KYBERNETIKA — VOLUME 15 (1979), NUMBER 3

Classification of Linear Estimators

PAVEL KOVANIC

An attempt is made to find a system for classification of linear estimators based only on mean values and covariances. As a tool for such classification a generalized estimator is used called the minimum penalty estimator. It makes it possible to show connections between estimators seeming to be different and differences between estimators seeming to be related. A number of examples of linear estimators demonstrates the suitability of the approach.

1. INTRODUCTION

It has been shown by Swerling [1] that all results obtained via the linear filter theory and by modern estimation methods are special cases of results obtainable from the viewpoint of properly elaborated Gauss method of least squares. Some of such special cases are mentioned in [1] explicitly: they include all problems in optimum linear filtering or prediction of random processes including recursive solutions. But literature of the field and of related fields brings extremely great number of elaborations of Gauss method of least squares, of its different extensions, modifications and generalizations (see e.g. [2]). This fact strengthens Swerling's statement as well as makes it more difficult to orient oneself in particular special cases. It is therefore worth trying to look out approaches covering as many as possible of particular cases. This paper is devoted to an attempt to use the minimum penalty estimator [3], [4] as a means for an unifying view on linear estimates based only on first two statistical moments.

2.1. Model

Observed data form an N × P random matrix

$$Y = Y_x + Y_e \tag{1}$$

where Y_x represents random signals and Y_e noise components. Mean values and second order statistical moments are supposed to be known a priori. A possible nonzero mean value $\langle Y_e \rangle$ of the noise matrix Y_e might be included into the signal matrix Y_x, therefore one is free to take it equalling zero. Noise is uncorrelated with signals. Knowledge of covariances makes it possible to evaluate matrices $\langle \hat{Y}_x Q \hat{Y}_x^T \rangle$ and $\langle Y_e Q Y_e^T \rangle$ for a given positive definite symmetrical weighting matrix Q. (The symbol $\hat{Y}_x$ states for the centered variable $Y_x - \langle Y_x \rangle$.) For some purposes an another form of (1) is useful

$$Y = XA + Y_e \tag{2}$$

where X and A have dimensions N × L and L × P, respectively. The matrix X is a given nonrandom matrix chosen to satisfy equation

$$\langle X \hat{A} Q \hat{A}^T X^T \rangle = \langle \hat{Y}_x Q \hat{Y}_x^T \rangle = XX^T \tag{3}$$

so that

$$\langle \hat{A} Q \hat{A}^T \rangle = I \tag{4}$$

for the centered value $\hat{A} = A - \langle A \rangle$ of the coefficient matrix A. Denote the rank of the matrices

$$\mathrm{r}\{\langle \hat{Y}_x Q \hat{Y}_x^T \rangle\} = \mathrm{r}\{X\} = \mathrm{M} \tag{5}$$

and

$$\mathrm{r}\{\langle Y_e Q Y_e^T \rangle\} = \mathrm{S}\,. \tag{6}$$

Ranks S and M are not necessarily full,

$$\mathrm{S} \leqq \mathrm{N}\,, \tag{7}$$

$$\mathrm{M} \leqq \mathrm{N}\,. \tag{8}$$

An estimate having a general linear form

$$Z = WY + C \tag{9}$$

will be considered where W and C are some constant matrices of dimensions T × N and T × P, respectively. Required results of estimation are defined separately for

operations performed on pure signals

$$Z_x = \mathscr{L}_x\{Y_x\} \tag{10}$$

and for operations on noise

$$Z_e = \mathscr{L}_e\{Y_e\} \tag{11}$$

where $\mathscr{L}_x$ and $\mathscr{L}_e$ are some given linear operators. Required results of estimation when operating on actual observed data (1) are thus

$$Z_0 = Z_x + Z_e\,. \tag{12}$$

The problem under consideration is to find matrices W and C minimizing a criterional function of deviations of actual estimates Z (9) from required results of estimation. Among possible criteria of optimality the one called the penalty is suitable for the purpose of this paper.

