Chapter 3
Noise Immunity and Convergence

According to the principle of self-organization, the depth of the minimum of the principal selection criterion (i.e., regularity, minimum bias, balance of variables) is taken as an indicator of the successful synthesis of a model. Suppose we have m input variables of x and an output variable y with N observations. In the combinatorial inductive setup, we make all possible partial structures from the reference function $y = f(x)$. The choice of the optimal model depends on the given external criterion and on the given partition of data sets. An unbiased equation can be obtained with the help of the minimum bias criterion η_{bs} as the principal selection criterion. The same result can also be obtained, for low noisy data, using the regularity criterion $\Delta(B)$. The deeper the minimum of the unbiasedness ($0 \leq \eta_{bs} \leq 0.05$) or regularity, the more reliable the prediction of the changing character of the process. Nevertheless, biased equations can be useful for approximating a process in the interpolation interval. If the global minimum is not achieved according to our expectations, it signifies that the problem is not solved. Then it is necessary to take measures like (i) reformulating the problem, (ii) changing the list of feasible variables, (iii) introducing new reference functions, (iv) increasing the freedom of choice for further evaluation, and so on.

Noisy data is characterized by its noise level α as a measure of noise-to-signal ratio. Noise intensity in the data plays an important role in obtaining the deep minimum. If a sufficiently deep minimum of the principal selection criterion is reached, it is possible to assume that the problem is solved. The results of potential noise stability indicate the exact limit of satisfactory modeling from the noisy data using an inductive algorithm that can be attained by using actual external selection criteria or multicriterion analysis. The degree of noise stability of the selection criterion can be determined by gradually increasing the noise level of data and finding its critical value α^*, above which the criterion fails. Before going into experimental studies, we give an analogy with the well-established information theory.

1 ANALOGY WITH INFORMATION THEORY

The concept of a signal and its noise stability are well studied and established in the field of information theory [111]. The importance of the studies in information theory exerts a favorable influence on other branches of science and technology—in particular, with the self-organization theory. The information theory assumes that input signal is frequency-band limited and that an additive noise is superimposed on it (even if the noise level is very high). According to the self-organization theory, usually only a small sample of data represents the system. It takes into account the fact that additive noise is superimposed on the output variable. Comparison of the properties of different systems in modulating a

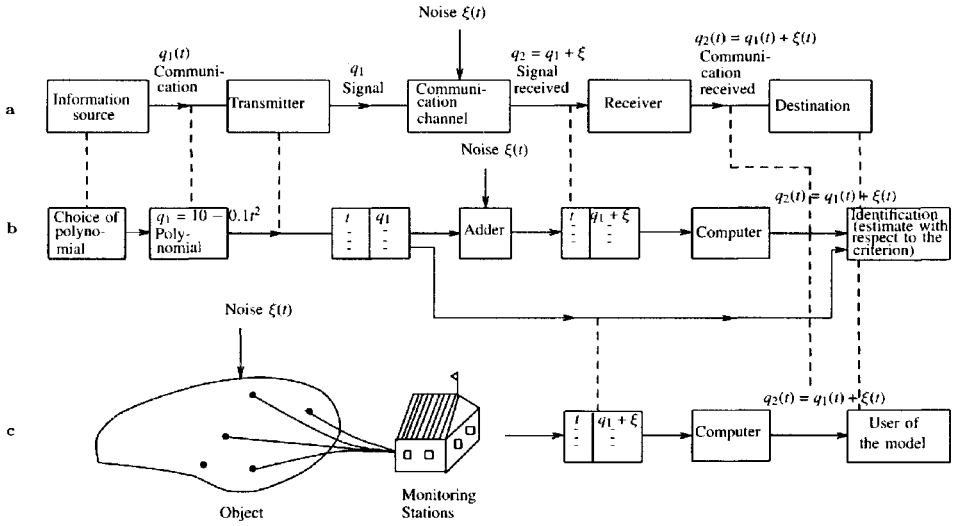

Figure 3.1. Schematic diagram of (a) a communication system, (b) a computational experimental setup, and (c) a self-organization modeling system

signal, which include Shannon's coding theory, constitutes an important part of information theory.

We give an analogy between the basic concepts of information theory and self-organization theory in identifying the processes. The main purpose of this analogy is to show the possibility of the exchange of basic ideas between these theories. We restrict our assumptions such that we are dealing only with simple amplitude modulation used in the communications and with the simplest polynomial (linear in weights) models of the form $q = a_0 + a_1x_1 + \cdots + a_mx_m$, where q is the dependent variable and x is the relabeled independent argument of nonlinear nature (for example, $q = 10 - 0.1t^2$).

In systems modeling, one usually considers the identification of a model only, and not the self-organization of predicting models, although communication theory does include a prediction method that is used for decreasing the redundancy of a signal. This does not restrict our study of drawing meaningful analogues between communication systems and self-organization modeling systems.

Let us put our analogy in the form of block diagrams as shown in Figure 3.1, where (a) is a communication system, (b) is a computational experimental setup, and (c) is a self-organization modeling system for obtaining an objective model (omitting the functions of specific elements).

In the communication system the information source chooses the particular form of communication from a set of possible communications. In the computational experimental scheme, we choose a polynomial (for example, $q = a_0 + a_1x = 10 - 0.1t^2$). In the self-organization system, the information source is the object of investigation (for example, ecological system) that "transmits communication" within a period of time.

In the communication system, the transmitter maps the space of communications into the space of non-noisy signals as $q_1(t) \rightarrow q_1$. In the computational experimental scheme, the polynomial $q_1(t)$ is represented in a data table with the columns of t and q_1. In the self-organizing system, the actual data is hidden in the system itself.

The communication channel in the communication system is the link at which noise intrudes. At its output, we obtain a copy of the signal; namely, the table of noisy data

with t and $q_2 = q_1 + \xi$. The noisy signal is received by the receiver and is mapped into the space of received communications $q_2 \rightarrow q_2(t)$. In all the systems, the data table with t and q_2 is transformed into a polynomial for $q_2(t)$, called the physical model. The receiver corresponds to the algorithm in self-organization modeling. The destination is the place where the communication (model) is expected to go.

Information theory studies the signal at the output of the communication channel; self-organization theory studies the experimental data sample at the output of the object of investigation. Overall, one can see that the most important parts of the systems from the communication channel to the destination or user is the same for all three systems.

Analogy between the approaches in information and self-organization theories. Both theories focus on the quasistatic part of the processes (known as the signal or trend) that consider noise as a dynamic component. Both of them assume that the data being processed contain information of true input signal that conceal the governing laws acting on the object. The objective goals concentrate on a receiving device for restoring as accurately as possible the original signal (governing laws); here the receiver corresponds to the modeling algorithm of self-organization modeling.

The information theory assumes that the signal at the input of a communication channel is frequency-band limited and that an additive noise is superimposed on it. The self-organization theory also takes into account that additive noise is superimposed on the output.

The communication theory pragmatically defines the "true input signal" $q_1(t)$ and the concept of noise $\xi(t)$; for example, a portion of the output voltage permitting transmission of communication appears in the signal. Similarly, in systems modeling, the useful part of the data is the part that is utilized for identification or prediction depending on the problem; everything else is noise. The noise hinders performance of modeling and lowers the minimum of criterion for selecting a model.

Information theory assumes that noise is independent of signal and additive with normal distribution. Self-organization theory asserts that if noise is independent, then the information theory is directly applicable; but if noise is dependent on the signal, it is applicable only to orthogonalized inductive algorithms.

1.1 Basic concepts of information and self-organization theories

Signal transmission time versus interval of data points. In the information theory, the signal at the input of a communication channel is characterized by the quantities: amplitude $q_1(t)$, power $P_1(t) = q_1^2$, frequency band ω_1, maximum transmitter frequency ω_{max}, signal-to-noise ratio as $\log_2(P_1/\xi^2)$, and volume $V_1 = \omega_1 T_1 \log_2(P_1/\xi^2)$.

The signal at the channel output is determined by the quantities: amplitude $q_2(t) = q_1(t) + \xi(t)$, power $P_2 = P_1 + \xi^2$, frequency band ω_2, signal-to-noise ratio as $\log_2(P_2/\xi^2)$, and channel volume $V_2 = \omega_2 T_2 \log_2(P_2/\xi^2)$. The signal duration T is analogous to the period of observations (length of experiment) of the modeling object; i.e., the total time interval of data observations from first observation to the last one. The divisions of data must be no wider than $1/(2\omega_1)$, where ω_1 is the frequency band. Consequently, the signal transmission time corresponding to the minimum length of the measurements is as follows:
when there is no noise,

$$T_1 = \frac{N_1}{2\omega_1} \text{ sec},$$

when there is noise,

$$T_2 = \frac{N_2}{2\omega_2} \text{ sec; here}$$

$$T_2 \leq T_1.$$

N_1 and N_2 are the algebraic minima of points required in self-organization modeling with and without noises, respectively. For polynomial models, the number of points is equal to the number of terms in the individual polynomials. For harmonic models, it is three times the number of harmonic components of the model. Here N specifies the number of terms in the polynomial. At the same time, it is also the minimum number of data points required to estimate the coefficients using the least-squares technique.

Transmission capacity versus minimization of external criterion. The transmission capacity C_t of a communication system in the sense of Hartley is logarithmic to the base two of the number of communications that can be transmitted per unit time with a given accuracy. The optimal admissibility of a communication system is given in terms of its transmission capacity (speed of transmission) as

$$C_t = \omega_2 \log_2 \left(\frac{P_1 + \xi^2}{\xi^2} \right) \text{ bits/sec.} \tag{3.1}$$

In time T, it is possible to transmit $J = C_t T$ bits of information through the communication system. The formula shows that for equal information, that is for J = constant, signal power P_1 can be traded off for bandwidth ω or for transmission time T, and so on.

In self-organization modeling, the problem is solved in a much more modest way. If we confine ourselves to stationary models with constant coefficients, we need to transmit only one communication; i.e., to construct a single model. The optimal system for obtaining a self-organizing linear model in the absence of noise requires a number of measurements equal to the algebraic minimum of the N_1 points.

We can treat the reciprocal of the minimum of the selection criterion as the analogue of the transmission capacity of the communication system ($C_t = k/\Delta(B)_{min}$), where k is an arbitrary constant. As noise increases, the minimum depth of the criterion decreases; i.e., the transmission capacity drops (Figure 3.2).

Transmission capacity versus noise stability. The noise stability of a communication system is determined by the minimum limiting admissible value of the signal-to-noise ratio for which it is still possible to receive the signal.

In self-organization modeling, one uses two limits. One of the limits is determined by the confidence level of the the external criterion through a computational experiment and the other by the polynomial structural changes.

The efficiency E of a communication system is directly proportional to the transmission capacity C_t and the maximum noise stability, and is inversely proportional to the signal observation time T.

The efficiency of an inductive learning algorithm is directly proportional to the ratio of the algebraic minimum number of points necessary for constructing the model to the number of points in the data table.

$$E = k \frac{N_1}{N_{max}} = k \frac{V_1}{V_{max}}. \tag{3.2}$$

The greater the ratio of the volume of the communication channel to that of the signal, the greater the noise stability, but the lower the efficiency of use of the given communication channel (or the efforts made to obtain the experimental data).

The efficiency of communication characterizes the possibility of transmission along channels with narrow-band with low energy expenditure. The efficiency of modeling character-

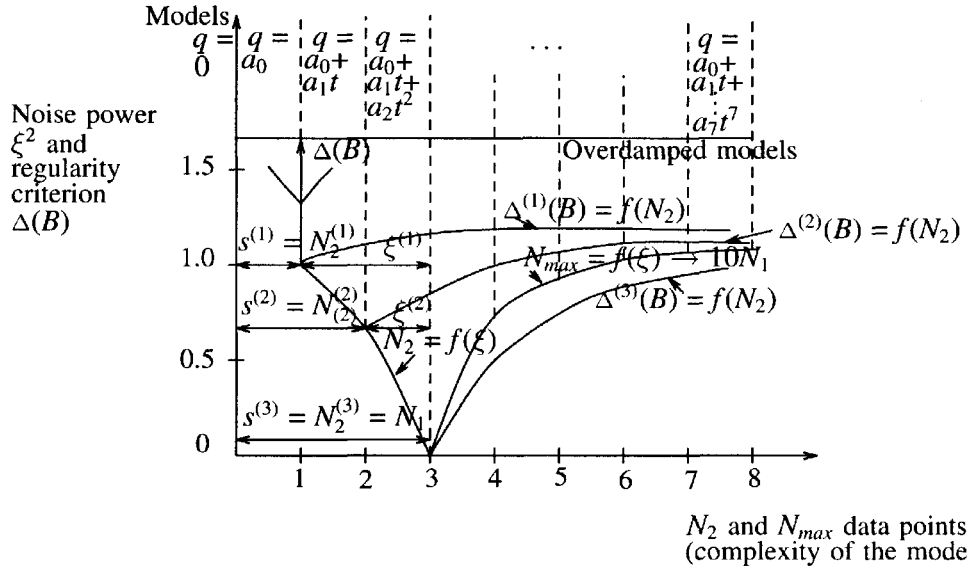

Figure 3.2. Decrease in the discrete values of the optimal complexity of model corresponding to N_2 and increase in the optimal length of the data sample N_{max} with increase in the noise power ξ^2 for a specified model complexity as at the point (3,0) as $q = 10 - 0.1t^2$

izes the possibility of constructing a sufficiently accurate model from a small number of points with a small expenditure of time on measurement, collection, and processing of data.

This can be applied directly to one-dimensional modeling problems, though two dimensional models require the introduction of two frequency bands as in two dimensional cases of communications [35].

1.2 Shannon's second theorem

The theorem is formulated as follows: Let P denote the signal power—supposing that the noise is independent—and white is the variance of ξ^2 in a frequency band ω. The optimal transmission speed attained is

$$C_{t\max} = \omega_2 \log_2\left(\frac{P + \xi^2}{\xi^2}\right) = \omega_2 \log_2\left(\frac{\omega_1}{\omega_1 - \omega_2}\right). \tag{3.3}$$

The greater the signal power in comparison with the noise variance, the greater will be the attainable transmission capacity. Thus, the theorem establishes a bound for the transmission capacity of the communication system that is attainable for optimal choice of the coding method and channel band ω_2 (the signal band ω_1 is assumed to be given) (Figure 3.3).

In self-organization modeling, the theorem enables us to choose the model with optimal complexity N_2 (complexity of the modeling object N_1 is given). The greater the noise, the lesser the depth of the minimum of the selection criterion, and the simpler the model (Figure 3.2). The theorem indicates the optimal (limiting attainable) values of the signal band (and the complexity of the models), and thus makes it clear why it is necessary, in the presence of noise, to use nonphysical models. The physical models correspond to the point (3,0) indicated in Figures 3.2 and 3.3.

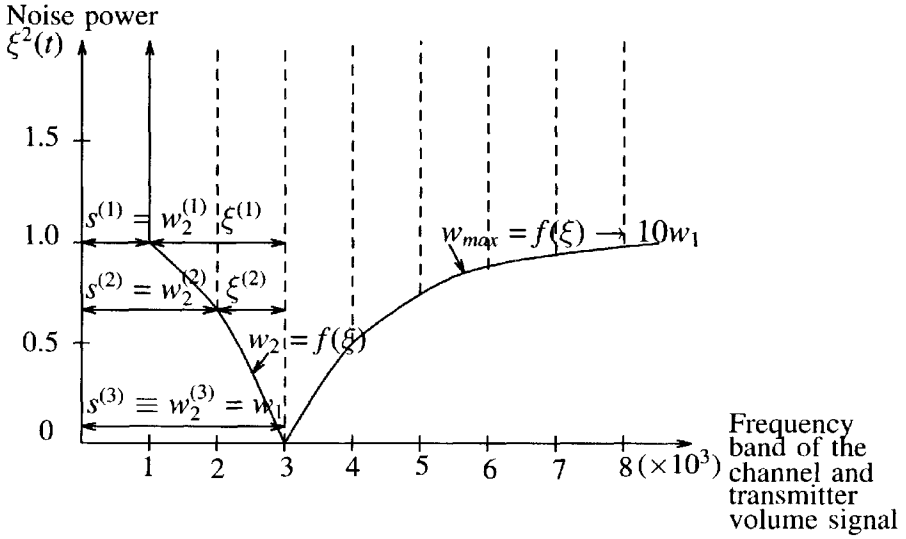

Figure 3.3. Decrease in the discrete values of the optimal band of the channel corresponding to ω_2 and increase in the optimal band of transmitter ω_{max} with increase in the noise power $\xi^2(t)$ for a given signal band of $\omega_1 = 3 \times 10^3 = \text{constant}$

Shannon's geometrical construction of the theorem

Shannon's geometrical construction is an interpretation of Shannon's second theorem (noisy coding theorem) about the limiting transmission capacity of a communication system. The input signal (like the table of input data in modeling) is at all times filled with new points with a discrete interval step of $\Delta \leq \frac{1}{2\omega_1}$. With the appearance of each new point, the dimension of the hyperspace increases by one. However, the mean value of the signal is stable. This is represented by a hypersphere with unit radius r_1 and with a center at zero. The noise is equal to the variance of the deviations of the signal from its average value. It is represented by a hypersphere with radius δ_{x_1} corresponding to noise and with the center at point A on the outer hypersphere with radius $r_2 (= 1 + \delta_{N_1}^2)$ and with the center at zero.

In the absence of noise, the number of models is infinite as they lie on the inner hypersphere of unit radius. With reference to the Figures 3.2 and 3.3, all of them correspond to the point (3,0) and are often combined into a single "physical model."

In the presence of noises, the number of models, called nonphysical, is finite and lies on the outer hypersphere of radius r_2, which satisfies the relationships of Shannon's limit theorem. In Figures 3.2 and 3.3 they correspond to the points (1,1) and (2,0.75).

- for Figure 3.2: $N_1 \equiv P + \xi^2$; $N_2 \equiv P$; $N_1 - N_2 \equiv \xi^2$
- for Figure 3.3: $\omega_1 \equiv P + \xi^2$; $\omega_2 \equiv P$; $\omega_1 - \omega_2 \equiv \xi^2$

so that

$$\frac{N_2}{N_1} = \frac{P}{P + \xi^2} \quad \text{or} \quad \frac{\omega_2}{\omega_1} = \frac{P}{P + \xi^2}. \tag{3.4}$$

If the noise power ξ^2 is given and the physical model corresponding to N_1 is known, then the theorem enables us to find a model of optimal complexity with N_2 deductively; i.e., without sorting the partial models. The theorems applies only to self-organization modeling on the basis of external criteria characterizing its accuracy (regularity, prediction, etc.).

Criterion of convolution stability

The basic purpose of geometric construction of the theorem is to find the area or chord lengths of the surface in the hypersphere (corresponding to noise) intersecting the inner hypershere of unit radius. Specifically, this is formulated as

$$h = \sqrt{2T\omega \frac{P\xi^2}{P+\xi^2}}, \tag{3.5}$$

where T is the signal duration, ω is the frequency band, P is the signal power, and ξ^2 is the noise variance.

This enables us to find a criterion for stability of convolution of chords that is convenient for self-organization modeling and that can be used for solving various problems such as pattern recognition and long-range predictions.

For example, in selecting the best predicting model, this is represented from analytical formulas of form:

$$h_{t_i} = \frac{2\sigma^2_{(t_i)}}{1+\sigma_i^2}, \tag{3.6}$$

where $\sigma_{(t_i)}$ and σ_i are the variances of the prediction models and the variable features correspondingly, and are calculated by averaging deviations. t is the prediction model numbers and i, the variable numbers $i = 1, 2, \cdots, m$.

h_t is computed as the mean chord length (convolution)

$$h_t = \sqrt{\frac{1}{m}(h^2_{t_1} + h^2_{t_2} + \cdots + h^2_{t_m})}. \tag{3.7}$$

For example, if there are ten prediction models, h_t is computed for each model. It is chosen so that the model for the convolution of the chords is least. If $h_{10} < h_t$, $t = \overline{1,9}$, then the optimal model according to the criterion of stability of chord length of Shannon's construction is the tenth.

Similar formulations are used to solve problems of pattern recognition and vector optimization. This criterion is also called Shannon's displacement criterion.

1.3 Law of conservation of redundancy

The properties of a communication system are determined by the value of its redundance. The properties are different for wide-band and narrow-band communication systems. In wide-band systems, the redundancy exceeds zero and the channel volume exceeds the signal volume $V_{max} > V_1$ or $\omega_{max} > \omega_1$. In wide-band systems for self-organizing models, the candidate models (from the very simplest models to the models whose complexity considerably exceeds the complexity of the actual or physical model) are put up for sorting according to a set of criteria. Algebraic models can serve as the analogue of a wide-band communication system in modeling. For them, increase or decrease of data points (with subsequent operations with the data table) is useless.

In narrow-band communication systems, the channel volume is less than the signal volume, and there is no redundancy; $V_{max} < V_1$ or $\omega_{max} < \omega_1$. The optimal relationships of Shannon's limit theorem (shown in Figures 3.2 and 3.3) are violated. In this case reduction of the signal proves feasible.

