, \quad u \geq 0.$$

where $\varphi(\tau)$ is the autocorrelation coefficient.

From the Wiener—Hopf integral equation we obtain for $\mathbf{K}^*$

$$\mathbf{K}^*(s) = \rho(s).$$

using this we have

$$\inf_{K \in L^2(0,\infty)} \lim_{T \to \infty} \frac{1}{T} \int_0^T \left| g(t) - \int_0^\infty \mathbf{K}(s) f(t-s)\, ds \right|^2 dt =$$

$$= 1 - \int_0^\infty \mathbf{K}^*(s)\, \rho(s)\, ds = 1 - \int_0^\infty \rho^2(s)\, ds \tag{11}$$

From (10) and (11) follows (2) and our theorem is proved.

Note

From (10) for the case investigated here follows the inequality:

$$\int_0^\infty \eta^2(t)\,dt \leq 1$$

that is the special case of the Bessel inequality and generalizes one consequence of Rényi's "large sieve" theorem ([8], p. 233).

3. Applications

On the basis of Rényi's theorem ([8], p. 288) the equality in (9) is reached in the case of these $\mathbf{G}$ for which

$$\mathbf{G}[f(t-s)] = a(s)\,\mathbf{M}[g(t) \mid f(t-s)] + b(s),$$

where $a(s) \neq 0$, $b(s)$ are constant in case of fixed s.

In case of non normalized $\mathbf{G}$ the normalized of $\mathbf{G}$ is

$$|\widehat{\mathbf{G}}[f(t-s)]| = \left| \frac{\mathbf{M}[g(t) \mid f(t-s)] - \mathbf{M}[g(t)]}{\sqrt{\Theta(s)}} \right| = |\widehat{\mathbf{M}}[g(t) \mid f(t-s)]|, \quad (11)$$

where $\Theta(s) = \eta^2(s) \cdot D^2[g(t)]$ is the Rajbman's dispersion function ([9]).

It can be seen that in a Hammerstein system of

$$y(t) = \int_0^\infty \widetilde{\mathbf{K}}(s)\,\mathbf{G}[f(t-s)]\,ds + \xi(t)$$

equation (here $\xi(t)$ is the noise process independent of $f(t)$), in the case of "white" noise input signals, from the viewpoint of the minimal mean square error the optimal estimation of the $\mathbf{K} \in L^2(0, \infty)$ weighting function is given by the cross correlation ratio, that of the $\mathbf{G} \in \mathcal{G}_f$ function is given by the $\mathbf{M}[g(t) \mid f(t-s)]$ conditional expected value. Thus for the optimal estimation

$$\hat{g}(t) = \int_0^\infty \widehat{\mathbf{G}}[f(t-s)]\,\widetilde{\mathbf{K}}(s)\,ds \rightarrow \hat{\hat{g}}(t) = \int_0^\infty \widehat{\mathbf{M}}[g(t) \mid f(t-s)]\,\eta(s)\,ds. \quad (12)$$

This result can be used for the identification of Hammerstein non-linear systems, it is analogous to the theory of linear systems. According to it if a "white" noise (test) signal is applied at the input of the system, then as the product of the cross correlation ratio and the normalized conditional expected value the optimal estimation of the product of the non-linear function and the weighting function is obtained on the basis of the mean square error (see [10]).

Note

If the input signal is a white noise process, in the case of a linear stationary system, when the behavior of the system is described by the

$$y(t) = \int_0^\infty \mathbb{K}(s) f(t-s)\, ds + \xi(t)$$

(where $\xi(t)$ is the noise process uncorrelated of $f(t)$) from the viewpoint of the minimum mean square error, the optimal estimation of the $\mathbb{K} \in L^2(0, \infty)$ weighting function is given by the cross correlation coefficient and for the optimal estimation we obtain that

$$\hat{g}(t) = \int_0^\infty \rho(s) \hat{f}(t-s)\, ds .$$

It is easy to understand that applying the estimation for linear systems, if $f(t)$ is unit white noise process then because of