2.2. The Penalty and Three Kinds of Errors Arising in Estimation

As a norm of a random matrix E we introduce a scalar quantity

$$\|E\| = [\mathrm{tr}\,\{\langle EQE^T\rangle\}]^{1/2}\,. \tag{13}$$

(This is the square root of the trace of the mean value of the quadratic form EQE^T.)

To evaluate the quality of results of estimation and to facilitate the discussion of different special cases we take into account three kinds of errors:

1) The error of transformation of pure signals

$$\|E_x\| = \|WY_x + C - Z_x\| \tag{14}$$

2) The error of transformation of signals contaminated by noise

$$\|E_0\| = \|W(Y_x + Y_e) + C - Z_0\| \tag{15}$$

3) The error arising in estimation by amplification of noise

$$\|E_e\| = \|WY_e\|\,. \tag{16}$$

It is necessary to note that the mean value is taken in (13) (and consequently in (14)–(16) as well) over the set of all realizations of the errors, i.e. over the set undefined fully before estimation, at the moment when matrices W and C should be calculated. Judgements on "future" behavior of random quantities must be therefore based on a priori known statistical characteristics (obtained from "past" observations) and on subjective factors or on an additional information. To incorporate the uncertainty of the particular significance of the error $\|E_x\|$ in relations to errors $\|E_0\|$

or $\|E_e\|$ a scalar quantity

$$p = \frac{p_0}{2p_0 + p_x} \|E_0\|^2 + \frac{p_x}{2p_0 + p_x} \|E_x\|^2 \tag{17}$$

called the penalty ([3], [4]) can be used. The weight p_0 is positive, the weight p_x has reasonable values when taken from the interval

$$-p_0 < p_x \leqq \infty . \tag{18}$$

Definition of the penalty (17) differs slightly from that of [3], [4] by the normalizing factor to ensure that (17) is finite for all values of p_x from the interval (18).

2.3. The Minimum Penalty Estimator

Minimum penalty estimator is the estimator minimizing the penalty (17). Under mentioned conditions it follows from [4] as a special case that (9) takes the form

$$Z = \langle Z \rangle + W(Y - \langle Y \rangle) \tag{19}$$

where

$$W = (s\langle \hat{Z}_e Q Y_e^T \rangle + \langle \hat{Z}_x Q \hat{Y}_x^T \rangle)(s\langle Y_e Q Y_e^T \rangle + \langle \hat{Y}_x Q \hat{Y}_x^T \rangle)^g \tag{20}$$

and where

$$s = p_0/(p_0 + p_x) \tag{21}$$

is the penalty factor. The matrix of the type K^g is the one-condition generalized inverse of the matrix K, i.e. a matrix satisfying the condition

$$KK^gK = K . \tag{22}$$

The estimate (20) exists always.

In (20) a generalized inverse of a matrix of the "big" size N × N appears. An equivalent of this formula can be developed as in [4] having form

$$\begin{aligned} W = W_x + W_e = \langle \hat{Z}_x Q \hat{Y}_x^T \rangle (X^T)^g [sX^gX + X^TB^gX]^g X^TB^g + \\ + \langle Z_e Q Y_e^T \rangle B^g[I - X(sX^gX + X^TB^gX)^g X^TB^g] \end{aligned} \tag{23}$$

where

$$B = \langle Y_e Q Y_e^T \rangle \tag{24}$$

and where I states for the unity matrix.

Two notes are in order here:

1) To obtain (23) from (20) one assumes that there is no subspace of the range-space $\mathscr{R}(X)$ of the matrix X within which no noise exists, i.e. one takes that

$$BB^gX = X \tag{25}$$

holds. In an opposite case it would be possible to treat the noise-free signals separately as may be shown analysing results of consideration a problem of a similar type [5].