In narrow-band self-organization systems, we choose models whose complexity is no greater than the complexity of the actual model $N_{max} \leq N_1$. Finite-difference models of

complexity lower than the optimal model can serve as an example of a narrow-band link. Difference models are the analogues of differential equations only for small steps in the data sample.

Here the sequence of the following operations (using seasonal and annual values of a system)—collecting average seasonal data points of the variables, expanding the data table with the average annual values, self-organization modeling for obtaining a model with optimal complexity, detailed identification (seasonal data), and smoothed identification (annual data)—can be extremely efficient. Without expanding the data table, the model with optimal complexity cannot be attained because of the insufficient number of points of initial data.

1.4 Model complexity versus transmission band

In self-organization modeling, one often uses the term "complexity of a model." The complexity of the models is gradually increased until the minimum of the selection criterion is found. In linear polynomials, the complexity of the model is determined by the number of terms on the right-hand side of the equations.

The complexity of models obtained from the inductive algorithms varies from zero to N_{max} and passes through the value N_2 sought. In connection with this, in self-organization modeling, it is convenient to look at the quantities $N_1 = 2\omega_1 T_1$ (the algebraic minimum of points necessary for obtaining the true physical model), $N_2 = 2\omega_2 T_2$ (the algebraic minimum of points necessary for obtaining the optimal model using the inductive learning algorithm), and $N_{max} = 2\omega_{max} T_{max}$ (the algebraic minimum of points necessary for the most complex model that can be obtained as a result of self-organization, or the number of data points actually represented in the data table).

The following laws (Figure 3.2) come into effect in self-organization modeling:

1. In the absence of noise, beginning with some complexity equal to the complexity of the actual model N_1, further increase in complexity is not required; for $\xi^2(t) = 0$, we have $N_2 = N_1$ and $N_{max} \geq N_1$.
2. In the presence of noise, the model with optimal complexity appears earlier. The algebraic minimum of points (the complexity of the optimal model) decreases; for $\xi^2 > 0$, we have $N_2 < N_1$ and $N_{max} \geq N_2$.

The analogous laws are known in information theory (Figure 3.3). Since the bandwidths ω_1 and ω_2 can only be approximated, every communication channel gives distortion, just as every data sample, even when $V_2 = V_1$:

1. For exact transmission of a communication, it is necessary for the channel volume to be at least equal to the signal volume; for $\xi^2(t) = 0$, we have $V_2 = V_1$ and $V_{max} \geq V_1$.
2. When there are noises, the optimal channel volume is somewhat less than the signal volume; for $\xi^2 > 0$, we have $V_2 < V_1$ and $V_{max} \geq V_2$.

This means that the transmission band of special receivers designed for operation under noise conditions is narrower than wide-band receivers intended for the case of small noise. Thus, the communication channel band is analogous to the model complexity estimated according to the algebraic minimum of points

$$V_1 = \omega_1 T_1 \log_2(P_1/\xi^2) = \frac{N_1}{2} \log_2(P_1/\xi^2),$$

$$V_2 = \omega_2 T_2 \log_2(P_2/\xi^2) = \frac{N_2}{2} \log_2(P_2/\xi^2). \tag{3.8}$$

The influence of noise on the model accuracy can be overcome to some extent by increasing the number of measurements. However, when the number of data points becomes excessive, the accuracy and noise stability of the model decrease. Thus, there exists an optimal number of data points for stationary and non-stationary signals. Because of the necessity of decreasing the influence of noise, one chooses the table length about 10 times as great as the algebraic minimum of the points $T_\xi = 10.T_1$ (Figure 3.3). During this interval, the system will collect $J = C_i T_\xi = \omega . \log_2(\frac{P_1+\xi^2}{\xi^2}).T_\xi$ bits of information.

An analogy between the optimal complexity of models for the inductive algorithm and the transmission band for a communication system is shown in Figures 3.2 and 3.3, where N_1 is the complexity of the physical model, N_2 is the complexity of the non-physical model of optimal complexity, N_{max} is the optimal range of complexity of model candidates, ω_1 is the band of the true signal, ω_2 is the optimal band of the receiver, and ω_{max} is the optimal volume of the transmitter signal.

The law of compromization. The important result of investigations arrived at through the information theory is the establishment of a connection between transmission capacity and noise stability. Increase in noise stability decreases transmission capacity. Here one varies the parameters of the communication system; for example, by varying the frequency band ω_2.

An analogous law holds for self-organization modeling using the selection criterion such as regularity; an increase in the power (amplitude) of the noise leads to the choice of simpler noise-stable models for which the algebraic minimum of points is less than that of the model of the object obtained under conditions of absence of noise. Here one varies the parameters of the model; for example, by varying the algebraic minimum of points N_2.

Here we conclude that the noisy coding theorem (Shannon's second theorem) plays a central role in this analogy between information theory and self-organization theory. In fact, the theorem states that it is possible to transmit information through the channel with as small a probability of error as desired if it is transmitted at any rate less than the channel capacity. In other words, it guarantees the existence of a code that may be transmitted at any rate close to but less than that of channel capacity and still be received and decoded with arbitrarily small probability of error. It proves that channel capacity is a fundamental property of a communication channel. This is conceptualized analogously to the theory of self-organization modeling. In particular, it shows that, in the presence of noise, non-physical models obtained by self-organization modeling are optimal.

2 CLASSIFICATION AND ANALYSIS OF CRITERIA

Let us assume that the initial data is given in the form of the matrices

$$y = \begin{pmatrix} y_A \\ - - - \\ y_B \\ - - - \\ y_C \end{pmatrix}, \quad X = \begin{pmatrix} X_A \\ - - - \\ X_B \\ - - - \\ X_C \end{pmatrix}, \quad \begin{array}{r} y\,[N \times 1] \\ X\,[N \times m] \\ N_A + N_B + N_C = N \\ N_A + N_B = N_W \end{array} . \tag{3.9}$$

The entire data sample is partitioned into three disjoint subsets A, B, and C. The set W is the union of A and B. The optimal dependence relation between output y and input variables X is sought by the inductive learning that are linear in the coefficients of $y = X\hat{a}$.

It is assumed that the submatrices X_A and X_B, which are used in the selection process to any particular model of complexity $s(\leq n)$, are of complete rank.

The external criteria used in the inductive algorithms can be expressed in terms of the estimates of the output variables of the models and their coefficients obtained on A, B, W, and C. Here the basic quadrature, and combined and correlational criteria are described.

All the external criteria that have the quadratic form can be grouped into two basic groups: (i) accuracy criteria, which express the error in the model being tested on various parts of the model and (ii) matching (consistent) criteria, which are a measure of the consistency of the estimates obtained on different sets. There are symmetric and nonsymmetric forms of the criteria in both the groups, where symmetric means one in which the information in sets A and B is used equally; otherwise, it is nonsymmetric.

2.1 Accuracy criteria

Regularity criterion (nonsymmetric)

This is the typical quadratic criterion and historically the first one.

$$\begin{aligned}\Delta^2(B) = \Delta^2(B|A) &= \|y_B - \hat{y}_B^A\|^2 = (y_B - X_B\hat{a}_A)^T(y_B - X_B\hat{a}_A) \\ &= \|y_B - X_B\hat{a}_A\|^2, \end{aligned} \tag{3.10}$$

where $\hat{a}_A = (X_A^T X_A)^{-1} X_A^T y_A$, and $\hat{y}_B^A = X_B\hat{a}_A$.

We can obtain another nonsymmetric regularity criterion by replacing A by B and, vice versa, $\Delta^2(A) = \Delta^2(A/B)$.

Regularity criterion (symmetric)

This can be built up using the both the nonsymmetric versions of the regularity criterion.

$$\begin{aligned} d^2 = d^2(A, B|B, A) &= \Delta^2(B|A) + \Delta^2(A|B) \\ &= \|y_B - X_B\hat{a}_A\|^2 + \|y_A - X_A\hat{a}_B\|^2, \end{aligned} \tag{3.11}$$

where sets A and B are used equally. It smooths out the influence of the noise that acts on both parts of the data sample.

Stability criterion (nonsymmetric)

If we require an optimal model, which must be sufficiently accurate on both the sets—training set A and testing set B for the coefficients estimated on the set A—then this compromise can be obtained by the criterion

$$\begin{aligned}\kappa^2 = \kappa^2(W|A) &= \|y_W - X_W\hat{a}_A\|^2 \\ &= \Delta^2(A|A) + \Delta^2(B|A) = \varepsilon^2(A) + \Delta^2(B), \end{aligned} \tag{3.12}$$

where $\varepsilon^2(A)$ is the least squares error or residual sum of squares.

Stability criterion (symmetric)

$$\begin{aligned} S^2 = S^2(W|A, B) &= \kappa^2(W|A) + \kappa^2(W|B) \\ &= \|y_W - X_W\hat{a}_A\|^2 + \|y_W - X_W\hat{a}_B\|^2. \end{aligned} \tag{3.13}$$

It is expedient to use this criterion if the finiteness of the data is considered. The sensitivity to the separation of data is lowered and the influence of noise is averaged (a kind of filtering takes place). In other words, this has higher noise immunity.

Averaged regularity criterion [122]

According to this criterion the mean value is calculated on N_W for each particular model being tested under the condition that each point in the set W is, in its turn, the testing sample and the remaining $N_W - 1$ points constitute the training sample.

$$\Delta^2_{av}(W) = \frac{1}{N_W} \| y_j - \hat{y}_j^{W_j} \|^2_{j \in W}, \tag{3.14}$$

where $\hat{y}_j^{W_j} = x_j^T \hat{a}_{W_j}$, x_j are the argument measures at the jth point, W_j is the training sample without jth point, and $\hat{a}_{W_j}$ is the estimate of the coefficients on W_j. It is expedient to use this criterion for a small number of points.

Step-by-step prediction criterion

In case of finite-difference equations, it is expedient to use this external integral criterion.

$$i^2(W) = i^2(W|W) = \| y_W - \hat{y}_W^W \|^2, \tag{3.15}$$

where the estimate $\hat{y}_W^W$ is obtained by step-by-step integration of the difference equation from the given initial conditions. This criterion can also have the forms of $i^2(A)$ and $i^2(B)$.

The above accuracy criteria, like all other types of criteria, are used in modeling of both static and dynamic objects.

2.2 Consistent criteria

The criteria in this group do not take into account the error of the model in explicit form, but measures the consistency of the model on two different data sets.

Criterion of minimum coefficient bias

This reflects the requirement that the coefficient estimates in the optimal model estimated on sets A and B, differ minimally; i.e. they are in agreement.

$$\eta_a^2 = \eta_a^2(A, B) = \| \hat{a}_A - \hat{a}_B \|^2. \tag{3.16}$$

Minimum bias criterion

This is the most widely used form of the criterion.

$$\begin{aligned} \eta_{bs}^2 = \eta_{bs}^2(W|A, B) &= \| \hat{y}_W^A - \hat{y}_W^B \|^2 \\ &= \| X_W \hat{a}_A - X_W \hat{a}_B \|^2 \\ &= (\hat{a}_A - \hat{a}_B)^T X_W^T X_W (\hat{a}_A - \hat{a}_B), \end{aligned} \tag{3.17}$$

which differs from η_a by the presence of the weight matrix $X_W^T X_W$ and expresses a different minimum requirement of consistency on the set W from the estimates of the model outputs that obtained coefficients from the sets A and B.

Absolute noise immune criterion

$$V^2 = V^2(W|A,B,W) = (\hat{a}_W - \hat{a}_A)^T X_W^T X_W (\hat{a}_B - \hat{a}_W)$$
$$= (\hat{y}_W^W - \hat{y}_W^A)^T (\hat{y}_W^B - \hat{y}_W^W). \tag{3.18}$$

This uses the estimates of the model output for the coefficients obtained on three sets—A, B, and W. It got its name because it satisfies the most important condition of noise immunity. It rejects excessively complex models under noise conditions [67].

The above minimum bias criteria are symmetric. It is easy to write nonsymmetric forms of η_{bs} and V^2. For example, on the set B [129]

$$\eta_{bs}(B) = \|\hat{y}_B^A - \hat{y}_B^B\|^2 = \|X_B\hat{a}_A - X_B\hat{a}_B\|^2, \tag{3.19}$$

$$V^2(B) = (\hat{a}_W - \hat{a}_A)^T X_B^T X_B (\hat{a}_B - \hat{a}_W). \tag{3.20}$$

One useful way is to clarify the connections among certain external criteria. One can easily show that $\eta_{bs}^2(W) = \eta_{bs}^2(A) + \eta_{bs}^2(B)$ and, in the same way, $V^2(W) = V^2(A) + V^2(B)$ because of the relation $X_W^T X_W = X_A^T X_A + X_B^T X_B$.

2.3 Combined criteria

In addition to the criteria $c1, c2$, and $c3$ introduced in chapter 1, here is another form of the combined criterion $c4$.

Minimum bias plus symmetric regularity

$$c4^2 = \bar{\eta}_{bs}^2 + \bar{d}^2. \tag{3.21}$$

It is recommended that the sequential use of two-criterion selection is preferred in the combined criteria. F number of models are selected using the consistent criterion like η_{bs}^2, then the best model is selected using a accuracy criterion like $\Delta^2(C)$. Such sequential application of the criteria increases the efficiency of the modeling, including noise immunity.

2.4 Correlational criteria

These criteria impose definite requirements on the relationship of correlation characteristics of the output variables of model and the object. Unlike the quadratic criteria, they can be both positive and negative. This is one of the reasons for separating them as a special group of criteria. Their applicability for model selection is ensured by the fact that coefficients of the model are estimated on the set A and the correlation relationship is computed with respect to the set B.

Correlational regularity criterion

$$K(B) = \frac{(y_B - \bar{y}_B)^T (\hat{y}_B^A - \bar{\hat{y}}_B^A)}{\|y_B - \bar{y}_B\| \cdot \|\hat{y}_B^A - \bar{\hat{y}}_B^A\|}, \tag{3.22}$$

where y_B is the actual output; $\hat{y}_B^A$ is the model output, the coefficients of which are estimated on set A; $\bar{y}_B$ and $\bar{\hat{y}}_B^A$ are the mean values of the actual and model outputs, respectively. The best model is based on the condition $K(B) \to 1$.

Table 3.1. Classification of external criteria

Type	Criterion	Criterion form	
		nonsymmetric	symmetric
Accuracy	regularity	$\Delta^2(B), \Delta_2(A)$	$d^2(W)$
	stability	$\kappa^2(B), \kappa^2(A)$	$S^2(W)$
	averaged regularity	-	$\Delta^2_{av}(W)$
	prediction	$i^2(B), i^2(A)$	$i^2(W)$
Consistent	minimum bias	$\eta^2_{bs}(A), \eta^2_{bs}(B)$	$\eta^2_{bs}(W)$
	abs. noise immune	$V^2(A), V^2(B)$	$V^2(W)$
Correlational	regularity	$K(B), K(A)$	$K(A) + K(B)$
	NL agreement	$J_s(B), J_s(A)$	$J_s(A) + J_s(B)$
Combined	bias + regularity	$\eta^2_{bs} + \Delta^2(B)$	$\eta^2_{bs} + d^2$
	bias + MSE	$\eta^2_{bs} + \varepsilon^2(A)$	$\eta^2_{bs} + \varepsilon^2(W)$
	bias + prediction	$\eta^2_{bs} + \Delta^2(C)$	-

Correlational criterion with nonlinear agreement [129]

This has three different components; one is equivalent to the correlational regularity, the other is the agreement criterion for the degree of nonlinearity, and the third is the agreement criterion for the mean values of the actual and estimated outputs. These components are based on the mean-squared error as follows:

$$\begin{aligned}\varepsilon^2 &= \frac{1}{N}(y-\hat{y})^T(y-\hat{y}) \\ &= (1 - J_c^2 + J_s^2 + J_m^2)z_v^2,\end{aligned} \tag{3.23}$$

where y_i and $\hat{y}_i$, $i \in N$ are the actual and estimated outputs of N data points. The quantities J_c, J_s, and J_m are expressed in terms of the centered vectors $v = y - \bar{y}$ and $\hat{v} = \hat{y} - \bar{\hat{y}}$ and the estimates of the variances as $z_v = \sqrt{(v^T v/N)}$ and $z_{\hat{v}} = \sqrt{(\hat{v}^T \hat{v}/N)}$.

$$J_c = r(\hat{v}, v) = \hat{v}^T v / N z_{\hat{v}} z_v \tag{3.24}$$

$$J_s = z_{\hat{v}}/z_v - r(\hat{v}, v) \tag{3.25}$$

$$J_m = (\bar{y} - \bar{\hat{y}})/z_v. \tag{3.26}$$

It was proposed in [129] that the components J_c, J_s, and J_m of the error vector can be used as independent selection criteria, calculated on the set B with the estimates $\hat{a}_A$ obtained on the set A. The component $J_c(B)$ coincides exactly with the criterion $K(B)$. The component $J_s(B)$ is called the agreement criterion for the degree of nonlinearity; this should satisfy the condition $J_s(B) \to 0$. The component $J_m(B)$ is also called the agreement criterion for the mean values, but it does not seem to have any independent significance. One can convert the criterion $J_c(B) \equiv K(B)$ into a minimization form $|1 - K(B)| \to$ min.

The above mentioned correlational criteria are nonsymmetric; to make them symmetric, an expression must be added to each in which the sample parts A and B swap roles. Various groups of criteria are given in Table 3.1.

2.5 Relationships among the criteria

In this section we derive the number of relationships that express the connection among different external criteria.

Let us consider the quadratic criteria of symmetric type. We can write the relationship of S^2 in terms of d^2.

$$S^2 = d^2 + \varepsilon^2(A) + \varepsilon^2(B), \tag{3.27}$$

where $\varepsilon^2(A) = \Delta^2(A|A)$ and $\varepsilon^2(B) = \Delta^2(B|B)$ are the mean square errors on the sets A and B, respectively. We can write the minimum bias criterion as

$$\begin{aligned} \eta_{bs}^2 &= \|(y_W - \hat{y}_W^B) - (y_W - \hat{y}_W^A)\|^2 \\ &= \|y_W - \hat{y}_W^A\|^2 + \|y_W - \hat{y}_W^B\|^2 \\ &\quad -2(y_W - \hat{y}_W^A)^T(y_W - \hat{y}_W^B), \end{aligned} \tag{3.28}$$

since $X_W^T X_W = X_A^T X_A + X_B^T X_B$, $X_W^T y_W = X_A^T y_A + X_B^T y_B$, $\hat{a}_B = (X_B^T X_B)^{-1} X_B^T y_B$, $\hat{a}_A = (X_A^T X_A)^{-1} X_A^T y_A$. The term from the above expansion can be evaluated further as

$$\begin{aligned} (y_W - \hat{y}_W^A)^T(y_W - \hat{y}_W^B) &= y_A^T y_A - y_A^T X_A \hat{a}_A + y_B^T y_B - y_B^T X_B \hat{a}_B \\ &= \varepsilon^2(A) + \varepsilon^2(B). \end{aligned} \tag{3.29}$$

Knowing this, we can obtain a relationship between S^2 and η_{bs}^2:

$$S^2 = \eta_{bs}^2 + 2(\varepsilon^2(A) + \varepsilon^2(B)); \tag{3.30}$$

between d^2 and η_{bs}^2

$$d^2 = \eta_{bs}^2 + \varepsilon^2(A) + \varepsilon^2(B). \tag{3.31}$$

Now let us consider the criterion V^2:

$$\begin{aligned} V^2 &= [(y_W - \hat{y}_W^A) - (y_W - \hat{y}_W^W)]^T[(y_W - \hat{y}_W^W) - (y_W - \hat{y}_W^B)] \\ &= (y_W - \hat{y}_W^A)^T(y_W - \hat{y}_W^W) + (y_W - \hat{y}_W^B)^T(y_W - \hat{y}_W^W) \\ &\quad -(y_W - \hat{y}_W^A)^T(y_W - \hat{y}_W^B) - \|(y_W - \hat{y}_W^W)\|^2. \end{aligned} \tag{3.32}$$

The term $(y_W - \hat{y}_W^A)^T(y_W - \hat{y}_W^B)$ is given above as $\varepsilon^2(A) + \varepsilon^2(B)$. Since $\hat{a}_W = (X_W^T X_W)^{-1} X_W^T y_W$, one can obtain the relation as

$$\begin{aligned} (y_W - \hat{y}_W^A)^T(y_W - \hat{y}_W^W) &\equiv (y_W - \hat{y}_W^B)^T(y_W - \hat{y}_W^W) \\ &\equiv \|y_W - \hat{y}_W^W\|^2 = \varepsilon^2(W) \end{aligned} \tag{3.33}$$

by establishing that $y_W^T \hat{y}_W^A \equiv \hat{y}_W^{WT} \hat{y}_W^A$, and $y_W^T \hat{y}_W^B \equiv \hat{y}_W^{WT} \hat{y}_W^B$. Ultimately, we can obtain the formula [35]

$$V^2 + \varepsilon^2(A) + \varepsilon^2(B) = \varepsilon^2(W). \tag{3.34}$$

Using the above examination, one can easily write the relationships

$$V^2 + S^2 = d^2 + \varepsilon^2(W) \tag{3.35}$$

$$V^2 + d^2 = \eta_{bs}^2 + \varepsilon^2(W). \tag{3.36}$$

One can show that the absolute noise immune criterion V^2 is a quadratic, not a nonnegative one, by the relation

$$\begin{aligned} V^2 &= (\hat{a}_W - \hat{a}_A)^T X_W^T X_W (\hat{a}_B - \hat{a}_W) \\ &= \hat{a}_A^T X_A^T X_A \hat{a}_A + \hat{a}_B^T X_B^T X_B \hat{a}_B - \hat{a}_W^T X_A^T X_A \hat{a}_W - \hat{a}_W^T X_B^T X_B \hat{a}_W. \end{aligned} \tag{3.37}$$

This can be expressed into the sum of two quadratic forms:

$$V^2 = \|\hat{y}_A^A - \hat{y}_A^W\|^2 + \|\hat{y}_B^B - \hat{y}_B^W\|^2$$
$$= (\hat{a}_A - \hat{a}_W)^T X_A^T X_A (\hat{a}_A - \hat{a}_W) + (\hat{a}_B - \hat{a}_W)^T X_B^T X_B (\hat{a}_B - \hat{a}_W), \quad (3.38)$$

so that we can always have $V^2 \geq 0$.