$$\eta(t) = |\rho(t)| \tag{13}$$

$$\widehat{\mathbb{M}}[g(t) \mid f(t-s)] = \operatorname{sign} \rho(s) \hat{f}(t-s), \quad \text{where } \operatorname{sign} \xi = \begin{cases} 1 & \text{for } \xi > 0 \\ 0 & \text{for } \xi = 0 \\ -1 & \text{for } \xi < 0 \end{cases}$$

we obtain

$$\hat{y}(t) = \int_0^\infty \eta(s) \widehat{\mathbb{M}}[g(t) \mid f(t-s)]\, ds = \int_0^\infty |\rho(s)| \operatorname{sign} \rho(s) \hat{f}(t-s)\, ds = \int_0^\infty \rho(s) \hat{f}(t-s)\, ds,$$

that is the (12) model at the testing of the linear dynamic system "automatically" ensures the optimal linear estimation with the correlation coefficient from the viewpoint of the mean square error.

Finally the relation obtained as a result of Theorem (2) provides an opportunity to consider the

$$N = \int_0^\infty (\eta^2(s) - \rho^2(s))\, ds \tag{14}$$

as degree of the non-linearity of the Hammerstein system [10].

From (10) it immediately follows

$$0 \leq N \leq 1 \tag{15}$$

$N = 0$ if the system is linear and $N = 1$ where there is a non-linear deterministic system where the linear model cannot be applied (e.g. the G function is of x^k form when k is an even number) and the test signal $f(t)$ is Gaussian white noise with unit intensity.

This measure of the non-linearity is a generalization of the "classical" K. Pearson degree of non-linearity for two random variables. In fact if "stochastic processes" $g(t)$ and $f(t)$ are independent of t, e.g. they are random variables, then obviously according to (14)

$$N = \eta^2 - \rho^2$$

where obviously

$$\inf_{a,b} \mathbf{M}(g - (a + bf))^2 - \inf_{G} \mathbf{M}(g - G(f))^2 = \eta^2 - \rho^2$$

with η correlation ratio and ρ correlation coefficient between the random variables g and f.

The relative degree of non-linearity can be interesting, it can be defined as follows

$$N_R = \frac{\int_0^\infty (\eta^2(s) - \rho^2(s))\, ds}{\int_0^\infty \eta^2(s)\, ds}. \tag{16}$$

It is easy to understand that the following inequality is extant:

$$0 \leq N_R \leq 1 \tag{17}$$

Example (1)

Let us consider the definition of non-linearity degree in the case of a Hammerstein system with quadratic non-linearity where the equation giving the relation between the input $x(t)$ and the output $y(t)$ stochastic processes has the form:

$$y(t) = \int_0^\infty h(s)\, G_2[x(t-s)]\, ds$$

where $G_2(x) = ax^2 + bx + c$, here a, b and c can be any constant (Fig. 1).

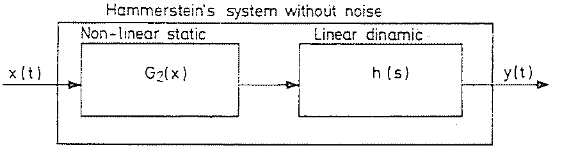

Fig. 1

Let the input (testing) signal be a Gaussian white noise with an expected value $\mathbf{M}[x(t)] = 0$ and with unit intensity. Then the conditional expected value of $\mathring{y}(t)\ (= y(t) - \mathbf{M}[y(t)])$ referring to $x(t)$ is

$$\mathbf{M}[\mathring{y}(t) \mid x(t-\tau)] = ah(\tau)\,(x^2(t-\tau) - 1) + bh(\tau)\, x(t-\tau)$$

whence the cross-correlation ratio

$$\eta^2_{yx}(\tau) = \frac{\mathbf{M}\{\mathbf{M}^2[\mathring{y}(t)\, x(t-\tau)]\}}{\mathbf{D}^2[y(t)]} = \frac{2a^2 h^2(\tau) + b^2 h^2(\tau)}{\mathbf{D}^2[y(t)]}$$

where the variance of $y(t)$ output:

$$\mathbf{D}^2[y(t)] = 2a^2 \int_0^\infty h^2(s)ds + b^2 \int_0^\infty h^2(s)\, ds$$

Obviously, the cross-correlation coefficient is:

$$\rho_{yx}(\tau) = \frac{bh(\tau)}{\mathbf{D}[y(t)]} .$$

The non-linearity degree according to (2), (14) is:

$$N_{yx} = \int_0^\infty \left(\eta^2_{yx}(t) - \rho^2_{yx}(t)\right) dt = \frac{\int_0^\infty 2a^2 h^2(s)\, ds}{2a^2 \int_0^\infty h^2(t)\, dt + b^2 \int_0^\infty h^2(t)\, dt} = \frac{2a^2}{2a^2 + b^2} .$$

Thus the non-linearity degree of the given dynamic system corresponds to the "non-linearity measure" of the G_2 function.