2) In the limit case $s \to 0$ the same formula (23) can be used as shown in [4].

3. CLASSIFICATION OF LINEAR ESTIMATES

Formulae given above may be used for an attempt to introduce a classification of linear estimates based only on first and second statistical moments. Three aspects may be of interest here:

1) The class of required result of estimation

2) Usage of a priori information on signal components

3) The choice of the value of the penalty factor s in (23).

3.1. Classification According to Goals of Estimation

3.1.1. *Class A: Zero operation on the Noise*

This is the most usual case for which the most possible suppression of noise is required

$$Z_e \equiv 0 \tag{26}$$

and the second part of (23) denoted W_e is a zero operator.

3.1.2. *Class B: Zero Operation on the Signal*

In this case the goal of estimation is defined by a nonzero Z_e and by

$$Z_x = 0\,. \tag{27}$$

Consider an example: It is required to obtain a best estimate $\tilde{Y}_e$ of the noise part Y_e of observed data so that $Z_e = Y_e$. We obtain from (23)

$$\tilde{Y}_e = BB^g\left[I - X(sX^qX + X^TB^gX)^g X^TB^g\right] \hat{Y}\,. \tag{28}$$

This is the same result as if $\tilde{Y}_x$ would be estimated by a corresponding estimator W_x of the class A for $\hat{Z}_x = \hat{Y}_x$ and if the formula

$$\tilde{Y}_e = \hat{Y} - W_x\hat{Y} \tag{29}$$

would be applied as usually. It might seem therefore that it is not necessary to introduce the concept of the estimator W_e and of classes differing from A. But an expression of the type (29) cannot be found for a general type of Z_e although usefulness

of such generalization is obvious. Let us demonstrate it by a simple practical example of prediction of the noise component,

$$(Z_e)_t = (Y_e)_{t+\tau} . \tag{30}$$

Predicted value $(Y_e)_{t+\tau}$ is dependent on correlations of $(Y_e)_t$ with $(Y_e)_{t+\tau}$ but they do not play any role for W_x. Therefore, an expression similar to (29) does not exist in this case.

3.1.3. *Class C: Non-Zero Operations on Both Data Components*

In this case neither Z_x nor Z_e is zero matrix. An example: it is required to predict observed data.

3.2. Classification According to Usage of a priori Information

In our problem is the a priori information represented by the mean values $\langle Y_x \rangle$ of signal component and by covariances appearing in the expression $\langle \hat{Y}_x Q \hat{Y}_x^T \rangle$. There is no necessity of having anything more to evaluate the expression $\langle \hat{Z}_x Q \hat{Y}_x^T \rangle$ as for the considered case when all operators are linear the quantity $\hat{Z}_x$ is obtainable via the formula

$$\hat{Z}_x = L\hat{A} \tag{31}$$

where L is a given matrix. It is not essential for the classification if such a priori information is available or not but if it is used in estimating formulae or not.

3.2.1. *Class a: Estimators Making Use of a priori Information*

In this case formulae (19), (20) and (23) hold without any change.

3.2.2. *Class b: Estimators Based on no a priori Information*

This case may be considered to be a limit case of the previous one when the mean value $\langle Y_x \rangle$ is taken to equal zero and when variances of the signal components increase unlimitedly. Instead of (3) one has

$$\langle \hat{Y}_x Q \hat{Y}_x^T \rangle = d^2 X_b X_b^T . \tag{32}$$

where X_b is a fixed matrix and d^2 reaches in a limit the infinity. Formulae (19) and (23) take form

$$Z = WY \tag{33}$$

and for an arbitrary but finite s

$$W = L(X_b^T B^+ X_b)^+ X_b^T B^+ + \langle Z_e Q Y_e^T \rangle B^+ [I - X_b (X_b^T B^+ X_b)^+ X_b^T B^+] \tag{33}$$

where the well-known Moore-Penrose pseudoinverse K^+ is used instead of a general inverse K^g of a matrix K. Pseudoinverse will be used also in all cases considered below because of its uniqueness and additional favorable features.