The relationships established between the criteria S^2, d^2, and V^2 interconnect all the symmetric quadratic criteria. In addition to this, the formulas reexpressed for the criteria S^2 and d^2 in terms of the minimum bias and mean-square errors allow one to group these criteria into the group of the combined criteria. They are, however, fundamentally different from the combined criteria because of the components included in them and there is no need for normalization.

Similarly, one can obtain the relationships connecting the nonsymmetric criteria. For example, the regularity criterion $\Delta^2(B)$ can be represented [129] as

$$\Delta^2(B) = \varepsilon^2(B) + \eta_{bs}^2(B), \quad (3.39)$$

where $\eta_{bs}^2(B)$ is the nonsymmetric form of the minimum bias criterion on the set B.

The connection can be established among the regularity criterion and the correlational criteria directly from the relationship $\varepsilon^2 = (1 - J_c^2 + J_s^2 + J_m^2) z_v^2$ as

$$\Delta^2(B) = (1 - K^2(B) + J_s^2(B) + J_m^2(B)) z_v^2 N. \quad (3.40)$$

The representations of some criteria in terms of other criteria enable us to determine the characteristics of unique models derived from the original ones. For example, after calculating the squared errors $\varepsilon^2(A), \varepsilon^2(B), \Delta^2(A), \Delta^2(B)$, and $\Delta^2(C)$, one can also determine d^2, S^2, and η_{bs}^2 directly; after estimating $\varepsilon^2(W)$ one can calculate V^2.

The reader can find the usage of canonical forms in analyzing the noise immunity of quadratic criteria in the works [135], [119]. Here the expected value of the criterion is considered the sum of two components: one takes into account the non-noisy data and decreases (possibly nonmonotonically) with the increase of complexity of models; the second reflects the presence of noise that is directly proportional to its variance and increases monotonically with the increase of complexity of the models. With an increase in the noise level, the minimum of the external criteria (V^2 and d^2) moves into the region of simpler structures, which is analogous to the behavior of the ideal external criterion.

3 IMPROVEMENT OF NOISE IMMUNITY

We assume that noise can be additive, multiplicative, or a combination of these two types and that it does not contain a regular component. When the noise intensity (amplitude) is very high, the external criterion used might select a model that does not correspond to the system under study. The criterion is called noise-immune if it selects the true model even at a significant level of noise immunity. The analytical properties of selection procedures based on certain selection criteria are given here. Emphasis is made on improving the noise stability of the criteria in extracting the optimal model with true structure in the presence of noise. This identifies the true structure by comparing different structures that determine the maximum allowed noise level.

Let us assume that y is an output variable with a normally distributed noise. Its unit variance is represented as

$$y = \overset{\circ}{y} + \xi, \quad E[\xi] = 0, \quad \sigma^2 = E[\xi^T \xi] = 1, \quad (3.41)$$

where $\overset{\circ}{y}$ is the noise-free output connecting a set of m arguments (input variables) as $\overset{\circ}{y} = \phi(z)$. Let us assume that we have some noise realization of ξ_0 with N values. We obtain a series of output data for varying intensity of this realization of the noise, yielding N values of the function

$$y_\xi = \overset{\circ}{y} + \xi_0. \tag{3.42}$$

The sample of noise-free data obtained from the function $\overset{\circ}{y} = \phi(z)$ can be called the signal and output samples for different variances of ξ can be called the signal with noise. Each sample of noisy data is characterized by a value of the noise-to-signal ratio or by the noise level as

$$\alpha^2 = \frac{\sigma^2}{s^2} = \sigma^2 / \sum_{j=1}^{N} (\overset{\circ}{y}_j - \bar{y}^\circ)^2, \tag{3.43}$$

where s^2 and σ^2 are the signal and noise variance or power, correspondingly; $\bar{y}^\circ$ is the average value of the signal. For a fixed signal the variance and noise level are connected by a one-to-one relationship $\sigma^2 = \alpha^2.s^2$ or $\sigma = \alpha.s$.

Suppose the function $\phi(z)$ is a linear (in coefficients) convolution of some number of functions (for example, a set of polynomial functions $f_1(z), f_2(z), \cdots, f_m(z)$), equivalent to the vector of arguments $x = f(z)$. Then for each noise level α, the exact model is restored by optimizing the structure and estimating the coefficients a of the model.

$$y = a^T f(z) = a^T x \tag{3.44}$$

for the given sample of input and output values.

Here two types of study results are presented to show the efforts in improving the noise immunity of various external criteria. The first part consists of the initial studies [129] conducted on the minimum-bias criterion. This reveals the importance of the extension of the time interval to the extrapolation region of the data and shows that the largest noise immunity was possessed by special forms of the criterion with some specified general properties. The second part is concerned with the finding of noise stability of various criteria (single- as well as two-criterion analysis) by increasing noise levels for different data divisions. This gives some comparative results on several most commonly applied criteria for obtaining single- and two-criterion choices of models.

3.1 Minimum-bias criterion as a special case

The original form of the minimum bias criterion is

$$\eta_{bs}^2 = \sum_{p=1}^{N_W} \frac{(\hat{y}^A - \hat{y}^B)_p^2}{y_p^2}, \tag{3.45}$$

where $\hat{y}^A$ is the estimated output of the model, the coefficients of which are obtained using the set A; $\hat{y}^B$ is the estimated output of the model, the coefficients of which are obtained using the set B; and y is the actual output.

Geometric interpretation of the minimum bias

Suppose in an N-dimensional space R^N (N is the length of the data sample), $\hat{y}_{LS}$ is an orthogonal projection of the vector $y^T = (y_1, y_2, \cdots, y_N)$ from the output of the linear model

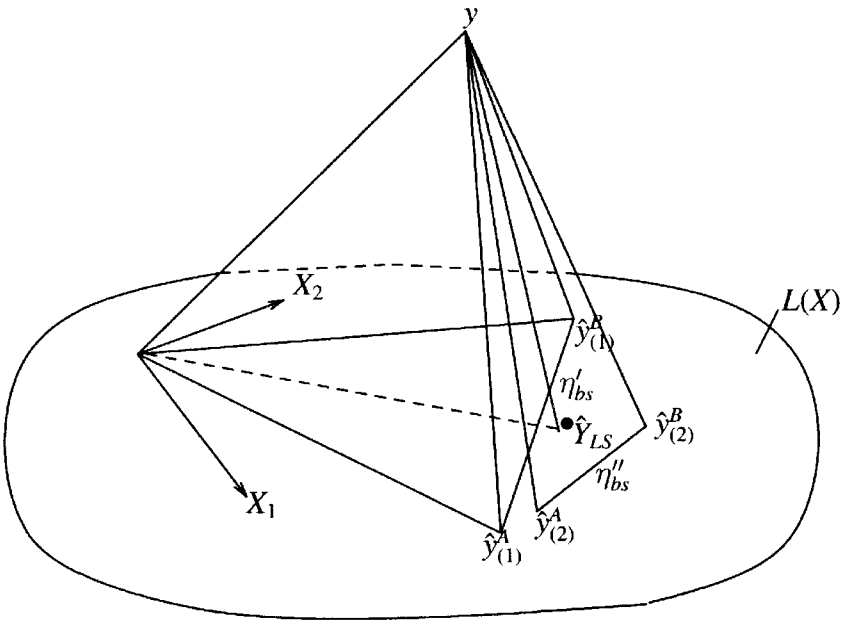

Figure 3.4. Minimum bias of solutions as a distance between projections of y for different divisions of data sample

$y = x^T a$, which is estimated by the least squares method onto a linear subspace $L(X)$ (Figure 3.4), formed by the vectors of m arguments $X_i^T = (x_1, x_2, \cdots, x_N)$, $i = 1, 2, \cdots, m$, ie., $a, x \in R^m$, and $x_i \in R^N$; and $X[N \times m]$ consists of the sets of matrices with real elements.

The data sets A and B are used as training sequence to estimate the coefficients of two models of similar structure and to the approximations of y as a total sample. Projections $\hat{y}^A$ and $\hat{y}^B$ of the same vector y on to the $L(X)$ are formed and these are usually non-orthogonal. Each ith version of the division of the data has corresponding vectors $\hat{y}^A_{(i)}$ and $\hat{y}^B_{(i)}$ belonging to $L(X)$. The ensemble of such vectors forms a "cluster of projections"; i.e., a set of points in $L(X)$; all points of the cluster are grouped around $\hat{y}_{LS}$. Models with false structures are more sensitive to the variations of the training sequence and, as a result, become significantly displaced from $\hat{y}_{LS}$—causing the cluster to widen. Different forms of the minimum bias criterion provide us with the possibility of estimating the dimensions of this cluster; i.e., an ability to compare different models.

(i) *Increasing the time interval of data in the criterion by introducing a noise immunity coefficient* δ_T. The minimum bias criterion has a relatively low noise-immunity because the approximating properties of the models are usually identical on the interval of interpolation. The squared errors are small for models of any structure except for the simplest linear models. The performance of models diverge in the extrapolation interval in which the differences between the model outputs become significant and, consequently, more immune to noise. Figure 3.5 shows an example of bias estimation for two polynomial models. The shaded areas indicate the differences of two sets of models (area between the integral curves $\hat{y}^A_{(1)}$ and $\hat{y}^B_{(1)}$ and area between the curves $\hat{y}^A_{(2)}$ and $\hat{y}^B_{(2)}$). Consequently, the bias estimate of the second polynomial is significantly smaller than the first one; i.e., $\eta^2_{bs_2} \ll \eta^2_{bs_1}$.

Here it is recommended that the minimum bias criterion with additional length of time

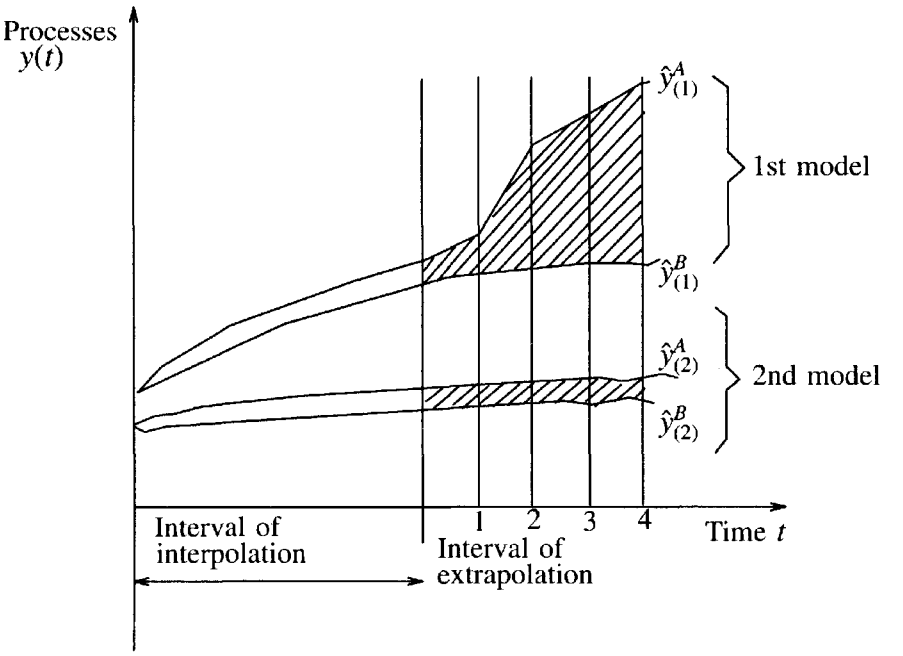

Figure 3.5. Explanation of estimation of bias of two models

interval should be

$$\eta^2_{bs} = \sum_{p=1}^{\delta_T N_W} \frac{(\hat{y}^A - \hat{y}^B)^2_p}{y^2_p}. \tag{3.46}$$

One can notice that it is applicable only to the nonlinear functions that have quadratic or higher order arguments.

(ii) *Extraction of first harmonic of the output variable.* The output variable is approximated with the harmonic equation using the sets A and B as

$$\begin{aligned} y^A &= a_0 + a_1 \sin(\omega_1 t + Q_1), \\ y^B &= b_0 + b_1 \sin(\omega_2 t + Q_2), \end{aligned} \tag{3.47}$$

where ω_1 and ω_2 are the fundamental frequencies, Q_1 and Q_2 are the phase shifts, and a's and b's are the estimated coeficients. It is assumed that the frequency expansion of the useful signal without noise occupies a portion of the spectrum which is different from the signal with noise. If this is justified, then the noise immunity of the minimum bias criterion can be increased because of the first harmonic. The first harmonic of set A should coincide as nearly as possible with the first harmonic of set B. The minimum bias criterion is recommended as

$$\eta^2_{bs} = \sum_{p=1}^{\delta_T N_W} \frac{(\hat{y}^A - \hat{y}^B)^2_p}{(\hat{y}^A + \hat{y}^B)^2_p}, \tag{3.48}$$

where $\hat{y}_A$ and $\hat{y}_B$ are the estimated outputs of the first harmonics. In practical situations, the fundamental frequencies ω_1 and ω_2 should be closer within the limits of the specified set

of structures. If the spectral content of the signal and noise are identical, then it is difficult to filter out the noise.

(iii) *Extraction of the linear trend of output variable.* If the outputs of the models are smooth and can be approximated by polynomials, extracting linear parts instead of first harmonics is recommended.

$$y^A_{(\text{lin})} = a_0 + a_1 t,$$
$$y^B_{(\text{lin})} = b_0 + b_1 t. \tag{3.49}$$

The two models to be identified have identical structures. The model $y^A_{(\text{lin})}$ should coincide as nearly as possible with the model $y^B_{(\text{lin})}$; i.e., the structure of the model is estimated in accordance to the minimum bias criterion

$$\eta^2_{bs} = \sum_{p=1}^{\delta_T N_W} \frac{(\hat{y}^A_{(\text{lin})} - \hat{y}^B_{(\text{lin})})^2_p}{(\hat{y}^A_{(\text{lin})} + \hat{y}^B_{(\text{lin})})^2_p}. \tag{3.50}$$

This provides justification to calculate the minimum bias criterion based on the linear parts. When the linear parts of the models are slightly dependent on the noise, such a criterion will have an increased noise immunity.

Example 1. An experiment is conducted to show the effect of the data interval on the noise immunity of the criterion. The true model is taken as $\overset{\circ}{y} = 2 - 0.1t^2$. Twelve values of the output variable $y(t)$ are taken, corresponding to $t = 1, 2, \cdots, 12$. The noise intensity is increased step by step and the optimal models are extracted for each set of data. The combinatorial algorithm is used with a reference function of the third-degree polynomial in t.

$$y = a_0 + a_1 t + a_2 t^2 + a_3 t^3. \tag{3.51}$$

The first minimum-bias criterion most immune to the noise is found by extracting the linear part with the noise immunity coefficient value $\delta_T \approx 2.0$. The second one most immune to noise has data points arranged according to variance. The lowest one has data points arranged as even and odd points.

In all the cases, preliminary extraction of linear parts or trends and widening of data interval with δ_T have significant effect; the noise immune coefficient δ_T is found in the range of 1.5 to 3.0.

3.2 Single and multicriterion analysis

Several qualitative estimates of the degree of noise stability can be obtained analytically by considering just one fixed structure of the model. Suppose the equation $y = X\hat{a}$ is written for the chosen structure. Consider the prediction problem using the prediction criterion

$$\Delta^2(C) = \sum_{i \in C} \frac{(\hat{y} - y)^2_i}{(y - \bar{y})^2_i} \leq 1.0, \tag{3.52}$$

where $\hat{y}$ is the estimated output, $\bar{y}$ is the average value of the output, y is the actual output, and C is the prediction data set. We obtain the estimates of the coefficients $\hat{a}$ using the data set $W = A \cup B$.

$$\hat{a} = (X^T X)^{-1} X(\overset{\circ}{y} + \xi_0) = a^0 + \sigma^2 a_\xi. \tag{3.53}$$

This is the sum of the exact value of the coefficient vector and the added quantity which depends linearly on the noise level (the value of a_ξ is independent of noise variance σ^2). The predictions based on this model have the form of

$$\hat{y}_i = \overset{\circ}{y}_i + \sigma^2 a_\xi^T x_i = \overset{\circ}{y}_i + \alpha^2 s^2 a_\xi^T x_i, \quad i \in C, \tag{3.54}$$

where σ is substituted as $\alpha.s$. The prediction accuracy is obtained as

$$\Delta^2(C) = \sigma^2 \sum_{i \in C} (a_\xi^T x_i)^2 / s^2 = \alpha^2 \sum_{i \in C} (a_\xi^T x_i)^2. \tag{3.55}$$

We obtain the critical noise level α^* on the basis of the condition $\Delta(C) = 1$ as

$$\alpha_p^* = 1 / \sqrt{\sum_{i \in C} (a_\xi^T x_i)^2}. \tag{3.56}$$

Thus, the critical noise level α_p^* depends on both the volume and grouping of the data, and on the realization of the noise. However, this estimate does not coincide with the limiting noise stability of any criterion since, with increase in the noise level, the inductive algorithm chooses another simpler model, which can predict a noise-free signal more accurately. Even this is true with the identification and filtering problems. The analytical study of critical noise levels for identification (α_i^*) and filtering (α_f^*) can be developed.

The combinatorial algorithm is used to obtain the optimal model of complexity by sorting all possible polynomials from the complete basis according to a given external criterion or set of criteria for the given partition of sets. The degree of noise stability of the selection criterion is determined by gradually increasing the noise level and finding the critical value of α in each case.

Example 2. An experiment on estimation of the noise stability of various selection criteria is made with $y^0 = 10 - 0.1t^2$, $t = 1, 2, \cdots, 22$—and a normally distributed white noise with unit variance is obtained for 22 values. This is realized for the output variable y for different noise levels of α with percentage values of 3, 5, 10, 20, 40, 60, 80, 100, 130, 160, 200, 230, 260, 300, 330, 360, 400. Four variants of partitioning of data are used: (i) $N_A + N_B + N_C = 7 + 7 + 8$, (ii) $N_A + N_B + N_C = 8 + 8 + 6$, (iii) $N_A + N_B + N_C = 9 + 9 + 4$, and (iv) $N_A + N_B + N_C = 4 + 4 + 3$ (in all the cases, the points are chosen successively).

The reference function considered here is the third-degree polynomial in t. Combinatorial algorithm is used for sorting all combinations of the structures (15 polynomials of varying structure). The following criteria are tested for their noise stability.

Regularity

$$\Delta^2(B) = \Delta^2(B/A) = \sum_{i \in B} \frac{(\hat{y} - y)_i^2}{(y - \bar{y})_i^2}, \tag{3.57}$$

where $\Delta^2(B/A)$ denotes the model calculated on the set B using the coefficients obtained on A.

Minimum bias

$$\eta_{bs}^2 = \sum_{i \in W} \frac{(\hat{y}^A - \hat{y}^B)_i^2}{(y - \bar{y})_i^2}, \tag{3.58}$$

where the estimates $\hat{y}^A$ and $\hat{y}^B$ correspond to the same model structure but with coefficients calculated on sets A and B, respectively.

Symmetric regularity

$$\begin{aligned} d^2 &= \Delta^2(A/B) + \Delta^2(B/A) \\ &= \|y_A - X_A\hat{a}_B\|^2 + \|y_B - X_B\hat{a}_A\|^2, \end{aligned} \tag{3.59}$$

where parts A and B are used equally.