The above statement stands also for a more general case, namely when the above mentioned white noise and the $G(x)$ function can be given in the following form:

$$\hat{G}(x) = \sum_{k=1}^{\infty} a_k H_k(x)$$

where $H_k(x)$ is the Hermite polinom.

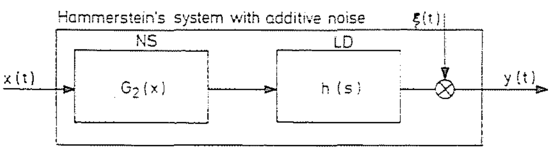

Fig. 2

If we consider the system in our example with $\xi(t)$ additive noise (Fig. 2) which is not correlated with the input process and $M[\xi(t)] = 0. D^2[\xi(t)] = \sigma^2$ i.e.

$$y(t) = \int_0^\infty h(s)\, G_2[x(t-s)]\, ds + \xi(t)$$

then the non-linearity degree of this system according to (14)

$$N_{yx} = \frac{2a^2 G}{2a^2 G + b^2 G + \sigma^2}$$

where $G = \int\limits_0^\infty h^2(t)\,dt$, furthermore the relative non-linearity degree according to (16)

$$N_R = \frac{\int\limits_0^\infty \left(\eta_{yx}^2(t) - \rho_{yx}^2(t)\right) dt}{\int\limits_0^\infty \eta_{yx}^2(s)\,ds} = \frac{2a^2}{2a^2 + b^2}.$$

It is obvious from this example that if

$$G\,[x(t-\tau)] = ax^2(t-\tau)$$

then

$$\rho_{yx}(\tau) = 0 \text{ and } N_{yx} = 1$$

In case of an additive noise which is not correlated with the input

$$N_R = 1.$$

Example (2)

Let us define the non-linearity degree of a Hammerstein system when the non-linear function is

$$G\,[x(t-\tau)] = ax^3(t-\tau).\ *$$

The input (test) process is the same as given in the previous example.

Then from [9] we can compute the cross-correlation ratio:

$$\eta_{yx}^2(u) = \frac{9\,h^2(u) + 6\,h^2(u)}{9\int\limits_0^\infty h^2(s)\,ds + 6\int\limits_0^\infty h^2(s)\,ds} = \frac{h(u)}{\int\limits_0^\infty h^2(s)\,ds}$$

and the cross-correlation coefficient:

$$\rho_{yx}(\tau) = \frac{3h(\tau)}{\sqrt{15\int\limits_0^\infty h^2(s)\,ds}}.$$

Thus the non-linearity degree of the dynamic system:

$$N_{yx} = \int\limits_0^\infty \left(\eta_{yx}^2(t) - \rho_{yx}^2(t)\right) dt = \int\limits_0^\infty \left(\frac{h^2(t)}{\int\limits_0^\infty h^2(s)\,ds} - \frac{9\,h^2(t)}{15\int\limits_0^\infty h^2(s)\,ds}\right) dt = \frac{2}{5}.$$

* For the sake of simplicity we consider the case where $a = 1$. It is easy to show that with any $a \neq 0$ we have the same result.

Let us consider the above case with any stationary additive noise $\xi(t)$ (not correlated with the input) where $\mathbf{M}[\xi(t)] = 0$ and $\mathbf{D}^2[\xi(t)] = \sigma^2$.