3.3. Classification According to the Relative Penalty Factor

Consider the interval of s corresponding to (21)

$$0 < s \leqq \infty \,. \tag{35}$$

There are three cases of special interest relating to this interval:

1) $s \to \infty$

2) $s = 1$

3) $s \to 0_+$

Three errors mentioned in Chapt 1. are functions of the factor s. Three cases considered here are connected with extremal values of some of the errors. Therefore, a fourth case may be included:

4) Other values of s

This classification is not usable for the estimates of class b for which the factor s does not play a role.

3.3.1. *Class 1: Zero Estimator*

For an s approaching the infinity all terms of (23) vanish excepting one

$$W = \langle Z_e \underline{Q} Y_e^T \rangle \, B^+ \,. \tag{36}$$

Note that this case differs essentially from the class B which should suppress the signal. Estimating according to class 1, one wants to obtain an non-zero estimate of the signal but one has no subjective confidence in "new" data. Therefore the best estimate of the signal is its a priori ("old") mean value, as results from (19) for the considered case in which

$$W_x \to 0 \,. \tag{37}$$

The error $\|E_e\|$ reaches its minimum value.

3.3.2. *Class 2: Unconditional Estimators*

The error $\|E_x\|$ is fully ignored in this case,

$$p_x = 0 \tag{38}$$

and the error $\|E_0\|$ reaches its minimum as shown in [4].

3.3.3. *Class 3: Conditional Estimators*

Subjective weight $p_x/p_0 \to \infty$ is given to the error $\|E_x\|$ in this case. As shown in [4], it is equivalent to the requirement of constrained minimalization of $\|E_0\|$ under condition that $\|E_x\|$ reaches its minimum. It is clear that an unconstrained minimum $\|E_0\|_{s=1}$ taking place for the class 2 can be smaller than the constrained one $\|E_0\|_{s\to 0_+}$ of the class 3.

3.3.4. *Class 4: Other Estimators*

This class includes all cases of estimators for which s differs from 1 and from both bounds of its interval. Single error of no type is minimized in this case but a compromise solution is possible of a particular problem when both error $\|E_x\|$ and $\|E_0\|$ are of importance. One example will be mentioned later in connection with the ridge regression.

4. DISCUSSION ON THE CHOICE OF THE PENALTY FACTOR

The choice of the relative penalty factor s is the matter of subjective approach always. Its role has been explained by its consequences on each of three estimating errors. But it is useful to take into account further explanation.

The penalty introduced by (17) as

$$p = \frac{s}{1+s}\|E_0\|^2 + \frac{1-s}{1+s}\|E_x\|^2 \tag{39}$$

may be presented also in an equivalent form

$$p = \frac{1}{1+s}\|E_x\|^2 + \frac{s}{1+s}\|E_e\|^2 . \tag{40}$$

The minimum penalty estimator minimizes the penalty for each particular value s. But we see that this minimal penalty value equals to $\|E_x\|_{s\to 0}$ for $s \to 0$ (class 3), to $\frac{1}{2}\|E_0\|_{s=1}$ for $s = 1$ (class 2) and to $\|E_e\|_{s\to\infty}$ for $s \to \infty$ (class 1).