Here is another form of symmetric criterion:

$$\begin{aligned} S^2 &= \Delta^2(W/A) + \Delta^2(W/B) \\ &= \|y_W - X_W\hat{a}_A\|^2 + \|y_W - X_W\hat{a}_B\|^2, \end{aligned} \tag{3.60}$$

which is an overall estimate on W for the same structure, but with coefficients estimated on different sequences (just as in the criteria η_{bs} and d).

The combined type: ("minimum bias plus prediction")

The noise immunity can be increased significantly by using the following criterion.

$$c3^2 = \eta_{bs}^2 + \Delta^2(C), \tag{3.61}$$

where η_{bs} is one of the realizations of the minimum bias criteria and $\Delta(C)$ is the prediction criterion that computes the sum of square errors using the set C. This criterion requires that a model be unbiased and is also the best predictive method.

A common difficulty with direct application of the combined criteria is the incommensurability of their input quantities. They evaluate different characteristics of the model, such as minimum bias and regularization or extrapolation. Therefore, using them requires choosing and applying weights for each problem.

$$c^2 = \gamma\eta_{bs}^2 + (1-\gamma)k^2, \tag{3.62}$$

where k^2 indicates a stabilizing term of the form $\Delta^2(C)$ or d^2. To obviate selection of weights, one uses a normalized form as

$$c^2 = \frac{\eta_{bs}^2}{\eta_{max}^2} + \frac{k^2}{k_{max}^2} = \bar{\eta}_{bs}^2 + \bar{k}^2, \tag{3.63}$$

where $\bar{\eta}_{bs}$ and $\bar{k}$ are the normalized values, and η_{max} and k_{max} are the maximum values of the criteria of all the models being compared.

All the criteria given above can be used individually as a single criterion choice; at the same time, the combined criteria can be used as two criterion choices. One can also use a stepwise choice; first choose F number of models with the minimum bias criterion, then choose the best model among them using the prediction criterion.

Noise stability of single-criterion selection

The combined criterion $c3$ with its normalized form exhibits the lowest noise stability. Individual criteria operate efficiently with successive application. The regularity criterion is

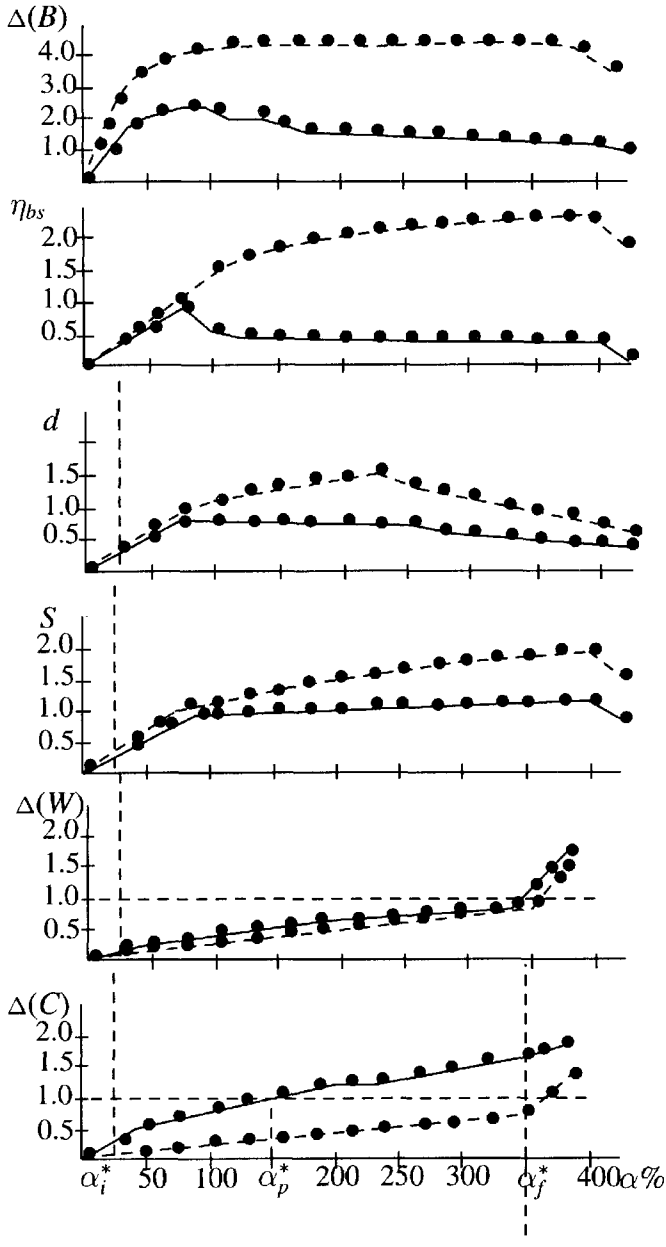

Figure 3.6. Relationship between selection criteria and percentage of the noise level α; the solid line is for models chosen with respect to the minimum of the criterion and the dashed line is for the model $y = 10 - 0.1t^2$

Table 3.2. Values of α_i^* for different selection criteria

Criterion	Different partitions of data			
	4+4+3	7+7+8	8+8+6	9+9+4
$\Delta(B)$	20	360	0	40
η_{bs}	20	60	20	80
$c3$	0	10	10	0
d	20	80	20	80
S	20	130	20	80

the most sensitive criterion to the partitioning of the data; care must be taken in using this criterion for noisy data.

The symmetric criteria (η_{bs}, d, and S) proved to be stable with respect to the partitioning of the data; they virtually have the same noise stability in the case of individual application.

The results of determining the limiting noise stability α_i^* (identification case) for which the original model structure was still acurately reproduced is shown in Table 3.2. Figure 3.6 shows the values of the criteria for the structures obtained on the division $N_A + N_B + N_C = 8+8+6$. The solid curve shows the optimal structures based on the minimum of the criterion and the dashed curve shows the actual structure $y = 10 - 0.1t^2$. The limiting noise levels α_f^* and α_p^* (filtering and prediction cases correspondingly) are considerably higher than the level of the structure for α_i^*.

Noise stability of two-criterion selection

Two-criterion selection is a widely used device in inductive self-organization modeling. Here we use external criteria of a different nature (for example, η_{bs} and $\Delta(B)$) that are related to different parts of data sets (for example, η_{bs} and $\Delta(C)$). There are two types of two-criterion analysis—one is in the form of convolution and another is successive in nature. Sometimes the former may lead to difficulties because of normalization of the criteria. It often turns out that η_{max} exceeds $\Delta(C)$ or d by higher magnitudes so that the bias becomes insignificant and the model is incorrectly chosen by the second criterion. This difficulty is avoided by the successive use of the criteria. The first criterion is used to select F number of models—the best one is chosen using the second criterion. The basic results of successive application of different combinations of two criteria based on the above example are discussed below.

(i) The combination $\Delta(B) \rightarrow \Delta(C)$. The noise stability increases to $\alpha_i^* = 60\%$ (for the criterion $\Delta(B)$, it is 0%).

(ii) The combination $\eta_{bs} \rightarrow \Delta(C)$. The noise stability is very significant; the correct structure of the model is better to a level of $\alpha_i^* = 260\%$ (just for η_{bs} separately, it is 20%).

(iii) The combination $d \rightarrow \Delta(C)$ *and* $S \rightarrow \Delta(C)$. They yield the same results as the preceding pair of criteria.

The use of two-criterion selection of models also increases the level of noise stability in predicting and filtering problems; in case of the combination $\eta_{bs} \rightarrow \Delta(C)$, the noise stability of filtering α_f^* is preserved at the level of 360%, and the noise stability of prediction α_p^* increases from 130% to 400%.

Usually it is impossible to determine the noise stability levels α^* for actual problems on the basis of noisy data because the information regarding the exact structure of the model and the characteristics of the noise is unknown. However, it can be controlled in the course of calculations. The values of the errors $\Delta(A+B)$ and $\Delta(C)$ are noticeably correlated with the ideal estimates of $R(A+B)$ and $R(C)$ (new notations). The difference between $\Delta(A+B)$ and $R(A+B)$ is the denominator term $\sum_{i\in W}(y_i-\bar{y})^2$, that represents the signal variance, similarly between $\Delta(C)$ and $R(C)$. In almost all cases, $\Delta(A+B) > R(A+B)$, and $\Delta(C) > R(C)$. This makes it possible to determine the prediction and filtering satisfactory with the additional conditions $\Delta(C) < 1$ and $\Delta(A+B) < 1$.

There are three ways of testing the operability of an inductive algorithm—with exact data, with a given noise distribution, and with noise distribution peculiar to the class of modeling objects. To verify the results, one has to perform a large number of tests in all the cases. Apparently, it is the only way to solve the problem of definitive verification of the algorithms, and this is done before it is recommended for practical use. Insufficient study of verification might lead to certain difficulties in the practical use of these algorithms. Nonetheless, they can be recommended for solving problems for which other algorithms are unsuited; for example, problems of detailed long-range predictions.

Correct choice of criteria and of the application sequence ensure achievement of qualitative noise stable modeling. Further increase in the noise stability is achieved through the use of multilevel schemes which are described in Chapter 2.

4 ASYMPTOTIC PROPERTIES OF CRITERIA

In this section we present the recent work of Stepashko [120] on asymptotic properties of external criteria for model selection.

The structural identification problem consists of choosing an estimate of the model $\overset{\circ}{y} = a_0^T \overset{\circ}{x}$, $\overset{\circ}{x} = (\overset{\circ}{x}_1, \cdots, \overset{\circ}{x}_{s^0})$, where $\overset{\circ}{y}$ and $\overset{\circ}{x}$ are the output and input vector actions, correspondingly, and a_0 is the actual parameter vector, which is optimal according to a specified combined criterion of minimum-bias plus regularity from a set of various models which contain all possible combinations of $m(\geq s^0)$ input variables. The best regression model is obtained according to the combined criterion from the $2^m - 1$ possible models under the conditions of noisy output $y_i = \overset{\circ}{y}_i + \sigma\xi_i$, $E[\xi_i] = 0$, $E[\xi_i\xi_j] = \sigma^2\delta_{ij}$, where E is the mathematical expectation, δ_{ij} is the kronecker delta, and σ^2 is an unknown finite variance. N_W is the number of points in the given data set.

A simplified version of this problem, which does not restrict the generality of the obtained conclusions about the asymptotic properties of the external criteria, is investigated here.

It is considered as searching an optimal model by successive inclusion of regressors x_s. The set of compared models consists of m various models of the form

$$\hat{y}_s = X_s\hat{a}_s, \quad s = 1, 2, \cdots, m, \tag{3.64}$$

where $X_s = (x_1, \cdots, x_s) = (X_{s-1}, x_s)$, s is the complexity of the model, and where the parameters are estimated by the least squares method as $\hat{a}_s = (X_s^T X_s)^{-1} X_s^T y$.

The structural identification problem is reduced to the determination of the optimal complexity of the model as

$$s^* = \arg\min_{s=\overline{1,m}} c2(y, \hat{y}_s), \tag{3.65}$$

where $c2$ is the combined criterion evaluated by using the actual and estimated values of

the output. The whole data set W is partitioned into two subsets A and B such that

$$X = \begin{pmatrix} X_A \\ X_B \end{pmatrix}, \quad y = \begin{pmatrix} y_A \\ y_B \end{pmatrix}, \quad \text{rank } \{X_A\} = \text{rank } \{X_B\} = m, \quad N = A \cup B. \tag{3.66}$$

The ideal (theoretical) criterion of minimum variance of the forecast J (and of its estimate, the combined criterion) are examined. Two variants of the theoretical criterion that are averaged over the number of points in the sample for which they are calculated, are given below:

$$J(s,N) = \frac{1}{N} E\| \overset{\circ}{y} - X_s \hat{a}_s \|^2, \tag{3.67}$$

$$J_B(s,N_A,N_B) = \frac{1}{N_B} E\| \overset{\circ}{y}_B - X_{Bs} \hat{a}_{As} \|^2, \tag{3.68}$$

and the external criteria; regularity and minimum-bias

$$\Delta_B(s,N_A,N_B) = \frac{1}{N_B} \|y_B - X_{Bs} \hat{a}_{As}\|^2, \tag{3.69}$$

$$\eta_{bs}(s,N) = \frac{1}{N} \|X_s \hat{a}_{As} - X_s \hat{a}_{Bs}\|^2. \tag{3.70}$$

An optimal smoothing model corresponds to the solution of the problem with respect to the minimum of $J(s,N)$, and an optimal forecasting model corresponds to the minimum of $J_B(s,N_A,N_B)$. The external criteria Δ_B and η_{bs} are their estimates. To investigate the behavior of the theoretical as well as external criteria as $N \to \infty$, it is assumed that the matrix X satisfies the strong regularity condition;

$$\lim_{N\to\infty} \frac{1}{N} X_N^T X_N = \bar{H}, \tag{3.71}$$

where H is a nonsingular finite $m \times m$ matrix.

The characteristic results of the solution of the structural problem according to the given criteria for $\sigma^2 = \text{var}, N = \text{const}$, and $\sigma^2 = \text{const}, N \to \infty$ are compared below for the adopted assumptions.

4.1 Noise immunity of modeling on a finite sample

When solving the above structural identification problem, one has to estimate the parameters for each set of regressors as $s = 1, 2, \cdots, m$. This can be done conveniently by the recursive algorithm presented in the preceding chapter for constructing the partial models of gradually increasing complexity, beginning with a single argument ("method of bordering"). (Refer to the section on "Recursive scheme for faster combinatorial sorting" in chapter 2.)

For quick reference, we briefly give the algorithm here. X^TX and X^Ty are denoted as H and g, correspondingly, and H_s, g_s, and $\hat{a}_s$ are represented in the form of

$$H_s = \begin{bmatrix} H_{s-1} & h_s \\ h_s^T & \vartheta_s \end{bmatrix}, \quad g_s = \begin{bmatrix} g_{s-1} \\ \gamma_s \end{bmatrix}, \quad \hat{a}_s = \begin{bmatrix} \hat{a}_{s-1}^* \\ \hat{\alpha}_s \end{bmatrix}, \tag{3.72}$$

where $h_s = X_{s-1}^T x_s$, $\vartheta_s = x_s^T x_s$, and $\gamma_s = x_s^T y$.

The following recursive algorithm is valid for the calculation of H_s^{-1} and $\hat{\alpha}_s$;

$$\beta_s = 1/(\vartheta_s - h_s^T c_s), \; c_s = H_{s-1}^{-1} h_s, \; H_0^{-1} \triangle = 0, \tag{3.73}$$

$$H_s^{-1} = \left[\begin{array}{c|c} H_{s-1} + \beta_s c_s c_s^T & -\beta_s c_s \\ \hline -\beta_s c_s^T & \beta_s \end{array}\right], \tag{3.74}$$

$$\hat{a}_s = \left[\begin{array}{c} \hat{a}_{s-1} - \hat{\alpha}_s c_s \\ \hline \beta_s(\gamma_s - g_{s-1}^T c_s) \end{array}\right] = \left[\begin{array}{c} \hat{a}_{s-1}^* \\ \hat{\alpha}_s \end{array}\right], \quad \hat{a}_0 \triangle = 0. \tag{3.75}$$

This algorithm can be used directly for the criterion $J(s, N)$, while for $J_B(s, N_A, N_B)$ and Δ_B, it is applied using the subset A (using the index A). For η_{bs}, the quantities are computed on both the subsets A and B.

Properties of the criteria $J(s, N)$ and $J_B(s, N_A, N_B)$

These criteria are reduced to the form;

$$J(s, N) = J^0(s) + J^*(s) = \frac{1}{N} \| \overset{\circ}{y} - X_s a_s \|^2 + \frac{\sigma^2}{N} s, \tag{3.76}$$

$$J_B(s, N_A, N_B) = J_B^0(s) + J_B^*(s) = \frac{1}{N} \| \overset{\circ}{y}_B - X_{Bs} a_{As} \|^2 + \frac{\sigma^2}{N_B} \operatorname{tr} (H_{As}^{-1} H_{Bs}). \tag{3.77}$$

The parameters a_s and a_{As} are estimated either by using the least-squares technique or by using the above recursive algorithm and substituting $\overset{\circ}{y}$ for y; i.e., $a_s = E[\hat{a}_s]$, $a_{As} = E[\hat{a}_{As}]$. Here, $J^0(s)$ and $J_B^0(s)$ characterize the structural bias, while $J^*(s)$ and $J_B^*(s)$ reflect the effect of noise. Obviously, $a_{s^0} = a_{As^0} = a_0$, so that $J^0(s^0) = J_B^0(s^0) = 0$.

Let us examine $J(s)$; one can obtain

$$J^0(s) = J^0(s-1) - \frac{1}{N} \alpha_s^2 / \beta_s = J^0(s-1) - \frac{1}{N} \beta_s (\gamma_s - g_{s-1}^T c_s)^2, \tag{3.78}$$

where $\beta_s = x_s^T (I - X_{s-1}(X_{s-1}^T X_{s-1})^{-1} X_{s-1}) x_s$ is positive and equal to the ratio of the determinants of the matrices H_{s-1} and H_s. Thus, $J^0(s)$ is a monotonically decreasing function of the complexity s so that in view of $J^*(s) = J^0(s)$ the function $J(s, n)$ for $\sigma^2 > 0$ always has a single minimum at the point $s^* \leq s^0$. As σ^2 increases, the complexity s^* decreases. A simpler model becomes J-optimal. This property is named "noise immunity"; i.e., the error in reconstructing the noise-free vector $\overset{\circ}{y}$ decreases due to the simplification of the model. This means that the model of s^0 loses its J-optimality for the variance $\sigma^2 > \sigma_{cr}^2(s^0) = \alpha_{s^0}^2 / \beta_{s^0}$. In general, for arbitrary complexity s, the transition from $s^* = s$ to $s^* = s - 1$ occurs for $\sigma^2 > \sigma_{cr}^2(s)$, where

$$\sigma_{cr}^2 = N(J^0(s-1) - J^0(s)) = \alpha_s^2 / \beta_s. \tag{3.79}$$

α_s is the coefficient of the sth regressor and 'cr' indicates the critical value.

Let us examine $J_B(s)$; one can obtain $J_B^0(s)$ as

$$\begin{aligned} J_B^0(s) = J_B^0(s-1) &- 2\alpha_{As}(\overset{\circ}{y}_B - X_{B,s-1} a_{A,s-1})^T (x_{Bs} - X_{B,s-1} c_{As}) \\ &+ \alpha_{As}^2 \| x_{Bs} - X_{B,s-1} c_{As} \|^2. \end{aligned} \tag{3.80}$$

Here, one cannot guarantee that the increment will be negative for every s (except when the regressors are orthogonal), so that in the general case the decrease from $J_B^0(1)$ to $J_B^0(s^0) = 0$

is not monotonic. The monotonicity of the dependence on s is preserved for the noise component $J_B^*(s)$. This fact is proven below on the basis of the recursive algorithm, with $h_{Bs} = X_{B,s-1}^T x_{Bs}$ and $\vartheta_{Bs} = x_{Bs}^T x_{Bs}$ taken into account.

$$J_B^*(s) = \frac{\sigma^2}{N_B} \text{ tr } (H_{As}^{-1} H_{Bs}) = J_B^*(s-1) + \frac{\sigma^2}{N_B} \beta_{As} (h_{As}^T H_{A,s-1}^{-1} H_{B,s-1} H_{A,s-1}^{-1} h_{As} -$$

$$-2h_{As}^{-1} H_{A,s-1} h_{Bs} + \vartheta_{Bs} = J_B^*(s-1) + \frac{\sigma^2}{N_B} \beta_{As} \| x_{Bs} - X_{B,s-1} H_{A,s-1}^{-1} h_{As} \|^2 \qquad (3.81)$$

The increment change in this equation is estimated, having determined the extremum of the vector argument $\varphi(z) = z^T H_{B,s-1} z - 2z^T h_{Bs} + \vartheta_{Bs}$, where $z \triangle = H_{A,s-1}^T h_{As}$. If we differentiate φ with respect to z and equate it to zero, we obtain $z_0 = H_{B,s-1}^{-1} h_{Bs}$. Since $H_{B,s-1}$ is a positive-definite matrix, $\varphi(z)$ for $z = z_0$ has a positive minimum; $\varphi(z_0) = \vartheta_{Bs} - h_{Bs}^T H_{B,s-1}^{-1} h_{Bs} = 1/\beta_{Bs}$. Thus, the minimal increment change of the trace in the equation of $J_B(s, N_A, N_B)$ equals β_{As}/β_{Bs}, and is attained when the relation

$$H_{A,s-1}^{-1} h_{As} = H_{B,s-1}^{-1} h_{Bs} \qquad (3.82)$$

is satisfied. In particular, this relation is satisfied when $H_{Bs} = \lambda^2 H_{As}$ or $X_B = \lambda X_A$, where λ is an arbitrary constant. Here $\text{tr}(H_{As}^{-1} H_{Bs}) = \text{tr } (H_{A,s-1}^{-1} H_{B,s-1}) + \lambda^2$, so that the rate of growth of the trace is proportional to the value of λ^2 as s is to one; i.e., even if the relation is satisfied, the examined increment may be arbitrary. Thus, as the component $J_B^0(s)$ in the equation of $J_B(s, N_A, N_B)$ decreases, and $J_B^*(s)$ increases monotonically as the complexity of the model s increases, the minimum of $J_B(s, N_A, N_B)$ is possible only for $s^* \leq s^0$. Qualitatively the behavior of the criteria $J(s, N)$, $J_B(s, N_A, N_B)$ is the same. Moreover, for a model of complexity s^0, one can establish the threshold of the J-optimality loss. For this, it is considered that $\overset{\circ}{y}_B = X_{Bs^0} a_0 = X_{B,s-1} a_{s^0-1}^0 + x_{Bs^0} \alpha_0$, where $a_0 = (a_{s^0-1}^{0T} \alpha_0)^T$. Furthermore, according to the recursive relation, $a_{s^0-1}^0 = a_{A,s^0-1} - \alpha_{As^0} c_{As^0}$, $\alpha_{As^0} \equiv \alpha_0$, and, consequently,

$$\overset{\circ}{y}_B - X_{B,s^0-1} a_{A,s^0-1} = \alpha_{As^0} (X_{Bs^0} - X_{B,s^0-1} c_{As^0}),$$

$$J_B^0(s^0) = J_B^0(s^0 - 1) - \alpha_{As^0} \| X_{Bs^0} - X_{B,s^0-1} c_{As^0} \|^2. \qquad (3.83)$$

From the conditions $J_B(s^0) = J_B(s^0 - 1)$, $J_B^0(s^0) = 0$, considering the equations for $J_B^0(s^0)$ and $J_B^*(s)$, we obtain

$$\sigma_{cr}^2(s^0) = \alpha_{As^0}^2 / \beta_{As^0}. \qquad (3.84)$$

Thus, the condition for losing the J-optimality for a model of actual complexity s^0 (with an unbiased structure) turns out to be completely identical in problems of search for optimal smoothing and prediction models. This property is determined solely by the properties of subset A. This result can also be obtained by using different transformations. It is noted that $\sigma_{cr}^2(s^0)$ does not depend on the number of points N_B but depends implicitly on N_A.