Hence the non-linearity degree is

$$N_{yx} = \int_0^\infty \left\{ \frac{15\,h^2(t)}{15\int_0^\infty h^2(s)\,ds + \sigma^2} - \frac{9\,h^2(t)}{} \right\} dt = \frac{6\int_0^\infty h^2(t)\,dt}{15\int_0^\infty h^2(s)\,ds + \sigma^2}$$

and the relative non-linearity degree according to (15)

$$N_R = \frac{6\int_0^\infty h^2(t)\,dt}{15\int_0^\infty h^2(s)\,ds} = \frac{2}{5}\,.$$

Appendix

In the following we summarize the most important and known measures of dependence and their most significant characteristics with regards to random variables as well as to stochastic processes. In this short summary we also mention some new results concerning measures of dependence applied to stochastic processes.

1. *Correlation coefficient*

The simplest and best known measure of dependence showing the relation between two random variables is the correlation coefficient

$$R(\xi, \eta) = \frac{\mathbf{M}[(\xi - \mathbf{M}(\xi))\,(\eta - \mathbf{M}(\eta))]}{D(\xi)\,D(\eta)} = \frac{\mathbf{M}(\xi\eta) - \mathbf{M}(\xi)\,\mathbf{M}(\eta)}{D(\xi)\,D(\eta)} \tag{I}$$

Its most important characteristics:

1. $0 \leq |R(\xi, \eta)| \leq 1$
2. $|R| = 1$ if and only if $\eta = a\xi + b$ (where a and b are constants).
3. If ξ and η are independent, then $R = 0$.
4. $R = 0$ (i.e. ξ and η are uncorrelated) does not mean that ξ and ξ are independent.

The 2nd and 4th characteristics are great disadvantages of this measure of dependence.

In the case of $\{\xi_n\}_1^\infty$ (stationary time series) the definition of the so-called autocorrelation coefficient (Wiener and Kolmogorov):

$$\varphi_j = R_j(\{\xi_n\}) = \lim_{N\to\infty} \frac{1}{2N-1} \sum_{k=-N}^{N} \xi_{k+j}\,\xi_k . \tag{II}$$

And the cross-correlation coefficient of the two processes is:

$$\psi_j = R_j(\{\xi_n\}.\{\eta_n\}) = \lim_{N\to\infty} \frac{1}{2N-1} \sum_{k=-N}^{N} \eta_{k+j}\,\xi_k . \tag{III}$$

The above definitions for stationary, stochastic processes in the continuous case are the following:

$$\varphi(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(t+\tau)\,\bar{f}(t)\,dt = R(f, f; \tau) \tag{II'}$$

$$\psi(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} f(t+\tau)\,\bar{g}(t)\,dt = R(f, g; \tau) \tag{III'}$$

Definition: $\{\xi_n\}$ is *stationary* if the T measure-preserving transformation exists: $\xi_n = \xi(T^n\omega)$.

Then according to the ergodic theorem of Birkhoff (II), (III), (II') and (III') are finite concerning almost every ω.

Definition: $\{\xi_t\}$ is stationary, if there is a measure preserving group $\{T^\lambda\}_{\lambda>0}$ of T transformation, i.e. $T^\lambda(T^\mu\omega) = T^{(\lambda+\mu)}(\omega)$ and also $\xi : \xi_t = \xi(T^t\omega)$ exists.

Definition: T and $\{T^\lambda\}$ is metrically transitive if their nontrivial invariant set does not exist.

Definition: $\{\xi_n\}$ and $\{\xi_t\}$ is ergodic if T and $\{T^\lambda\}$ is metrically transitive.

Then according to the theorem of Birkhoff:

$$R(f, f; \tau) = \varphi(\tau) = \int_\Omega f(T^\tau\omega)\,\bar{f}(\omega)\,d\omega$$

and

$$R(f, g : \tau) = \psi(\tau) = \int_\Omega f(T^\tau\omega)\,\bar{g}(\omega)\,d\omega$$

gives (1).

In the $\{\xi_t\}_{t>0}$, $\{\eta_t\}_{t>0}$ continuous case we will use (after Wiener) the symbols $f(t) = f(t, \omega)$ and $g(t) = g(t, \omega)$ in place of $\xi_t(\omega)$ and $\eta_t(\omega)$.