Properties of the external criteria

The mathematical expectations of the regularity ($\Delta_B(s, N_A, N_B)$) and minimum-bias ($\eta_{bs}(s, N)$) criteria are equal to

$$\bar{\Delta}_B(s) = \frac{1}{N_B} \| \overset{\circ}{y}_B - X_{Bs} a_{As} \|^2 + \frac{\sigma^2}{N_B} (N_B + \text{ tr } (H_{As}^{-1} H_{Bs})), \qquad (3.85)$$

$$\bar{\eta}_{bs}(s) = \frac{1}{N}\|X_s a_{As} - X_s a_{Bs}\|^2 + \frac{\sigma^2}{N}(2s + \text{ tr } (H_{As}^{-1}H_{Bs} + H_{Bs}^{-1}H_{Bs})). \tag{3.86}$$

Comparison of the equation Δ_B with its expectation $\bar{\Delta}_B$ yields

$$\bar{\Delta}_B(s) = J_B(s, N_A, N_B) + \sigma^2. \tag{3.87}$$

This means that the minimum of the regularity criterion Δ_B gives an unbiased J-optimal model, since the minima of $J_B(s)$ and $\bar{\Delta}_B(s)$ always correspond to the same optimal complexity s^*. Hence, the regularity criterion has the necessary property of noise immunity and other properties of the criterion $J_B(s, N_A, N_B)$; for example, the actual structure is optimal for $\sigma^2 < \sigma_{cr}^2(s^0)$.

The criterion $\bar{\eta}_{bs}(s)$ was worked out in detail in the work [118]; it was shown that, if the condition $(\tilde{X}_A^T\tilde{X}_A)^{-1}\tilde{X}_A^T \overset{\circ}{X}_A \neq (\tilde{X}_B^T\tilde{X}_B)^{-1}\tilde{X}_B^T \overset{\circ}{X}_B$ is satisfied, then it has a single global minimum. Biased values of its model structures decrease from $\eta_{bs}(1)$ to $\eta_{bs}(s^0) = 0$ (possibly nonmonotonically) while the noise component increases monotonically. Consequently, the minimum-bias criterion $\eta_{bs}(s)$ has the noise immune property.

4.2 Asymptotic properties of the external criteria

As $N = N_A \cup N_B$, one has to examine the case of $N \to \infty$ as well as its variants: $N_A \to \infty$, $N_B \to \infty$, and $N_A, N_B \to \infty$. In addition to the assumption that $\lim_{N\to\infty} \frac{1}{N}X_N^T X_N = \bar{H}$, let us assume that the matrices X_A and X_B are regular and are formed independently.

$$\lim_{N_A\to\infty} \frac{1}{N_A} X_{AN}^T X_{AN} = \bar{H}_A, \quad \lim_{N_B\to\infty} \frac{1}{N_B} X_{BN}^T X_{BN} = \bar{H}_B, \tag{3.88}$$

where $\bar{H}_A$, $\bar{H}_B$ are finite nonsingular matrices. The limits in the above equations exist for individual element of the matrices and for each of their blocks. Thus,

$$\lim_{N\to\infty} \frac{1}{N} x_{iN}^T x_{jN} = \bar{h}_{ij}, \quad i,j = 1, 2, \cdots, m,$$

$$\lim_{N\to\infty} \frac{1}{N} X_{sN}^T X_{sN} = \bar{H}_s = \begin{bmatrix} \bar{h}_{11} & \cdots & \bar{h}_{1s} \\ & \cdots & \\ \bar{h}_{s1} & \cdots & \bar{h}_{ss} \end{bmatrix}, \tag{3.89}$$

where $\bar{h}_{ij}$ and $\bar{H}_s$ are the finite numbers and matrices, respectively.

Taking the actual model ($\overset{\circ}{y} = a_0^T \overset{\circ}{x}$, $\overset{\circ}{x} = (\overset{\circ}{x}_1, \cdots, \overset{\circ}{x}_{s^0})^T$) of the object into account, one can write the following relation in matrix form as

$$\overset{\circ}{y}_N = \overset{\circ}{X}_N a_0 = \overset{\circ}{X}_N a_0 + \tilde{X}_N \emptyset \triangle = X_N a^*, \tag{3.90}$$

where $\emptyset$ is the zero or empty vector of dimension $(m - s^0)$, and $a^* = (a_0^T, \emptyset^T)^T$ is the finite vector of the actual parameters.

The assumption of limiting transition as $N \to \infty$ implies the existence of the finite limits as

$$\lim_{N\to\infty} \frac{1}{N} \overset{\circ}{y}_N^T \overset{\circ}{y}_N = a^{*T} \bar{H} a^* \triangle = \bar{\mu}, \tag{3.91}$$

$$\lim_{N\to\infty} X_N^T \overset{\circ}{y}_N = \bar{H} a^* \triangle = \bar{g}. \tag{3.92}$$

Similarly, the existence of $\bar{\mu}_A, \bar{\mu}_B, \bar{g}_A$, and $\bar{g}_B$ for other assumptions is $N_A \to \infty$ and $N_B \to \infty$.

In asymptotic problems of control theory, the condition of mean-square summation (integrability) of functions is usually a common assumption. From the well-known relation between the elements of the normal matrix X^TX: $x_i^Tx_j < (x_i^Tx_i + x_j^Tx_j)/2$, the mean square summation of observations of all the individually taken regressors is written as

$$\lim_{N\to\infty} \frac{1}{N} x_{Ns}^T x_{Ns} = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} x_{is}^2 = \bar{\vartheta}_s, \quad s = 1, 2, \cdots, m, \tag{3.93}$$

based on the convergent sequences for all $\frac{1}{N} x_{Ni}^T x_{Nj}$, $i \neq j$; $i, j = 1, 2, \cdots, m$.

Properties of the criteria $J(s, N)$ and $J_B(s, N_A, N_B)$

The structural components of these criteria are represented in the form

$$J^0(s, N) = \frac{1}{N} \left(\overset{\circ}{y}_N^T \overset{\circ}{y}_N - a_{sN}^T X_{sN} \overset{\circ}{y}_N \right), \tag{3.94}$$

$$J_B^0(s, N_A, N_B) = \frac{1}{N_B} (\overset{\circ}{y}_{N_B}^T \overset{\circ}{y}_{N_B} - 2a_{sN_A}^T X_{sN_B}^T \overset{\circ}{y}_{N_B} + a_{sN_A}^T X_{sN_B}^T X_{sN_B} a_{sN_A}). \tag{3.95}$$

The convergence of the parameters in these equations is established on the basis of the limiting transitions given as

$$\lim_{N\to\infty} a_{sN} = \bar{H}_s^{-1} \bar{g}_s \triangle = \bar{a}_s, \quad \lim_{N_A\to\infty} a_{sN_A} = \bar{H}_{As}^{-1} \bar{g}_{As} \triangle = \bar{a}_{As}. \tag{3.96}$$

Taking $J^*(s, N) = \sigma^2 s/N$ into account, we obtain

$$\tilde{J}(s) = \lim_{N\to\infty} J(s, N) = \lim_{N\to\infty} J^0(s, N) = \bar{\mu} - \bar{g}_s^T \bar{H}_s^{-1} \bar{g}_s. \tag{3.97}$$

It is obvious that for $s = 1, 2, \cdots, s^0$, $\tilde{J}(s)$ decreases monotonically, while for $s \geq s^0$ the quantity $a_s = a^*$, so that $J(s) = 0$. Thus, if the quantity $J(s, N)$ for $\sigma^2 > 0$, and $N > \infty$ has a minimum of $s^* \leq s^0$, then there is a compromise between its structural $J^0(s, N)$ and noise $J^*(s, N)$ components. As $N \to \infty$, the component $J^*(s, N)$ disappears (an increase in the amount of information removes the uncertainty) and the minimum of $\tilde{J}(s)$ corresponds to the actual unbiased model structure.

In the case of $J_B(s, N_A, N_B) = J_B^0(s, N_A, N_B) + J_B^*(s, N_A, N_B)$, one has to consider the limiting transitions for $N_A \to \infty$ and $N_B \to \infty$. Finite values are obtained for the structural component $J_B^0(s, N_A, N_B)$ by taking into account its convergence property of parameters.

$$\tilde{J}_B^0(s, N_B) = \lim_{N_A\to\infty} J_B^0(s, N_A, N_B) = \frac{1}{N_B} (\mu_B - 2\bar{a}_{As}^T g_{Bs} + \bar{a}_{As}^T H_{Bs} \bar{a}_{As}), \tag{3.98}$$

$$J_B^0(s, N_A) = \lim_{N_B\to\infty} J_B^0(s, N_A, N_B) = \bar{\mu}_B - 2a_{sN_A}^T \bar{g}_{Bs} + a_{sN_A}^T \bar{H}_{Bs} a_{sN_A}, \tag{3.99}$$

where $\mu_B = \overset{\circ}{y}_{N_B}^T \overset{\circ}{y}_{N_B}$, $\bar{\mu}_B = \lim_{N_B\to\infty} \mu_B$. Convergence of the noise component $J_B^*(s, N_A, N_B)$ is determined by the asymptotics of the trace $\mathrm{tr}(H_{As}^{-1} H_{Bs})$ in the equation given for $J_B(s, N_A, N_B)$.

$$\begin{aligned} \tilde{J}_B^*(s, N_B) &= \lim_{N_A\to\infty} J_B^*(s, N_A, N_B) \\ &= \frac{\sigma^2}{N_B} \lim_{N_A\to\infty} \frac{1}{N_A} \mathrm{tr} \left(\frac{1}{N_A} H_{As} \right)^{-1} H_{Bs} = 0, \end{aligned} \tag{3.100}$$

$$J_B^*(s, N_A) = \lim_{N_B \to \infty} J_B^*(s, N_A, N_B) = \sigma^2 \text{ tr } (H_{N_A s}^{-1} \bar{H}_{Bs}) < \infty. \tag{3.101}$$

Thus, as $N_A \to \infty$, the noise component disappears and the properties of $\tilde{J}_B(s, N_B)$ become analogous to the properties of $\tilde{J}(s)$. However, as $N_B \to \infty$, the uncertainty caused by the parameter estimates a_{As} from a finite sample is not removed, and the criterion $\tilde{J}_B(s, N_A) = \tilde{J}_B^0(s, N_A) + \tilde{J}_B^*(s, N_A)$, where N_A is finite has the same properties as the equation given for $J_B(s, N_A, N_B)$. This means that if the parameters of the model obtained on a finite sample A and the model is applied on a infinite sample B, then the minimal variance of the forecast, in general, is achieved by a model according to noise immunity (J-optimal), rather than to an unbiased model which depends on σ^2.

If $N_A \to \infty$ and $N_B \to \infty$, we obtain from the above equations

$$\tilde{J}_B^*(s) = \lim_{N_B \to \infty} \lim_{N_A \to \infty} J_B^*(s, N_A, N_B) = 0$$

$$\begin{aligned} \tilde{J}_B(s) = \tilde{J}_B^0(s) &= \lim_{N_B \to \infty} \lim_{N_A \to \infty} J_B^0(s, N_A, N_B) \\ &= \bar{\mu}_B - 2\bar{a}_{As}\bar{g}_{sB} + \bar{a}_{As}\bar{H}_{Bs}\bar{a}_{As} < \infty, \end{aligned} \tag{3.102}$$

where $\tilde{J}_B(s|s \geq s^0) = 0$, as it is for $\tilde{J}_B(s, N_B)$.

This follows that the criteria $\tilde{J}(s)$ for $N \to \infty$, $\tilde{J}(s, N_B)$ for $N_A \to \infty$, and $\tilde{J}_B(s)$ for $N_A, N_B \to \infty$ converge for any s. Their minima equal to zero which corresponds to the actual model; $s^* = s^0$. This result is because of the consistency of the least squares estimates of the parameters of unbiased structures $s \geq s^0$ and the convergence of these estimates for the biased structures $s < s^0$. This established regularity of the asymptotic behavior of the criteria $J(s, N)$ and $J_B(s, N_A, N_B)$ is called "consistency property." As the sample length increases, the actual model which corresponds to the minimum or zero value of the criterion, becomes the limit of the optimal smoothing and forecasting model. Because there is no appearance of the expression concerning σ^2, the indicated property is valid for any variance. This means that the critical values of the variances $\sigma_{cr}^2(s)$ and $\sigma_{cr}^2(s^0)$ (the expressions given above) should approach infinity as $N \to \infty$ and $N_A \to \infty$, accordingly. In view of the established convergence, the concerned parameters turn out to be finite: $\bar{\alpha}_s = \lim_{N \to \infty} \alpha_{sN} < \infty$, as well as $\bar{\alpha}_{As^0} < \infty$. At the same time, for any s, the given equation for β in the recursive algorithm is obtained as

$$\begin{aligned} \bar{\beta}_s &= \lim_{N \to \infty} \beta_{sN} \\ &= \lim_{N \to \infty} \left[1 / \left\{ N \left(\frac{1}{N} \vartheta_{sN} - \frac{1}{N} h_{sN}^T \left(\frac{1}{N} H_{s-1,N} \right)^{-1} \frac{1}{N} h_{sN} \right) \right\} \right] = 0. \end{aligned} \tag{3.103}$$

Analogously, the relation $\beta_{As^0} = 0$ can be established by virtue of the limiting transitions that proves the assertion that $\sigma_{cr}^2(s, N) \to \infty$ for $\sigma_{cr}^2(s^0, N_A) \to \infty$ and $N_A \to \infty$.

It is obvious that any estimates of the criteria $J(s, N)$ and $J_B(s, N_A, N_B)$ used in practice must have the "consistency property."

Properties of the external criteria

The convergence of the regularity criterion $\bar{\Delta}_B(s) = \bar{\Delta}_B(s, N_A, N_B)$ for the cases $N_A \to \infty$, $N_B \to \infty$ and $N_A, N_B \to \infty$, and any s follows from the relation $\bar{\Delta}_B(s) = J_B(s, N_A, N_B) + \sigma^2$. The first and third cases are of interest with regard to "consistency property." One can

obtain any s by taking into account the obtained finite values on $\tilde{J}_B^0(s, N_B)$ and $\tilde{J}_B^0(s)$.

$$\tilde{\Delta}_B(s, N_B) = \lim_{N_A \to \infty} \bar{\Delta}_B(s, N_A, N_B) = \tilde{J}_B^0(s, N_B) + \sigma^2, \tag{3.104}$$

$$\tilde{\Delta}_B(s) = \lim_{N_B \to \infty} \lim_{N_A \to \infty} \bar{\Delta}_B(s, N_A, N_B) = \tilde{J}_B^0(s) + \sigma^2. \tag{3.105}$$

For the limiting transitions considered, taking into account the properties of the quantities $\tilde{J}_B^0(s, N_B), \tilde{J}_B^0(s)$, the minimum of the criterion $\bar{\Delta}_B(s, N_A, N_B)$ is

$$\min \tilde{\Delta}_B(s, N_B) = \min \tilde{\Delta}_B(s) = \sigma^2. \tag{3.106}$$

Thus, the minimum of the mathematical expectation of the criterion $\Delta_B(s, N_A, N_B)$ is an asymptotically unbiased estimate with the unknown noise variance and corresponds in the limit to the actual model, which has the "consistency property."

The asymptotic properties of the consistency criterion $\eta_{bs}(s) = \eta_{bs}(s, N_A, N_B)$ are to be determined by performing on the established relation $\bar{\eta}_{bs}(s)$ a double limiting transition such as $N_A \to \infty$ and $N_B \to \infty$. It is convenient to adopt the commonly applied condition $N_A = N_B$; then $N = 2N_A = 2N_B$. First, the deterministic component is considered and represented as

$$\eta_{bs}^0(s, N_A, N_B) = \frac{1}{N}(a_{As} - a_{Bs})^T H_s(a_{As} - a_{Bs}). \tag{3.107}$$

Taking the limits, we have

$$\tilde{\eta}_{bs}^0(s) = \lim_{N_A \to \infty} \lim_{N_B \to \infty} \eta_{bs}^0(s, N_A, N_B) = (\bar{a}_{As} - \bar{a}_{Bs})^T \bar{H}_s(\bar{a}_{As} - \bar{a}_{Bs}) < \infty, \tag{3.108}$$

where $\tilde{\eta}_{bs}^0(s | s \geq s^0) = 0$.

Second, the noise component (from the equation $\bar{\eta}_{bs}(s)$) is considered for performing the limiting transformations. This is analogous to the relation we got for $\tilde{J}_B^*(s, N_B)$;

$$\tilde{\eta}_{bs}^*(s) = \lim_{N_B \to \infty} \lim_{N_A \to \infty} \eta_{bs}^*(s, N_A, N_B) = 0. \tag{3.109}$$

This means that the mathematical expectation of the minimum-bias criterion also converges for any s and has the "consistency property." Moreover, this criterion can be viewed as the asymptotically unbiased estimate of the values of $\tilde{J}(s)$. Hence, using the minimum-bias criterion is preferred in searching for the optimal smoothing model, while it is better to use the regularity criterion in the search for an optimal forecasting model. The regularity and minimum-bias criteria, in addition to the noise immunity, also has the "consistency property" that permits the applicability of the inductive algorithms for complex problem-solving by using small as well as large samples of data observations. The reader can also refer to works on analogous results of Dyshin [13] [14] and Aksenova [2] for further study.

4.3 Calculation of locus of the minima

This section describes a procedure for calculating the locus of the minima (LM) for ideal criteria [121]. This is important because extrapolation of LM allows one to detect a true signal from the noisy data.

In the course of a numerical study of simulating properties of noise immunity, the following computational experiment is considered. An actual model of an object is given by $\overset{\circ}{y} = a_0^T \overset{\circ}{x}$, where input vector $\overset{\circ}{x} = (\overset{\circ}{x}_1, \overset{\circ}{x}_2, \cdots, \overset{\circ}{x}_{s^0})$. Based on this the noisy observations of the

true output at N points are calculated; $y_i = \overset{\circ}{y}_i + \sigma\xi_i$, where σ is an arbitrarily selected variance of the noise and ξ_i are known realizations of uncorrelated noise with $E[\xi_i] = 0$, $E[\xi_i^2] = 1$. In the experiment, the number of points N, the realization of ξ, and the variance σ may vary. Moreover, it is assumed that there is an extended vector of input variables $x = (\overset{\circ}{x}, \tilde{x}^T)$ with the dimensionality of $m \geq s^0$. Models of different combinations of m input variables are compared; $\hat{y} = \hat{a}^T x$. This corresponds to an application of the combinatorial algorithm for modeling the actual signal $\overset{\circ}{y} = (\overset{\circ}{y}_1, \overset{\circ}{y}_2, \cdots, \overset{\circ}{y}_N)^T$ by comparing all the models of the above form or in the matrix notation

$$\hat{y}_s = X_s \hat{a}_s, \quad s = \overline{1, 2^m - 1}, \tag{3.110}$$

where parameters $\hat{a}$ are the least-squares estimates

$$\hat{a}_s = (X_s^T X_s)^{-1} X_s^T y, \tag{3.111}$$

calculated using the noisy output vector.

The aim of this computational experiment is to compare the efficiency of various criteria for selecting models in relation to the ideal criterion

$$\Delta_N(s) = \| \overset{\circ}{y} - \hat{y}_s \|^2 = \| \overset{\circ}{y} - X_s \hat{a}_s \|^2, \tag{3.112}$$

that gives a measure of precision in recovering the actual signal $\overset{\circ}{y}$ by means of the model $\hat{y}_s$ obtained using the noisy data for each s. By varying s one can obtain optimal value s^0 (optimal structure) and the corresponding minimum value of the criterion $\Delta_N(s^0)$ for different ξ and σ. It is convenient to pose the above problem according to the complexity of the models as per their number of input variables. Evidently, there are C_m^1 models of complexity for $s = 1$, C_m^2 models of complexity for $s = 2$, and so on to $C_m^m = 1$ models of complexity m. The minimum value of $\Delta_N(s)$ is determined for each s. It will then constitute the characteristic $\Delta_N(s)$ and $\Delta_N(s) \equiv \Delta_N(s^0)$, so that the optimal value s^0 corresponds to a model with the minimal variance.

Let us assume that the values of $\Delta_N(s)$ are obtained for $s = 1, 2, \cdots, m$ by successive inclusion of regressors x_s and that the properties of the functional J as the mathematical expectation of the $\Delta_N(s)$ is

$$J(s) = E\| \overset{\circ}{y} - X_s \hat{a}_s \|^2. \tag{3.113}$$

It is shown before that

$$J(s) = J^0(s) + J^*(s) = E\| \overset{\circ}{y} - X_s a_s \|^2 + \sigma^2 s, \tag{3.114}$$

where $a_s = (X_s^T X_s)^{-1} X_s \overset{\circ}{y} \equiv E(\hat{a}_s)$.

$J^0(s)$ is a monotonically decreasing function with complexity s, where $J(s | s \geq s^0) = 0$; $J(s)$ has a unique minimum for certain optimal complexity $s^* \leq s^0$. This minimum shifts to the left as σ increases (refer to Figure 5.3). $\Delta_N(s)$ possesses the same properties, as shown before (for regularity criterion).

$$\Delta_N(s) = \Delta_N^0(s) + \Delta_N^*(s) = J^0(s) + \sigma^2 \hat{a}_{\xi s} X_s^T X_s \hat{a}_{\xi s}, \tag{3.115}$$

where $\hat{a}_{\xi s} = (X_s^T X_s)^{-1} X_s^T \xi$, and $\Delta_N^*(s)$ is a monotonically increasing function.

Optimal intervals of each model can be calculated by using systematic algorithms.

The notion of the "locus of the minima" of a criterion is defined as a function $J_{min}(s)$ or $\Delta_{min}(s)$ whose value corresponds to the critical value $\sigma_{cr}(s)$ for which the model with complexity of $s - 1$ becomes optimal instead of the model with s.

Algorithm for calculating the LM of $J_{min}(s)$

Let us assume that the regressors are included in the model in order of the correlation coefficients between the regressors and the actual output.

$$r_{x_s \overset{\circ}{y}} = \frac{\sum_{j=1}^{N}(x_{js} - \bar{x}_s)(\overset{\circ}{y}_j - \overset{\circ}{\bar{y}})}{\left(\sum_{j=1}^{N}(x_{js} - \bar{x}_s)^2\right)^{\frac{1}{2}} \left(\sum_{j=1}^{N}(\overset{\circ}{y}_j - \overset{\circ}{\bar{y}})^2\right)^{\frac{1}{2}}}, \quad s = \overline{1,m}, \tag{3.116}$$

and the regressors are ranked in decreasing order of the correlation coefficients.

The algorithm consists of the following steps:

1. calculating the matrices X^TX, $X^T\overset{\circ}{y}$ for the Gaussian normal equations for the full model;
2. computing the least-squares estimates of the parameters a_s using the equation $X_s^TX_sa_s = X_s^T\overset{\circ}{y}$;
3. determining the quadratic error of the estimator $\overset{\circ}{y}$ using the least-squares method as

$$J^0(s) = \overset{\circ}{y}^T\overset{\circ}{y} - a_s^TX_s^T\overset{\circ}{y}; \tag{3.117}$$

4. calculating the estimate of a_{s+1};
5. determining the $J^0(s+1)$;
6. calculating the decrease in error due to the inclusion of regressors by one:

$$\delta_{s+1}^2 = J^0(s) - J^0(s+1),\ J^0(s^0) = 0; \tag{3.118}$$

7. determining the ordinate of LM of the ideal criterion at point s: $J_{min}(s) = J^0(s) + s\delta_{s+1}^2$;
8. increasing the complexity by one unit. Return to step 4.

Note that these calculations can also be conducted by using the recursive algorithm.

Algorithm for calculating the LM of $\Delta_{min}(s)$ for an individual realization of the noise vector

As in the above algorithm, the regressors are assumed to be ranked in decreasing order according to their correlation coefficients.

The algorithm consists of the following steps:

1. calculating the matrices X^TX, X^Ty, $X^T\xi$;
2. determining the estimator a_s and the errors of this estimator $\hat{a}_{\xi s}$ due to the presence of noise. Here a_s is the solution of the equation $X_s^TX_sa_s = X_s^T\overset{\circ}{y}$, and $\hat{a}_{\xi s}$ is the solution of the equation $X_s^TX_s\hat{a}_{\xi s} = X_s^T\xi$;
3. calculating $J^0(s)$ using the formula $J^0(s) = \overset{\circ}{y}^T\overset{\circ}{y} - a_s^TX_s^T\overset{\circ}{y}$ as well as the quantity

$$\lambda_s^2 = \hat{\xi}_s^T\hat{\xi}_s = \hat{a}_{\xi s}^TX_s^TX_s\hat{a}_{\xi s}, \tag{3.119}$$

   as it is given in the equation $\Delta_N(s) = \Delta_N^0(s) + \Delta_N^*(s) = J^0(s) + \sigma^2\hat{a}_{\xi s}^TX_s^TX_s\hat{a}_{\xi s}$;
4. determining the estimators of a_{s+1}, and $\hat{a}_{\xi,s+1}^T$;

5. calculating the quantity $J^0(s+1)$ and $\lambda^2_{s+1} = \hat{\xi}^T_{s+1}\hat{\xi}_{s+1} = \hat{a}^T_{\xi,s+1}X^T_{s+1}X_{s+1}\hat{a}_{\xi,s+1}$;
6. calculating the increments of δ^2_{s+1} and the amount of increment of the random component λ of the criterion $\Delta_N(s)$ with complexity:

$$\delta^2_{\xi_{s+1}} = \lambda^2_{s+1} - \lambda^2_s; \tag{3.120}$$

7. obtaining the ordinate of the LM of the criterion $\Delta_N(s)$ as

$$\Delta_{min}(s) = J^0(s) + \delta^2_{s+1}\lambda^2_s/\delta^2_{\xi_{s+1}}; \tag{3.121}$$

8. increasing the complexity by one unit. Return to step 4.

Note: The above algorithms describe the calculating LM for two forms of an ideal criterion. Extrapolation of LM allows one to detect the true signal from the noisy data [45]. To develop an algorithm for calculating LM of the minimum-bias criterion, certain conditions are imposed on the subsets A and B. The criterion is represented in the form of a difference of LM of two ideal non-quadratic criteria as

$$\eta_{bs} = |\,\|\hat{y}^A - \overset{\circ}{y}\| - \|\hat{y}^B - \overset{\circ}{y}\|\,| \to \min. \tag{3.122}$$

In the same way, one can also eliminate $\overset{\circ}{y}$ for special data samples. If all these are possible, then the inductive learning algorithms can be replaced by analytical calculation of LM for number of criteria. This leads to additional investigation.

5 BALANCE CRITERION OF PREDICTIONS

The criterion of balance-of-variables is the first of several kinds developed as a balance criterion. It is the simplest criterion to use to find a definite relationship (a physical law) of several variables of the process being simultaneously predicted. This has opened the basis for long-range predictions using the ring of 'direct' and 'inverse' functions and is similar to the balance-of-variables criterion [117].

The balance criterion is designed to choose models of optimal complexity with respect to several interrelated variables being modeled. This occupies an important place among the external criteria because of its nature as a system criterion and because it is used in two-level algorithms. It is still in its basic form in the multilevel modeling of different practical problems. Let us give a general form of the criterion. Later, we should delve into the nature of change in position of the minimum with increase in noise intensity.

First, we give the balance criterion in a set of interrelated variables to be modeled. Let us assume that some connections are known or established between the variables at every instant of modeling; for example,

$$\phi_k = f(y_1, y_2, \cdots, y_L); \quad k \in W \tag{3.123}$$

is a known connection, where $y_1, y_2, \cdots, y_L$ are the interrelated variables which are independently identified. The balance criterion is written as

$$B^2_H = \sum_{k\in C}\left[\hat{\phi}_k - f(\hat{y}_{jk})\right]^2, \tag{3.124}$$

where $\hat{\phi}$ and $\hat{y}$ are the predicted values of ϕ and y. The established connection is a constraint that all the functions $y_{jk}, j = 1, 2, \cdots, L$ are assumed to satisfy both in the interpolation region $k = 1, 2, \cdots, N_W$, and in the prediction region $k = 1, 2, \cdots, N_C$.

The balance criterion B_H is intended to reflect either nonlinear or linear connection between the variables. The nonlinear balance connection is known in the form of the ring of differences of the "direct" and "inverse" functions. The linear relationship among the variables being modeled can be established as

$$\phi_k = \sum_{j=1}^{L} \beta_{jk} y_{jk}; \quad k = 1, 2, \cdots, N, \tag{3.125}$$

where β are the balance coefficients which are determined from the experimental data. This enables us to generate a linear balance criterion of the form

$$B^2_{(lin)} = \sum_{k \in G} [\hat{\phi}_k - \sum_{j=1}^{L} \beta_{jk} \hat{y}_{jk}]^2, \tag{3.126}$$

where G is the set that belongs to an arbitrary part or prediction part of initial measurements.

The linear type of the balance criterion is widely used in inductive learning algorithms. They are often based on a precisely known relationship; for example, the change in the population of a city is always to the population increment minus its decrement during a certain period; total biomass of a plant is always equal to the sum of the biomasses of the parts above and below the surface. In these examples, the balance coefficients are unity.

Second, given here is the balance criterion using the relationship of moving or sliding average as a variable and its elements. This can be used successfully in algorithms for separately predicting the chosen time functions defined from the series data $y_k, k = 1, 2, \cdots, N$. The balance connection is

$$\bar{y}_k = \frac{1}{L} \sum_{j=-\frac{1}{2}(L-1)}^{+\frac{1}{2}(L-1)} y_{k+j}; \quad k = 1, 2, \cdots, N_W. \tag{3.127}$$

The relationship holds between the measured and averaged values of length L. The balance criterion is written as

$$B^2 = \sum_{k \in C} [\hat{\bar{y}}_k - \frac{1}{L} \sum_{j=-\frac{1}{2}(L-1)}^{+\frac{1}{2}(L-1)} \hat{y}_{k+j}]^2, \tag{3.128}$$

which is based on the predictions of the $\bar{y}$ and $y_1, y_2, \cdots, y_L$ of the process.

The moving averages $\bar{y}_k, k = 1, 2, \cdots, N - L$ from the initial data $y_j, j = 1, 2, \cdots, N$ can be obtained by using the matrix $\sigma\,[N - L \times N]$ form [130] as

$$\bar{y} = \frac{1}{L} \sigma_{N-L,N} y, \tag{3.129}$$

where $y^T = (y_1 y_2 \cdots y_N)$; $\bar{y}^T = (\bar{y}_1 \bar{y}_2 \cdots \bar{y}_{N-L})$; and

$$\sigma_{N-L,N} = \begin{bmatrix} 11 \ \ldots \ 10 \ \ldots \ 00 \ \ldots \ 00 \\ 01 \ \ldots \ 11 \ \ldots \ 00 \ \ldots \ 00 \\ \quad \ldots \quad \ldots \quad \ldots \\ 00 \ \ldots \ 00 \ \ldots \ 11 \ \ldots \ 10 \\ 00 \ \ldots \ 00 \ \ldots \ 01 \ \ldots \ 11 \end{bmatrix}.$$

The matrix used here has consecutive 1's of length L in each row. In adjacent rows the set of 1's is shifted one place to the right. The averaged vector $\bar{y}$ is of length $N - L$. The above

criterion can be formally put in the form as below, though y_{ik} are not individual variables to be modeled.

$$B_{av}^2 = \sum_{k \in G} (\hat{Y}_k - \sum_{i=1}^{L} \beta_i \hat{y}_{ik})^2, \tag{3.130}$$

where $Y_k = \bar{y}_k$; $\beta_i = 1/L$; and $y_{ik} = y_{i-(k-(L-1)/2)}$.

The linear concept of the balance criterion is extended further as a balance-of-predictions criterion in modeling of time series data which is cyclic in nature. This is used in algorithms of two-level predictions, in which the connection between the predictions of artificial variables $q_j, j = 1, 2, \cdots, L$ is obtained from the time series data q_k, $k = 1, 2, \cdots$, by averaging on different time intervals; for example, season and year, month and year, and hour and day.

In general, we assume that a year contains L arbitrary intervals, the months. At the lower level of the algorithm one predicts the mean monthly values of the process and at the upper level, the mean yearly values. This means that we use two-dimensional time readout in months t and years T instead of actual one-dimensional time readout, in the usual manner of continuous mean monthly data. There is a unique pair of values (t, T); $t = 1, 2, \cdots, L$; $T = 1, 2, \cdots, N$, where L is number of months (twelve months), and N is number of years, for each observation corresponding to the original measurement. The average annual values Q_T and the monthly values $q_{t,T}$ are connected by the relationship called calender averaging.

$$Q_T = \frac{1}{L} \sum_{t=1}^{L} q_{t,T}. \tag{3.131}$$

The balance-of-predictions criterion has the form

$$B_{year}^2 = \sum_{T \in G} (\hat{Q}_T - \frac{1}{L} \sum_{t=1}^{L} \hat{q}_{t,T})^2. \tag{3.132}$$

This type of criterion is used in various applications; for example, predictions of river flows, air temperature [65], and the elements of the ecosystem of a lake [48].

The operation of calculating the mean annual values Q_T can be represented in the matrix form as

$$Q_T = \frac{1}{L} \sigma_{N,K} q, \tag{3.133}$$

where q is the vector of K elements, Q_T is the vector of N elements, and σ is the matrix of $[N \times K]$ as given below:

$$\sigma_{N,K} = \begin{bmatrix} 111 \ldots 100 \ldots 000 \ldots 000 \ldots 000 \ldots 000 \\ 000 \ldots 011 \ldots 100 \ldots 000 \ldots 000 \ldots 000 \\ 000 \ldots 000 \ldots 011 \ldots 100 \ldots 000 \ldots 000 \\ \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ 000 \ldots 000 \ldots 000 \ldots 000 \ldots 001 \ldots 111 \end{bmatrix}.$$

Each row of the matrix $\sigma_{N,K}$ contains L 1's and each column contains a single one; a fact that differentiates calender from the moving averages.

In this modeling, different monthly models are obtained with the consideration of both the other monthly values and the annual values; i.e., the delayed arguments in months as well as in years are considered in obtaining the monthly models. Therefore, the expression B_{year} has formal equivalence as in $B_{(lin)}$. Any balance criterion can be called a criterion of

balance of predictions as long as the predictions of different variables are compared as in $B_{(lin)}$.

One can consider the general form of external balance criterion of linear type with the given balance coefficients β_j, $j = 1, 2, \cdots, L$ in vector notation as

$$\phi = (y_1|y_2|\cdots|y_L)\beta = Y\beta, \tag{3.134}$$

where ϕ is the N-dimensional vector; Y is the $[N \times L]$ matrix; and $\beta^T = (\beta_1, \beta_2, \cdots, \beta_L)$ is the vector of balance coefficients. The balance criterion is written as

$$B^2_{(lin)} = (\hat{\phi}_G - \hat{Y}_G\beta)^T(\hat{\phi}_G - \hat{Y}_G\beta) = \|\hat{\phi}_G - \hat{Y}_G\beta\|^2, \tag{3.135}$$

which is used on set G.

The important thing one has to note is that the balance criterion which is established among the variables ϕ and y_k, $k = 1, 2, \cdots, L$ indicates the linear dependence and has to be taken into account when using the balance criterion while process modeling.

5.1 Noise immunity of the balance criterion

Here we assume that the variable ϕ appearing in the balance criterion is physically measurable and that the balance coefficients are given (as a special case, $\beta_i = 1/L$).

Let us assume that the measurements of all the jointly modeled variables ϕ, y_1, y_2, $\cdots$, y_L are noisy.

$$\phi = \overset{\circ}{y}_0 + \xi_0; \quad y_i = \overset{\circ}{y}_i + \xi_i; \quad i = 1, 2, \cdots, L, \tag{3.136}$$

where $\overset{\circ}{y}_0, \overset{\circ}{y}_1, \cdots, \overset{\circ}{y}_L$ are the vectors of nonnoisy measurements, all the noise vectors ξ are independent of each other, and they normally have distributed independent components with mean zero and given variances.

$$\begin{aligned} &E[\xi_j] = 0, \; E[\xi_j\xi_j^T] = \sigma_j^2 I_N; \; j = 0, 1, \cdots, L, \\ &E[\xi_j^T\xi_i] = 0, \quad E[\xi_j\xi_i^T] = 0; \quad j \neq i; \; j, i = 0, 1, \cdots, L, \end{aligned} \tag{3.137}$$

where E is the mathematical expectation and I is the identity matrix. The exact models of the variables $\overset{\circ}{y}$ have the forms

$$\overset{\circ}{y}_j = \overset{\circ}{X}_j b_j^0; \quad j = 0, 1, 2, \cdots, L, \tag{3.138}$$

where $\overset{\circ}{X}_j$ are the $[N \times s_j^0]$ matrices of true independent arguments, b_j^0 are the $[s_j^0 \times 1]$ exact vectors of coefficients, and s_j^0 denote the complexities of the true models.

In self-organization modeling, one seeks for the optimal approximations to the true models from the noisy observations. The partial models are generated by sorting among the basis sets of $N \times m_j$ arguments X_j in which there are also, by assumption, true arguments $\overset{\circ}{X}_j$; that is, $m_j \geq s_j^0$, $j = 0, 1, \cdots, L$. Thus, in the sorting, one determines the coefficients in the partial models of differing complexity for each of the $L + 1$ variables in the conditional equations of the form

$$X_{j(s_j)}a_{j(s_j)} = y_j; \quad X_{j(s_j)}[N \times s_j], \; a_{j(s_j)}[s_j \times 1], \tag{3.139}$$

where s_j denotes the complexity of the partial model for the jth variable considering y_0 as ϕ for uniformity. All s_j vary independently, and the vectors of coefficients are determined by the least squares method using the noisy data.

$$\hat{a}_{j(s_j)} = [X^T_{j(s_j)} X_{j(s_j)}]^{-1} X^T_{j(s_j)} y_j. \tag{3.140}$$

For the existence of the inverse matrices, we assume that $N \geq \max(m_j, j = 0, 1, 2, \cdots, L)$ and that all the X_j have full rank.

The data sample (of length N) is not partitioned into number of sets because the balance criterion can be used on the interpolation region of the data. This means that all N points of the data are taken into consideration in all operations of the modeling.