2. *Correlation ratio*

The correlation ratio is one of the most general measures of dependence showing how strong the (non-linear) relation is between two random variables:

$$K_\xi(\eta) = \frac{\sqrt{\mathbf{M}(\mathbf{M}^2(\eta/\xi) - \mathbf{M}(\eta))^2}}{\mathbf{D}(\eta)} = \frac{\mathbf{D}(\mathbf{M}(\eta/\xi))}{\mathbf{D}(\eta)} = \frac{\sqrt{\mathbf{D}^2(\eta) - \mathbf{M}[\mathbf{D}^2(\eta/\xi)]}}{\mathbf{D}(\eta)} \quad \text{(IV)}$$

since e.g. according to $\mathbf{D}^2(\eta) = \mathbf{D}^2(\mathbf{M}(\eta/\xi)) + \mathbf{M}(\mathbf{D}^2(\eta/\xi))$ which shows that $\mathbf{D}^2(\mathbf{M}(\eta/\xi))$ is a part of $\mathbf{D}^2(\eta)$ relative to ξ, since $\mathbf{M}(\mathbf{D}^2(\eta/\xi))$ depends usually on η only (it is a homoscedastic relation).

(If conditional variance $\mathbf{D}(\eta/\xi)$ depends on ξ, then we call it heteroscedastic relation according to Bernstein.)

The most important features of the correlation ratio:

1. $0 \le K_\xi(\eta) \le 1$.
2. $K_\xi(\eta) = 1$ if and only if $\eta = G(\xi)$ where G is Borel-measurable (which is advantageous for the correlation coefficient).
3. If ξ and η are independent then $K = 0$ but from $K = 0$ independence does not follow. However, it is true that $K = 0 \rightarrow R = 0$.
4. If $\eta = a\xi + b + \nu$ where ν is a random variable with Gaussian distribution, then $K = |R(\eta, \xi)|$.

Note : If ξ and η are random variables, $\mathbf{D}^2(\eta) < \infty$ and G is Borel-measurable, then the value of "mean square error of estimate"

$$\mathbf{M}([\eta - G(\xi)]^2) \quad \text{(V)}$$

will be minimal in the case of $G(\xi) = \mathbf{M}(\eta/\xi)$ and this error is

$$\mathbf{D}^2(\eta) - \mathbf{D}^2(\mathbf{M}(\eta/\xi)) \quad \text{(VI)}$$

It is easy to show that

$$R^2 \le K^2$$

is true in any circumstances.

The most important theorem in connection with the correlation ratio is the Rényi—Gebelein theorem, according to which

$$K_\xi^2(\eta) = \sup_{G \in G_\xi} R^2(G(\xi), \eta). \quad \text{(VII)}$$

$G_\xi =: \{$Borel-measurable function, to which $\mathbf{M}(G(\xi))$ and $\mathbf{D}(G(\xi)) < \infty\}$.

The theorem of Rényi makes it possible to apply K to stochastic processes since R is already defined for processes.

There is equation in (7) if and only if $G(\xi) = a\mathbf{M}(\eta/\xi) + b$ where a and b are constants (for normalized processes $\mathbf{D} = 1, \mathbf{M} = 0$, these are defined by their absolute value).

$$a^2 = \frac{1}{\mathbf{D}^2(\mathbf{M}(\eta/\xi))}, \quad b^2 = \frac{\mathbf{M}^2(\eta)}{\mathbf{D}^2(\mathbf{M}(\eta/\xi))}.$$

Definition of correlation ratio for (stationary) stochastic processes (Rajbman, 1963 [9])

$$K_{f(t)}(g(t): \tau) =: \int_{\Omega} \mathbf{M}[g(\omega)/f(T^\tau\omega)] \cdot g(\omega)\, d\omega$$

supposing that

$$\mathbf{D}\{\mathbf{M}[g(\omega)/f(T^\tau\omega)]\} = 1, \ \mathbf{M}(g) = 0, \ \mathbf{D}(g) = 1$$

normalization is fulfilled.