The estimates for each variable being modeled can be written as

$$\begin{aligned} \hat{y}_{j(s_j)} &= X_{j(s_j)} \hat{a}_{j(s_j)} \\ &= X_{j(s_j)} [X^T_{j(s_j)} X_{j(s_j)}] X^T_{j(s_j)} y_j \\ &= P_{j(s_j)} y_j = P_{j(s_j)} (\overset{\circ}{y}_j + \xi_j); \quad j = 0, 1, 2, \cdots, L, \end{aligned} \tag{3.141}$$

where $\hat{y}_{0(s_0)}$ and y_0 are considered as $\hat{\phi}_{(s_0)}$ and ϕ, respectively, for uniformity, and $P_{j(s_j)}$, $j = 0, 1, 2, \cdots, L$ are the projection matrices. The balance criterion can be obtained as

$$\begin{aligned} B^2_{(lin)} &= \|\hat{y}_{0(s_0)} - (\hat{y}_{1(s_1)} | \cdots | \hat{y}_{L(s_L)}) \beta\|^2 \\ &= \|P_{0(s_0)} y_0 - (P_{1(s_1)} y_1 | \cdots | P_{L(s_L)} y_L) \beta\|^2 \\ &= \|[P_{0(s_0)} \overset{\circ}{y}_0 - (P_{1(s_1)} \overset{\circ}{y}_1 | \cdots | P_{L(s_L)} \overset{\circ}{y}_L) \beta] \\ &\quad + [P_{0(s_0)} \xi_0 - (P_{1(s_1)} \xi_1 | \cdots | P_{L(s_L)} \xi_L) \beta]\|^2. \end{aligned} \tag{3.142}$$

The objective of the balance criterion is to obtain consistent optimal predictive models for each of the $L + 1$ variables connected by the balance law. The criterion $B_{(lin)}$ is to be calculated for all variants of the partial models of varying complexity. The total amount of partial models is calculated as

$$p_B = \prod_{j=0}^{L} p_{m_j} = \prod_{j=0}^{L} (2^{m_j} - 1). \tag{3.143}$$

If $m_j = m$, then the complete sorting is proportional to 2^{mL}; obviously, in complex problems, for large m and L, complete sorting becomes impossible. We can tentatively assume that, as in the case of the combinatorial inductive approach, the complete sorting is efficient when $mL < 20$. For four seasons ($L = 4$), we can allow $m = 5$ arguments in each model of the seasons. For more complex problems, it is essential to apply proper way of sorting. Here let us assume that complete sorting is made. We shall seek the minimum of the balance criterion.

$$B^2_{(min)} = \min_{i=\overline{1,p_{m_j}};\ j=\overline{0,L}} B^2_{(lin)} [\hat{a}_{ij}(X_j, y_j)]. \tag{3.144}$$

The value of $B^2_{(min)}$ determines the set of models of optimal complexity s^*_j, $j = 0, 1, 2, \cdots, L$ for each of $L+1$ variables. Now let us see how the choice of models of optimal complexity changes with increase in the noise variances σ^2_j, $j = 0, 1, 2, \cdots, L$; i.e., let us investigate the

noise immunity of the choice with respect to the balance criterion. This is analyzed in the mean value sense; i.e., with respect to its mathematical expectation.

Keeping in mind the noise properties, the fact that $P_j^T P_j = P_j^2 \equiv P_j$; $j = 0, 1, \cdots, L$, and applying the mathematical expectation to the above derived balance criterion

$$\bar{B}^2_{(lin)} = E[B^2_{(lin)}] = \|P_{0_{(s_0)}} \overset{\circ}{y}_0 - (P_{1_{(s_1)}} \overset{\circ}{y}_1 \mid \cdots \mid P_{L_{(s_{L'})}} \overset{\circ}{y}_L)\beta\|^2 + \sigma_0^2 s_0 + \sum_{i=1}^{L} \sigma_i^2 \beta_i^2 s_i = B^2_{\underset{y}{\circ}} + B^2_{\xi}. \tag{3.145}$$

Thus, the expected value of the balance criterion $B^2_{(lin)}$ has two components: $B^2_{\underset{y}{\circ}}$, imbalance in the modeling of exact data, and B^2_{ξ} reflects the action of the noise.

First, let us look in greater detail at the component $B^2_{\underset{y}{\circ}}$ and then at the $B^2_{(lin)}$ as a whole.

$$\begin{aligned} B^2_{\underset{y}{\circ}} &= \|P_{0_{(s_0)}} \overset{\circ}{y}_0 - (P_{1_{(s_1)}} \overset{\circ}{y}_1 \mid \cdots \mid P_{L_{(s_{L'})}} \overset{\circ}{y}_L)\beta\|^2 \\ &= \|X_{0_{(s_0)}} \hat{b}_{0_{(s_0)}} - (X_{1_{(s_1)}} \hat{b}_{1_{(s_1)}} \mid \cdots \mid X_{L_{(s_{L'})}} \hat{b}_{L_{(s_{L'})}})\beta\|^2 \\ &= \|\hat{y}^{\circ}_{0_{(s_0)}} - (\hat{y}^{\circ}_{1_{(s_1)}} \mid \cdots \mid \hat{y}^{\circ}_{L_{(s_{L'})}})\beta\|^2 \\ &= \|\hat{y}^{\circ}_{0_{(s_0)}} - \hat{Y}^{\circ}\beta\|^2. \end{aligned} \tag{3.146}$$

Considered with the exact data, it is necessary to determine $B^2_{\underset{y}{\circ}} = 0$ and to check whether or not this corresponds to obtaining true models for which the balance relationship

$$\overset{\circ}{y}_0 = (\overset{\circ}{y}_1 \mid \overset{\circ}{y}_2 \mid \cdots \mid \overset{\circ}{y}_L)\beta = \overset{\circ}{Y}\beta; \quad \overset{\circ}{Y}\ [N \times L] \tag{3.147}$$

holds, and which is actually reflected in the exact initial data $\overset{\circ}{y}_0, \overset{\circ}{y}_1, \overset{\circ}{y}_2, \cdots, \overset{\circ}{y}_L$.

With increasing complexity of the models on the attainment of the true complexity s_j^0 (for each one of $L+1$ models), the coefficients are restored exactly to $\hat{b}_{j_{(s_j^0)}} \equiv \hat{b}_j^0$. Even with further increase in s_j, the coefficients $\hat{b}_{j_{(s_j \mid s_j \geq s_j^0)}} \equiv b_j^0$ do not change because the coefficients of the extra arguments are equal to zero. The models of all the variables are attained true and the value of the criterion $B_{\underset{y}{\circ}} = 0$. It is also possible, as it might turn out, that the criterion assumes the value $B_{\underset{y}{\circ}} = 0$ in the cases when the sorting among the different combinations of models for all the variables discloses partial models with coincidence structures such as $X_{0_{(s')}} = X_{1_{(s')}} = \cdots = X_{L_{(s')}} = X^*$. As $\overset{\circ}{y}_0 = \overset{\circ}{Y}\beta$, $P_{0_{(s')}} = P_{1_{(s')}} = \cdots = P_{L_{(s')}} = P^*$, $B_{\underset{y}{\circ}}$ becomes

$$B^2_{\underset{y}{\circ}} = \|P^*(\overset{\circ}{y}_0 - \overset{\circ}{Y}\beta)\|^2 \equiv 0 \tag{3.148}$$

for arbitrary values of the coefficients.

This property of the criterion is mentioned by Ihara [27] in his correspondence with the editor of the journal "avtomatika" (Soviet Journal of Automatic Control). Later, the idea arose as to nonuniqueness of the choice of models according to the balance criterion [36].

Theorem 1. Estimation of the coefficients of the prediction models with regard to the prediction balance criterion is an incorrect problem because it has an enormous set of solutions.

Proof: Considering β as a unit vector, the prediction balance criterion can be written as

$$B^2 = \|X_0 a_0 - (X_1 a_1 | \cdots | X_L a_L)\|^2, \tag{3.149}$$

where a_j, $j = 0, 1, 2, \cdots, L$ are the vectors of coefficients in the models of averaged data and the detailed prediction models, correspondingly.

This is to be minimized with respect to the coefficients a_0 and a_j to obtain the optimal set of prediction equations. The system of normal equations in Gaussian form can be obtained as

$$\frac{\partial B^2}{\partial a_j} = 0; \; j = 0, 1, 2, \cdots, L, \tag{3.150}$$

Assuming that the structures of the models are already known,

$$\begin{aligned} 2X_0^T X_0 a_0 - 2X_0^T (X_1 a_1 | \cdots | X_L a_L) &= 0, \\ -2X_j^T X_0 a_0 + 2X_j^T (X_1 a_1 | \cdots | X_L a_L) &= 0. \end{aligned} \tag{3.151}$$

These matrix equations are linearly dependent. Each of the equations has an infinite set of solutions and the system of equations yields the trivial solution $\hat{a}_j = 0$; $j = 0, 1, 2, \cdots, L$.

Corollary 1. In modeling the exact data, it is necessary to obtain the value of $B^2_{\underset{y}{\circ}} = 0$, which is sufficient for structural identification of the true models ($\overset{\circ}{y}_j = \overset{\circ}{X}_j \, b_j^0$; $j = 0, 1, 2, \cdots, L$), if (i) the exact data of the variables $\overset{\circ}{y}_j$; $j = 0, 1, 2, \cdots, L$ satisfy the balance relationship ($\overset{\circ}{y}_0 = \overset{\circ}{Y} \beta$); (ii) the arguments of the matrices X_j contain all true arguments of the $\overset{\circ}{X}_j$, $j = 0, 1, 2, \cdots, L$; and (iii) the common basis of the arguments is nondegenerate; i.e., sorting does not reveal a complexity s' such that $X_{1_{(s')}} = \cdots = X_{L_{(s')}}$. These three conditions are neither excessively stringent nor idealized. The first two make the problem of modeling several variables connected by the linear balance relationship well posed and the third establishes the conditions for correct application of the balance criterion (uniqueness of choice of models), which can be ensured algorithmically.

Attainment of the value of the criterion ($B^2_{\underset{y}{\circ}} = 0$) is achieved with increase in the complexity, which is always monotonic. This can be explained by the calculation of $B^2_{\underset{y}{\circ}}$ on the same data as it is used to estimate the coefficients. This can be represented graphically as the dependence of the criterion on the complexity of the partial models of the different variables in a multidimensional space.

Let us look at the second part B^2_ξ of the balance criterion. The component of the criterion $\bar{B}^2_{(lin)}$ which reflects the influence of noise is

$$B^2_\xi = \sigma_0^2 s_0 + \sum_{j=1}^{L} \beta_j^2 \sigma_j^2 s_j. \tag{3.152}$$

This increases linearly with an increase in the complexity of the partial models $\hat{y}_{j(s_j)}$; $j = 0, 1, 2, \cdots, L$; i.e., it is the plane whose inclination in multidimensional space is determined by the noise variances σ_j^2; $j = 0, 1, 2, \cdots, L$. This inclination increases according to the increase in the noise variances.

Keeping in mind the properties of the component $B^2_{\underset{y}{\circ}}$, we conclude that (i) the criterion $B^2_{(lin)}$ being an $(L+1)$-dimensional function of the variables (in numbers) s_j; $j = 0, 1, 2, \cdots, L$

always has a unique minimum, (ii) this minimum is always in the hypercube $1 \le s_j \le s_j^0$; $j = 0, 1, 2, \cdots, L$ (i.e., the overly complex superfluous models are always weeded out), and (iii) with increase in the variances of the noises (at least in one of them), the minimum is displaced on the side of decrease in the complexity of the models (with respect to at least one of the variables).

These properties can be represented graphically. One can observe the decrease in the optimal complexity of the models, which is typical of external criteria, using a multi-dimensional surface whose sections (isolines) for different noise variances are σ_j^2; $j = 0, 1, 2, \cdots, L$.

Corollary 2. If the three conditions which are asserted in corollary 1 for $B_{\underset{y}{\circ}}^2$ also hold in the modeling of noisy variables y_j; $j = 0, 1, 2, \cdots, L$, then the following properties of the selected models in optimal complexity as per the balance criterion $\bar{B}_{(lin)}^2$ hold: (i) for arbitrary nonzero noise variances σ_j^2; $j = 0, 1, 2, \cdots, L$, the minimum of the criterion as a function of different complexities s_j; $j = 0, 1, 2, \cdots, L$ exists and is unique, (ii) the achieved minimum always lies in the bounded region $1 \le s_j \le s_j^0$, $j = 0, 1, 2, \cdots, L$, where s_j^0 is the complexity of the true models, and (iii) with increase in the variances of the noises σ_j^2; $j = 0, 1, 2, \cdots, L$ the minimum is displaced in the direction of decrease in the complexity of the optimal models.

Theorem 2. The problem of estimating the coefficients for the prediction models as per the balance criterion becomes correct (having a unique solution) if the quadratic stability criterion S is considered along with the balance criterion; i.e., by forming the combined criterion as

$$c5^2 = B^2 + S^2. \tag{3.153}$$

Proof: Let us consider the stability criterion as a stabilizing functional, the sum of the quadratic criteria giving the quality of the output vector error on each of the prediction levels.

$$S^2 = \sum_{j=0}^{L} \|y_j - X_j a_j\|^2. \tag{3.154}$$

The combined criterion $c5$ is the combination of "prediction balance plus stability criterion."

$$c5^2 = \|X_0 a_0 - (X_1 a_1 | \cdots | X_L a_L\|^2 + \sum_{j=0}^{L} \|y_j - X_j a_j\|^2. \tag{3.155}$$

Let us determine the estimates of the vectors a_j, $j = 0, 1, 2, \cdots, L$ by minimizing the combined criterion. We obtain the system of normal equations as

$$\frac{\partial c5^2}{\partial a_j} = 0; \quad j = 0, 1, 2, \cdots, L. \tag{3.156}$$

Then, for the prediction model of the first level $\hat{y}_0$ or $\hat{\phi}$, we have:

$$\frac{\partial c5^2}{\partial a_0} = 2X_0^T X_0 a_0 - X_0^T (X_1 a_1 | \cdots | X_L a_L) - X_0^T y_0 = 0, \tag{3.157}$$

and from this,

$$\hat{a}_{0_{(c5)}} = \frac{1}{2}(X_0^T X_0)^{-1} X_0^T [y_0 + (X_1 a_1 | \cdots | X_L a_L)]. \tag{3.158}$$

For the models of the second level $\hat{y}_j$, $j = 1, 2, \cdots, L$:

$$\frac{\partial c5^2}{\partial a_j} = 2(X_j^T X_1 a_1 | \cdots | X_j^T X_L a_L) - X_j^T X_0 a_0 - X_j^T (y_1 | \cdots | y_L) = 0; \quad (3.159)$$
$$j = 1, 2, \cdots, L$$

From this,

$$\hat{a}_{j(cs)} = \frac{1}{2}(X_j^T X_j)^{-1} X_j^T [(y_1 | \cdots | y_L) + X_0 a_0]; \quad j = 1, 2, \cdots, L. \quad (3.160)$$

The matrices $X_j^T X_j$, $j = 0, 1, 2, \cdots, L$ are assumed to be nonsingular.

Solving the system of two matrix equations of $c5^2$ and $\hat{a}_0$, we obtain the estimates of the coefficients of the prediction models of both levels as

$$\hat{a}_{0(cs)} = \frac{1}{3}(X_0^T X_0)^{-1} X_0^T [2y_0 + (y_1 | \cdots | y_L)] = \frac{2}{3}\hat{a}_0 + \frac{1}{3}(X_0^T X_0)^{-1} X_0^T (y_1 | \cdots | y_L)$$
$$\hat{a}_{j(cs)} = \frac{1}{3}(X_j^T X_j)^{-1} X_j^T [2(y_1 | \cdots | y_L) + y_0] = \frac{2}{3}\hat{a}_j + \frac{1}{3}(X_j^T X_j)^{-1} X_j^T y_0; \quad (3.161)$$
$$j = 1, 2, \cdots, L$$

where a_j, $j = 0, 1, 2, \cdots, L$ are the estimates of the coefficients of respective models as per the least squares method. Thus, the goal of the regularization is achieved.

Corollary 3. On regularization of the problem of selection of structure for prediction models by the balance-of-prediction criterion with the help of stability criterion.

The problem of structure choice of prediction models by the balance criterion becomes correct (i.e., achieving a unique solution) if the quadratic stability criterion S^2 in the combined criterion $c5^2$ is used as regularizing the operator as

$$c5^2(s_0^*, s_j^*) = \min_{s_j \in \mathbf{m}_j; j=0,L} [B^2(s_0^*, s_j^*) + S^2(s_0^*, s_j^*)], \quad (3.162)$$

where $\mathbf{m}$ denotes the set of arguments taking part in the predictive models of both levels, s_0^* and s_j^* are the notations for optimal structures.

The stability criterion in the above formulation makes it possible to reduce the region of solution of the problem of selecting structures by using the prediction balance criterion to a compact subset which leads to a unique solution of the problem.

Interpretation of the results in the case of B_{year}^2. It follows from the comparison of $B_{(lin)}^2$ and B_{year}^2 that the number of variables is equal to, for example, the number of seasons ($L = 4$); the vector of connected variable is the vector of second-level variable $\phi = Q = (Q_1, Q_2, \cdots, Q_N)^T$. The remaining variables are the seasonal variables associated with the first level $y_j = q_t = (q_{t,1}, q_{t,2}, \cdots, q_{t,N})^T$, $t = 1, 2, \cdots, L$. All the balance coefficients are equal $\beta_i = 1/L$, $\beta = (1/L, \cdots, 1/L)^T$. However, for the results of the investigation of noise immunity of the criterion $B_{(lin)}^2$ to be applicable to the criterion B_{year}^2 (or what amounts to the same thing,) to

$$B_{year}^2 = \|\hat{Q} - (\hat{q}_1 | \cdots | \hat{q}_L)\beta\|^2, \quad (3.163)$$

it is necessary to show the validity of noise conditions specified before. The vectors Q and q_t; $t = 1, 2, \cdots, L$ are obtained from the measured time series data q_k; $k = 1, 2, \cdots, LN$. Let us suppose that a noise with the usual properties is imposed on these measurements,

$$q_k = \overset{\circ}{q}_k + \zeta_k; \quad E[\zeta_k] = 0, \; E[\zeta_k^2] = \sigma^2, \; E[\zeta_i \zeta_j] = 0, \; i \neq j. \quad (3.164)$$

The components of each vector q_t are taken from q_k at a period of length L.

$$q_t = (q_{k=t}, q_{k=t+L}, \cdots, q_{k=t+(N-1)L})^T; \quad t = 1, 2, \cdots, L. \tag{3.165}$$

Therefore, for each q_t the noise vector ξ_t satisfies the original conditions.

$$q_t = \overset{\circ}{q}_t + \xi_t; \; E[\xi_t] = 0, \; E[\xi_t \xi_t^T] = \sigma^2 I_N, \; E[\xi_j^T \xi_i] = 0, \; j \neq i, \tag{3.166}$$

where, for all t vectors of the seasonal values, the noise variances are equal to σ^2 and are independent of t. Further, with reference to the mean annual values, we obtain

$$Q = \overset{\circ}{Q} + \xi_0, \quad \overset{\circ}{Q} = (\overset{\circ}{q}_1 \mid \cdots \mid \overset{\circ}{q}_L)\beta,$$

$$\xi_0 = (\xi_1 \mid \cdots \mid \xi_L)\beta, \quad E[\xi_0] = 0, \quad E[\xi_0 \xi_0^T] = \frac{\sigma^2}{L} I_N; \tag{3.167}$$

i.e., the noise variance for the second-level mean annual variable is $1/L$ times the variance of the original noise $\sigma_0^2 = \sigma^2/L$ which increases the noise immunity of modeling at the second level. The components of the noise vector ξ_0 are also independent. This means that one of the conditions fails to be satisfied; i.e., the condition of independence of ξ_0 and all ξ_t; $t = 1, 2, \cdots, L$.

$$\begin{aligned} E[\xi_0^T \xi_t] &= \beta^T E[(\xi_1 \mid \cdots \mid \xi_L)^T \xi_t] \\ &= \frac{1}{L} E[\xi_t^T \xi_t] = \frac{\sigma^2}{L}. \end{aligned} \tag{3.168}$$

This reflects on the calculation of the mathematical expectation of the criterion B_{year}^2, which is obtained in somewhat different form

$$\begin{aligned} \bar{B}_{year}^2 = E[B_{year}^2] &= \|\hat{Q}^\circ - (\hat{q}_1^\circ \mid \cdots \mid \hat{q}_L^\circ)\beta\|^2 \\ &+ \frac{\sigma^2}{L} s_0 + \frac{\sigma^2}{L} \sum_{i=1}^{L} s_i - 2\frac{\sigma^2}{L} \sum_{i=1}^{L} \operatorname{tr}\left(P_{0(s_0)}^T P_{i(s_i)}\right). \end{aligned} \tag{3.169}$$

The trace of the $(P_0^T P_i)$ can be written as

$$\operatorname{tr}(P_0^T P_i) = \operatorname{tr}[X_0 (X_0^T X_0)^{-1} X_0^T X_i (X_i^T X_i)^{-1} X_i^T],$$

which cannot be calculated in the general case.

It is difficult to determine how $\bar{B}_{year}^2$ behaves with complication of the models of both levels. It may fail to be unimodal, or its global minimum may not be displaced in the direction of simplification of models with increase in the noise variances.

The criterion $\bar{B}_{year}^2$ will have the same properties as $\bar{B}_{(lin)}^2$ when the matrices X_0 and X_i are orthogonal; i.e., $X_0^T X_i = 0$; $i = 1, 2, \cdots, L$. Then

$$\bar{B}_{year}^2 = \|\hat{Q}^\circ - (\hat{q}_1^\circ \mid \cdots \mid \hat{q}_L^\circ)\beta\|^2 + \frac{\sigma^2}{L}\left(s_0 + \sum_{i=1}^{L} s_i\right). \tag{3.170}$$

Although this condition usually contradicts the condition of linear connection of variables of the two levels ($Q_T = \frac{1}{L}\sum_{t=1}^{L} q_{t,T}$), one can interpret it as the specific nature of the two-level modeling problem of a single variable given by its seasonal and annual values. The orthogonality condition $X_0^T X_i = 0$ will in fact hold when the seasonal models of the

first level are not used at the second level. Here, the mean annual models are constructed independently of the seasonal models, and these are used in the selection of the best two-level models according to the balance criterion.