In the non-normalized case

$$K^2_{f(t)}(g(t), \tau) = \frac{\mathbf{M}[\mathbf{M}^2(g(\omega))/f(T^\tau(\omega))]^2 - \mathbf{M}^2(g)}{\mathbf{D}^2(g)}.$$

Theorem (Generalization of the Rényi—Gebelein theorem for processes):

$$\eta^2(\tau) = K^2_{f(t)}(g(t), \tau) = \sup_{h \in H_f} R^2(h(f), g, \tau) \tag{VIII}$$

$$H_f =: \left\{ h \text{ Borel-measurable functions, for which } \begin{matrix} \mathbf{M}(h(f(\omega))) < \infty \\ \mathbf{D}(h(f(\omega))) < \infty \end{matrix} \right\}.$$

Proof: applying Rényi's theorem for $\xi = f(T^\tau\omega)$ and $\eta = g(\omega)$ the expression (VIII) follows directly.

$\eta^2(\tau)$ can be computed on basis of the theorem of Birkhoff and Rényi, in the following way:

$$\eta^2(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \mathbf{M}(g(t)/f(t-\tau))\, g(t)\, dt$$

in normalized case.

Here it is also desirable in practice to compute $\mathbf{M}(g(\cdot)/f(\cdot))$ from a realization!

This can be done in the following way:

$$\frac{\int_{-T}^{T} g(s)\, I_\Delta(f(s-\tau))ds}{\int_{-T}^{T} I_\Delta(f(s-\tau))ds} \to \mathbf{M}(g(t)/f(t-\tau)) \text{ if } \begin{cases} T \to \infty \\ \Delta \to f(t-\tau) \end{cases}$$

where I is the indicator function

$$I_\Delta(f(s-\tau)) = \begin{cases} 1 & \text{if } f(s-\tau) \in \Delta \\ 0 & \text{if } f(s-\tau) \notin \Delta \end{cases}.$$

3. *Maximal correlation*

We call maximal correlation the following measure of dependence:

$$\psi(\xi, \eta) =: \sup_{u,v} |R(u(\xi); v(\eta))| \tag{IX}$$

where u and v are Borel-measurable functions.

Characteristics:

1. $0 \leq \psi(\xi, \eta) \leq 1$.
2. $\psi(\xi, \eta) = \psi(\eta, \xi)$.
3. η and ξ are independent if and only if $\psi = 0$ (this is one of the advantages of maximal correlation on the other measures of dependence discussed so far).
4. If $\eta = a\xi + b$ then $\psi(\xi, \eta) = |R(\xi, \eta)|$.
5. $\psi = 1$ if and only if $u(\xi) = v(\eta)$ where u and v are Borel-measurable.

Its generalization for stochastic processes is evident, (see Kolmogorov, Rozanov). Its calculation from realisation for example is a bit difficult, that is why it is not used in practice, although theoretically it is a significant measure of dependence (theoretically it can be regarded the best measure of dependence — see Rényi [13]).

4. *Modulus of dependence according to Rényi*

The modulus of dependence by pair of a $\{\xi_n\}$ random variable series is defined in the following way:

$$A =: \inf \{c : |\sum_n \sum_m \psi(\xi_n, \xi_m)\, x_n x_m| \leq c \sum_n x_n^2\}.$$

This can be generalized for $A = A(\{\xi_t\})$ stochastic processes

$$A(\xi_t) =: \inf \{c : |\int_0^\infty \int_0^\infty \psi(\xi_t, \xi_s)\, K(t)\, K(s)\, dt\, ds| \leq c \int_0^\infty K^2(s)\, ds\}$$

where $\psi(\xi_t, \xi_s)$ is the coefficient of maximal correlation.

Characteristics:

1. $1 \leq A < \infty$.
2. If it is independent by pair, then $A(\{\xi_n\}) = 1$.

Definition: $\{\xi_t\}$ is called "white noise" for any t and s, where $t \neq s$, then ξ_t and ξ_s are independent, the autocorrelation function of the process is $R(\xi_t, \xi_s) = N\delta(t - s)$ where δ is a Dirac-function and N is the intensity of the "white noise". It is evident that its density spectrum (i.e. the Fourier transformation of the autocorrelation function) is constant i.e. the white noise cannot be realized physically.

3. It can be shown that if ξ_t is white noise, then $A(\xi_t) = 1$.

Theorem

It can be proved that if $f(t)$ and $g(t)$ are stationary stochastic processes then

$$\int_0^\infty K^2_{f(t)}(g(t); \tau)\, d\tau \leq A(\{f(t)\}) \tag{X}$$

Note: if $\{f(t)\}$ is white noise, then

$$A(\{f(t)\}) = 1 \text{ and } \{\mathbf{M}(g(t)/f(t-\tau))\}, \quad \tau > 0$$

is an orthogonal normalized system for any t in $L^2(\Omega)$ as we previously mentioned, this not being the case, then

$$\frac{\mathbf{M}(g(t)/f(t-\tau)) - \mathbf{M}(\mathbf{M}(g(t)/f(t-\tau)))}{\mathbf{D}(\mathbf{M}(g(t)/f(t-\tau)))}$$

will already be normalized.

The Fourier series of $g(t)$ according to

$$g(t) = \int_0^\infty C(\tau)\, \mathbf{M}(g(t)/f(t-\tau))\, d\tau, \quad \text{in} \quad L^2(\Omega)$$

and here for the correlation ratio $C(\tau) = K_{f(t)}(g(t); \tau)$ it is clear, that it is independent of t, since if $g(t)$ is normalized and for $M(./\cdot)$

$$C(\tau) = \mathbf{M}(g(t)\, \mathbf{M}(g(t)/f(t-\tau))) = K_{f(t)}(g(t); \tau)$$

Hence (X) gives the Bessel-unequality.

When the relation between $g(t)$ and $f(t)$ is written in the following form

$$g(t) = \int_0^\infty K(s)\, H[f(t-s)]\, ds$$

(with the Hammerstein operator) then the above mentioned orthonormal systems are complete and in (X) we have equation.

Acknowledgement

We are indebted to the late Prof. Rajbman N. S. (1921—1981) for the stimulating discussions and valuable suggestions during the writing of this paper.

Summary

The paper deals with an important connection between measures of dependence (correlation coefficient and ratio, maximal correlation etc. for stationary stochastic processes) and identification error (by mean square error) of linear and nonlinear Hammerstein models.

On the basis of obtained results it can be proved, that the identification of Hammerstein models with the help of these statistical characteristics (i.e. regression function, correlation ratio between input and output processes) is optimally defined by the mean square error criterion in the case of a white noise input process.

References

1. Wiener, N.: Extrapolation, Interpolation and Smoothing of Stationary Time Series (With Engineering Applications). John Wiley, New York, 1950.
2. Kolmogorov, A. N.: Interpolation und Extrapolation von stationären zufälligen Folgen, Bulletin de l'academie des sciences de USSR, Ser. Math. 5, 1941.
3. Rajbman, N. S. (ed.): Identification and System Parameter Estimation, Proc. of the 4th IFAC Symposium, North Holland, Amsterdam, New York—Oxford, 1978.
4. Vámos, T.: (The control of large industrial processes) Nagy ipari folyamatok irányítása. Akadémiai Kiadó, Budapest, 1970.
5. Riesz, F.—Szőkefalvi-Nagy, B.: Leçons d'Analyse functionelle. Akadémiai Kiadó, Budapest, 1955.
6. Wiener, N.: Response of a nonlinear device to noise. Report. No. 129, Radiation Laboratory, M. I. T., 1942.
7. Wiener, N.: Nonlinear Problems in Random Theory. John Wiley, New York, 1958.
8. Rényi, A.: Valószínűségszámítás. (Theory of probability) Tankönyvkiadó, Budapest, 1966. (in Hungarian)
9. Rajbman, N. S.: Тезисы доклада на 7 совещании о теории вероятностей и математической статистике об автодисперсионной и взаимной дисперсионной функциях. Киевский Математический Институт, 1963.
10. Várlaki P.—Rajbman N. S.: Оценка степени нелинейности динамического объекта. Изв. АН СССР сер. Техническая кибернетика 1980 № 1.
11. Balakrisnan, A. V.: Applied functional analysis. Springer Verlag, New York—Heidelberg—Berlin, 1967.
12. Rényi, A.: A new version of the probabilistic generalization of the large sieve. Acta Math. Acad. Sci. Hung., V. 10, No. 1—2, 1959.
13. Rényi, A.: On measures of dependence. Acta Math. Acad. Sci. Hung., V 10, No. 4, 1959.
14. Åström K. J.: Introduction to stochastic control theory Academic Press, New York and London 1970.

Dr. István Joó
Dr. Péter Várlaki
} H-1521 Budapest