Corollary 4. The problem of selecting structures for predictive models by the prediction balance criterion becomes correct if a different (in nature and composition) set of arguments is used for constructing models of the two levels.

Indeed, in reconciliation of seasonal and annual predictions using the same source of measured data, the balance criterion is inefficient and leads to trivial results. On the other hand, if one uses a different set of arguments for constructing two-level models, then the balance criterion in the choice of structure becomes efficient.

In practice, such a case is ensured, for example, by constructing the seasonal models in the form of a system of L difference equations and the annual models in the form of a harmonical functions [48]. This corresponds to the condition of independence of the informational bases of the models of different levels [36], which is necessary for efficiency of algorithms for multilevel modeling.

6 CONVERGENCE OF ALGORITHMS

First we give the definition for canonical formulation of the external criteria [135] and then proofs for internal convergence of two multilayer algorithms; one is the original multilayer algorithm; the other is the algorithm with propagating errors [42], [79] [134], [136].

6.1 Canonical formulation

The canonical form of the external criteria is an analytical tool to investigate various properties of the criteria. This is not convenient from a practical standpoint for calculating the value of a criterion in cases involving large numbers of observations, but it can be used directly for model selection of a small number of observations.

Definition

The canonical form of the criterion is defined as the expression $y^T Dy$, where D is a symmetric strictly positive-semidefinite matrix—strictly in the sense that (a) $\forall y \neq \Theta,\ y^T Dy \geq 0$ and (b) $\exists y \neq \Theta,\ y^T Dy = 0$.

The matrix D is determined by the corresponding criterion and the partitioning of data.

Residual sum of squares

We give here the canonical form of residual sum of squares (RSS) used in the least-squares method. Suppose we have a system of conditional equations of the form $y = Xa$. The parameters a are estimated as

$$a = (X^T X)^{-1} X^T y. \tag{3.171}$$

The RSS is calculated as

$$\varepsilon^2 = (y - Xa)^T (y - Xa). \tag{3.172}$$

This can be written in the canonical form using the notation $P_{NN} = X(X^TX)^{-1}X^T$ as

$$\varepsilon^2 = y^T(I - P_{NN})y = y^T D_{LS} y, \tag{3.173}$$

where $D_{LS} \hat{=} (I - P_{NN})$ is a symmetric positive semidefinite matrix for $N > m$, I is the unit matrix, and N indicates the total number of data points.

Regularity criterion

This is given as

$$\Delta^2(B) = (y_B - \hat{y}_B)^T(y_B - \hat{y}_B), \tag{3.174}$$

where $\hat{y}_B = X_B(X_A^TX_A)^{-1}X_A^Ty_A$. Using the notation $P_{BA} \hat{=} X_B(X_A^TX_A)^{-1}X_A^T \hat{=} (p_{ij})$, the criterion can be written as

$$\Delta^2(B) = \sum_{i \in B}(y_{B_i} - \sum_{j \in A} p_{ij} y_{A_j})^2, \tag{3.175}$$

where A and B are the training and the testing sets correspondingly. By expanding this algebraically, we get

$$\Delta^2(B) = \sum_{i \in B} y_{B_i}^2 - 2\sum_{i \in B}\sum_{j \in A} y_{B_i} p_{ij} y_{A_j} + \sum_{i \in B}\sum_{j \in A}\sum_{k \in A} p_{ij} p_{ik} y_{A_j} y_{A_k}, \tag{3.176}$$

or the matrix form

$$\Delta^2(B) = \left(y_A | y_B\right) \left(\begin{array}{c|c} \sum_{i \in B} p_{ij} p_{ik} & (-p_{ij}) \\ \hline (-p_{ij}) & I \end{array}\right) \left(\begin{array}{c} y_A \\ \hline y_B \end{array}\right)$$

$$= y^T \left(\begin{array}{c|c} P_{BA}^T P_{BA} & -P_{BA}^T \\ \hline -P_{BA} & I \end{array}\right) y = y^T D_{reg} y. \tag{3.177}$$

This is the canonical form for the regularity criterion. The matrix D_{reg} depends on the sequencing of the training and testing sets—so does vector y.

Minimum bias criterion

This is given as

$$\eta_{bs}^2 = (\hat{y}^A - \hat{y}^B)^T(\hat{y}^A - \hat{y}^B)_W, \tag{3.178}$$

where W indicates that the criterion is computed on the set W; $\hat{y}_W^G = X_W(X_G^TX_G)^{-1}X_G^Ty_G$; G corresponds to either A or B and $W = A \cup B$.

Let us define the notations as

$$X_W(X_G^TX_G)^{-1}X_G^Ty_G \hat{=} P_{WG}y_G, \quad G = A \text{ or } B. \tag{3.179}$$

The criterion can be rewritten as

$$\eta_{bs}^2 = (P_{WA}y_A - P_{WB}y_B)^T(P_{WA}y_A - P_{WB}y_B). \tag{3.180}$$

The canonical form can be obtained as

$$\eta_{bs}^2 = y^T \left(\begin{array}{c|c} P_{WA}^T P_{WA} & -P_{WA}^T P_{WB} \\ \hline -P_{WB}^T P_{WA} & P_{WB}^T P_{WB} \end{array}\right) y = y^T D_{bs} y. \tag{3.181}$$

This is the canonical form for the minimum bias criterion.

Analogously, one can obtain canonical forms for other criteria.

6.2 Internal convergence

Defining multilayer algorithm with propagating outputs

Let us assume that there are m input variables of x $(x_1, x_2, \cdots, x_m)$, y is the output variable, G_r is the set of q input variables at the rth layer $(z_1, z_2, \cdots, z_q)$, $(q \geq m)$, N is the number of initial data points. Mapping $\mathcal{R}$ takes place from layer r to the layer $r+1$; i.e., $\mathcal{R} : G_r \rightarrow G_{r+1}$.

First, the elements $z_k^{(r)}, k = 1, 2, \cdots, F$ are the column vectors of the matrix $z^{(r)}$ of the transformed experimental data. They are determined from the condition

$$z_k^{(r)} = P_k y, \tag{3.182}$$

where $P_k = (z_i^{(r-1)}|z_j^{(r-1)})[(z_i^{(r-1)}|z_j^{(r-1)})^T(z_i^{(r-1)}|z_j^{(r-1)})]^{-1}(z_i^{(r-1)}|z_j^{(r-1)})^T$, is the projection operator of the least-squares method; and y is the observation vector of output variable.

Second, the N-dimensional vector z is a partial description of the rth layer as its kth component is expressed by

$$z_{(k)}^{(r)} = g\left(z_{i(k)}^{(r-1)}, z_{j(k)}^{(r-1)}\right), \tag{3.183}$$

where i, j vary as per their representation from the $(r-1)$st layer. The partial polynomial in its simplest form is

$$g(z_{i(k)}, z_{j(k)}) = a_1 z_{i(k)} + a_2 z_{j(k)}; \quad i = 1, 2, \cdots, q-1; \; j = i+1, i+2, \cdots, q \tag{3.184}$$

where a_1 and a_2, as the arbitrary coefficients, assumes an iterative process.

Finally, from the set of elements $z_{(k)}^{(r)}$ of the following layer that is obtained, a subset $z^{(r)}$ is singled out according to an external criterion. The external criterion gives to these solutions qualitatively new properties that the modeler finds desirable.

Suppose the regularity criterion $\Delta^2(B)$ is considered as the external criterion that has the sum of squares of the deviations on the testing set B

$$y = \begin{pmatrix} y_A \\ - - - \\ y_B \end{pmatrix}, \quad Z^{(r-1)} = \begin{pmatrix} Z_A^{(r-1)} \\ - - - \\ Z_B^{(r-1)} \end{pmatrix}. \tag{3.185}$$

The algorithm stops when the criterion achieves the minimum in the layer r compared with the layer $r+1$ for a particular component; it is then said that it is converged; i.e.

$$y^T D_{reg}^{(r)} y \leq y^T D_{reg}^{(r+1)} y, \tag{3.186}$$

where $D_{reg}^{(r)}$ and $D_{reg}^{(r+1)}$ are the positive-semidefinite canonical matrices formed based on the components at r and $r+1$ layers, correspondingly.

Internal convergence is an especially important property of multilayer algorithms. If the external criterion becomes the internal criterion (i.e., the regularity criterion $\Delta^2(B)$ becomes the residual sum of squares (RSS) ε^2), the result of the algorithm must be equivalent to the result of multiple regression analysis, at least when the function of y is linear in variables and coefficients.

Here the internal convergence is considered (i) towards a solution and (ii) with respect to the structures.

Convergence to a solution. Suppose that stopping is not envisioned and the class of functions formed by superposition of the function g includes a function $h(x_{(k)}) = y_{(k)}$, $k = 1, 2, \cdots, N$ where $x_{(k)} = (x_{1(k)}, x_{2(k)}, \cdots, x_{m(k)})$, then the algorithm converges to a solution if the sequence of vectors $z_i^{(r)}$ has a limit as $r \rightarrow \infty$ and if this limit is y.

Convergence with respect to structures. Suppose that stopping is not envisioned and the class of functions formed by superposition of the function g includes a unique function $h(x)$ such that $h(x_{(k)}) = y_{(k)}$, $k = 1, 2, \cdots, N$, then the algorithm converges with respect to the structure if the sequence of functions $z_i^{(r)}(x)$ has a limit as $r \to \infty$ and if it is equal to $h(x)$. Unlike the above case, here it takes the measure of distance between the functions. In the class of linear polynomials, a natural measure for distance between two functions is the sum of squares of the distances between similar terms involved in them. The distance between two arbitrary functions is measuerd as the sum of the squares of distances between their values from the initial data. Based on this, the definitions of convergence to a solution and with respect to structure are equivalent.

Definition 1. An algorithm converges in a finite number of steps if, beginning with some layer, $z_l^{(r)}$ are equal to their limiting value.

Definition 2. There is effective convergence if the algorithm converges in a finite number of steps; i.e., the layer with which $z_l^{(r)}$ is the first one and equal to its limiting value; the next layer has the divergent characteristics.

Definition 3. It is referred to the convergence under the condition that $\Delta^2(B) = \text{RSS}$, where RSS is calculated on the initial data, as internal convergence.

The internal convergence to the solution and in structure is ensured by the following theorem.

Theorem 3. Suppose that y^* is the projection of the vector y on to the linear space $L(X)$, formed by the columns of the matrix X. Suppose the criterion is calculated on the set W. Then, F number of partial descriptions with the sequence of vectors $z_k^{(r)}$ converges to y^* as $r \to \infty$. If the X^TX is nonsingular, the model corresponding to the limiting vector coincides with the regression equation for y as a function of X.

Assume that the best model in optimal complexity is being sought, that means it is the case of $F = 1$.

Let us look at the numerical sequence of $\|y - z^{(r)}\|$, which can be shown as nonincreasing. In the multilayer algorithm with the propagating outputs, the vector $z^{(r+1)}$ is formed by

$$z^{(r+1)} = a_1 z_i^{(r)} + a_2 z_j^{(r)}, \tag{3.187}$$

where a_1 and a_2 are found by minimizing the quantity $\|y - a_1 z_i^{(r)} - a_2 z_j^{(r)}\|$. It follows that the vector $z^{(r+1)}$ is the projection y onto $L(z_i^{(r)} | z_j^{(r)})$; i.e., the linear hull of the vectors $z_i^{(r)}$ and $z_j^{(r)}$.

$$\|y - z^{(r+1)}\| \leq \|y - z^{(r)}\|, \quad r = 0, 1, \cdots \tag{3.188}$$

Thus, the sequence $\|y - z^{(r)}\|$ is nonincreasing and as a sequence of norms it is lower bounded. Therefore it has a limit that is denoted by ϱ.

Let us look at the sequence $\|z^{(r)}\|$. By the definition, $z_k^{(r)} \in L(X)$ for all r. Consequently, $z \in L(X)$. Further more, $(y - z)$ is orthogonal to $L(X)$; i.e., $(y - z)^T X = 0$. It follows from the above that $\|z^{(r+1)}\| \geq \|z^{(r)}\|$ and one can easily see that $\|z^{(r)}\| \leq \|y\|$. Thus, the sequence $\|z^{(r)}\|$ is nondecreasing and higher bounded. It has a limit, which is denoted by τ.

Let us look at the sequence $z^{(r)}$. The existence of the limits of the sequences $\|y - z^{(r)}\|$ and $\|z^{(r)}\|$ implies that with increasing r, the vectors $z^{(r)}$ become arbitrarily closer to the

manifold defined by the system of equations

$$\begin{aligned} \|y - z\| &= \varrho \\ \|z\| &= \tau. \end{aligned} \tag{3.189}$$

It is shown that there exists a unique vector z^* belonging to this manifold, which is the limiting vector of the sequence $z^{(r)}$. It follows that

$$\begin{aligned} z^{(r+1)} &= (z_i^{(r)}|z_j^{(r)})\,[(z_i^{(r)}|z_j^{(r)})^T(z_i^{(r)}|z_j^{(r)})]^{-1}(z_i^{(r)}|z_j^{(r)})^T y \\ &= \frac{z_j^{(r)^T} z_j^{(r)} z_i^{(r)} z_i^{(r)^T} - z_i^{(r)^T} z_j^{(r)} z_i^{(r)} z_j^{(r)^T} - z_i^{(r)^T} z_j^{(r)} z_j^{(r)} z_i^{(r)^T} + z_i^{(r)^T} z_i^{(r)} z_j^{(r)} z_j^{(r)^T}}{z_i^{(r)^T} z_i^{(r)} z_j^{(r)^T} z_j^{(r)} - (z_i^{(r)^T} z_j^{(r)})^2} y \\ &= P_{ir} y, \end{aligned} \tag{3.190}$$

where P_{ir} denotes the corresponding projection matrix.

It follows from the convergence of the sequence $\|y - z^{(r)}\|$ that by choosing r suitably, the equation $\|y - P_{ir}y\| = \varrho$ can be satisfied to any desired closeness for all $i = 1, 2, \cdots, m$.

From the above, the following equation

$$\begin{aligned} (y^T y - \varrho^2)[z_i^{(r)^T} z_i^{(r)} z_j^{(r)^T} z_j^{(r)} - (z_i^{(r)^T} z_j^{(r)})^2] = {} & y^T (z_j^{(r)^T} z_j^{(r)} z_i^{(r)} z_i^{(r)^T} - z_i^{(r)^T} z_j^{(r)} z_i^{(r)} z_j^{(r)^T} \\ & - z_i^{(r)^T} z_j^{(r)} z_j^{(r)} z_i^{(r)^T} + z_i^{(r)^T} z_i^{(r)} z_j^{(r)} z_j^{(r)^T}) y \end{aligned} \tag{3.191}$$

will be satisfied for any desired accuracy by noting that $z_i^{(r)^T} z_i^{(r)} z_j^{(r)^T} z_j^{(r)} - (z_i^{(r)^T} z_j^{(r)})^2 \neq 0$.

The unknowns $z_i^{(r)^T} z_j^{(r)}$ can be determined to an arbitrary degree of accuracy from the above equation because of its dependence on coefficients in terms of the unknowns. The solution can be found as

$$z_i^{(r)^T} z_j^{(r)} = z_i^{(r)^T} y, \quad i = 1, 2, \cdots, m \tag{3.192}$$

using the relationships $\varrho^2 + \tau^2 = y^T y$, and $y^T z^{(r)} = z^{(r)^T} z^{(r)}$. This is satisfied with an arbitrary accuracy as the quantities $\|y - z^{(r)}\|$ and $\|z^{(r)}\|$ tending to their limits ϱ and τ, respectively. This determines uniquely the limiting vector $z^* \in L(X)$. It can be written as

$$X^T(z^* - y) = 0. \tag{3.193}$$

Thus, z^* is the orthogonal projection of y on to $z(X)$ or, what amounts to the same thing, $z^* = y^*$.

Let us look at the case $F > 1$. It shows that the distances between the partial descriptions belonging to the same layer get arbitrarily smaller as $r \to \infty$; i.e., $\|z_k^{(r)} - z_l^{(r)}\|$, $k = 1, \cdots, F-1$, $l = k+1, k+2, \cdots, F$ gets arbitrarily small. Let us define $\|y - z_1^{(r)}\| = \varrho + \delta_r$, and it leads to

$$\|y - z_F^{(r+1)}\| \leq \|y - z_1^{(r)}\|. \tag{3.194}$$

We consider F partial descriptions of the form $a_1 z_i^{(r)} + a_2 z_j^{(r)}$ at the $(r+1)$st layer, for which the above inequality holds. Therefore,

$$\varrho + \delta_r \leq \|y - z_k^{(r)}\| \leq \varrho + \delta_{r-1}, \quad k = 1, 2, \cdots, F. \tag{3.195}$$

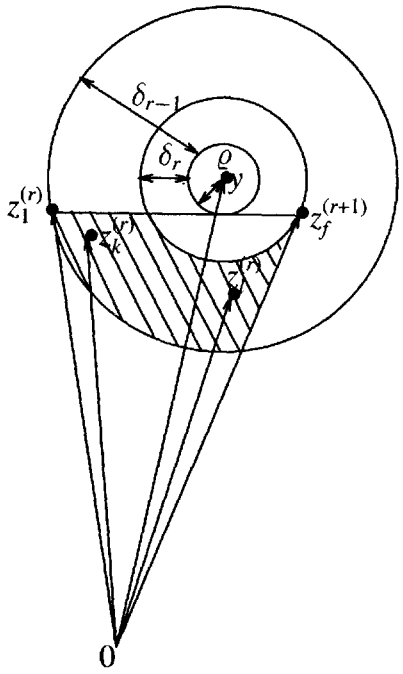

Figure 3.7. Geometrical interpretation of the sequences used in the internal convergence

Also, for arbitrary values of a_1 and a_2 we can have the relation $\|y - a_1 z_k^{(r)} - a_2 z_l^{(r)}\| \geq \varrho$. From the above inequalities, one can obtain the estimate of (Figure 3.7)

$$\|z_k^{(r)} - z_l^{(r)}\| \leq \sqrt{(\varrho + \delta_{r-1})^2 - \varrho^2} + \sqrt{(\varrho + \delta_r)^2 - \varrho^2}. \tag{3.196}$$

The right side quantity of the inequality can become arbitrarily small for a suitable r. This completes the proof for the internal convergence of the algorithm to the solution and in structure.

Defining multilayer algorithm with propagating errors

The function g has the form

$$g(x_i, x_j) = x_i + a x_j, \quad i = 1, 2, \cdots, F, \quad j = F+1, F+2, \cdots, Q, \tag{3.197}$$

where $Q = F + m$ and a is determined by the least-squares method using the set A or W.

Here the algorithm is described in its simplest way. In the first step the partial descriptions of the form

$$z_i = a x_i, \quad i = 1, 2, \cdots, m \tag{3.198}$$

of which the residual errors are computed as $\Delta^{(1)} z_{ij} = z_i - y$ and F best of the descriptions are chosen.

In the rth step ($r > 1$), the partial descriptions of the form

$$z_{ij}^{(r+1)} = z_i^{(r)} + a_j x_j, \quad i = 1, 2, \cdots, F, \quad j = 1, 2, \cdots, m \tag{3.199}$$

of which the residuals are computed as $\Delta^{(r+1)}z_{ij} = z_{ij}^{(r+1)} - z_i^{(r)}$, and F best models are chosen.

The process continues until the value of the criterion decreases significantly. Suppose we are required to reproduce a dependence of the form $y = h(x) + \xi$. The approximation is achieved as

$$h(x) = g_1(x) + g_2(x) + \cdots, \tag{3.200}$$

where $g_i(x)$, $i = 1, 2, \cdots$ correspond to the chosen equations at each step.

The internal convergence of the algorithm to the solution and in the structure is ensured by the following theorem.

Theorem 4. Suppose y^* is the projection of y onto $L(X)$ and the criterion is computed on the set W. For any F number of partial descriptions, the sequence of vectors $z^{(r)}$ converges to y^* as $r \to \infty$. If the matrix X^TX is nonsingular, the model corresponding to the limiting vector coincides with the regression equation for y as function of x.

The proof of this theorem differs from the theorem 1 because the vectors $(y - z^{(r)})$ and $z^{(r)}$ are not orthogonal in this case. We shall follow the preceding scheme.

When $F = 1$, the sequence $\|y - z^{(r)}\|$ is nonincreasing and lower bounded; it is denoted by the limit ϱ. As per the step-by-step iterations in the algorithm, we have

$$\|y - z^{(r)}\| - \|y - z^{(r+1)}\| \leq \delta, \tag{3.201}$$

and we note that

$$z^{(r+1)} = z^{(r)} + ax_i = z^{(r)} + \frac{x_i x_i^T}{x_i^T x_i}(y - z^{(r)}). \tag{3.202}$$

From the above inequality, $\|x_i^T y - x_i^T z^{(r)}\| \leq \delta(2\|y - z^{(r)}\| - \delta)$, $i = 1, 2, \cdots, m$ can be obtained. Thus, the sequence $z^{(r)}$ has a limit τ.

The rest of the proof is analogous to the preceding one.