It has been already mentioned that there are two different sets of realizations of random variables under consideration. The first one ("old") is used to evaluate mean values and covariances for calculations of the matrices W and C: The second set of realizations includes values which variables attain during applications of these matrices i.e. during estimation. This "new" set relates thus to observations "future" in relation to the moment when the choice of s should be made. A question therefore must be answered: will the statistical characteristics of the "future" set equal to that of the "old" set? If there is no confidence in stationarity of processes it is necessary to choose the estimator of the class "a". But consider case when one has reasons to suppose that mean values of processes are constant but that the matrix $X^T B^+ X$

determined from the "old" set of observations will change so that the "future" value of this matrix will equal to $k^2 X^T B^+ X$ where k^2 is a scalar. The estimator (23) will change only in that instead of s the value s/k^2 will take place. The quantity k^2 is related with anything like a "signal-to-noise" ratio. For $k^2 = 1$ we take a "conservative" point of view: The signal-to-noise in the future will be the same as it was in the past. For $k^2 > 1$ and $k^2 < 1$ we take an optimistic and pesimistic point of view, respectively. We have already mentioned that the estimating error $\|E_0\|$ is minimal for $s = 1$. Thus, the class 2 of estimators is connected with a conservative expectation on processes. The class 1 represents an extremaly pesimistic point of view, while the class 3 corresponds to an unlimited optimism: furure signal-to-noise ratio will be much better than it has been up to now.

For stationary conditions, unchanged measuring techniques and so on, the "conservative" class 2 may be quite reasonable especially because of the minimal estimating error $\|E_0^2\|$.

Another comment on the factor s is mentioned below in connection with Swerling's estimator.

5. CLASSIFICATION OF PARTICULAR TYPE OF ESTIMATORS

5.1. Gauss-Markov Estimator and its Generalizations

The well-known Gauss-Markov estimator exists if it is possible to find such matrix W that

$$\|E_x\| = 0 . \tag{41}$$

If it does not exist then a generalization of Lewis and Odell [6] can be used according to which the error $\|E_x\|$ should be made as small as possible. The Lewis-Odell estimator is thus classified as an estimator of the type Ab for which

$$L = I . \tag{42}$$

5.2. Discrete Zadeh-Ragazzini (Blum's) Estimator

Zadeh and Ragazzini generalized in 1950 the Wiener problem of filtering and prediction of continuous signals. Blum [7] gave the solution of a discrete analogue of Zadeh-Ragazzini problem. In our notation is the Blum's estimator of class Ab. An extension (which can be characterized as Ca 3) is described in [8].

5.3. Discrete Semyonov Estimator

As mentioned in [9] Semyonov (1954) formulated and solved an another generalization of Wiener problem for continuous variables. A discrete version of Semyonov estimator has been given in [9] which can be classified as Aa2.

5.4. Random Coefficient Regression

There is a lot of statistical results on this subject reviewed in [2]. Classification is again Aa2.

5.5. Goodman's Estimator

According to [10] the a priori information represented by means and covariances of coefficients of a linear model can be incorporated into least squares estimation. The quadratic form to be minimized includes not only squares of estimating errors but also squares of deviations of the new estimated coefficients from their a priori estimates weighted by an inverse of their a priori covariance matrix. Goodman's formula seems to be of similar form as estimators Aa2 but actually it corresponds to the class Aa3 if the variables have zero mean values and Ab if they have not.

5.6. Discrete Version of Karhunen-Lôewe Expansion

For the solution of a prediction problem a discrete Karhunen-Lòewe expansion has been applied in [11]. The idea is to find eigenvectors of the covariance matrix of the "old" observed vectors, to fit a part of them into the last data using the least squares method and to a best linear combination of the ends of fitted eigenvectors to be estimates of future data. An analysis shows that this case is again Aa2.

5.7. Swerling's Estimator and Discrete Filters

The review of linear estimators in [1] relates to estimators of the class Aa2. As the criterion of optimality a quadratic form has been used including squares of reziduals as well as the squares of deviations of estimates from their a priori means weighted by the inverse of a priori covariance matrix. Introduce a subjective weighting factor to give to each of both parts of Swerling's quadratic form a weight. Then an additional explanation of the role of the choice of the relative penalty factor is obtained: Increasing the factor s in (20), one decreases the weight given to a priori information. (Formula (20) is a generalization of the case considered in [1]).

5.8. Ridge Regression

In [12] an biased estimator of Hoerl and Kennard is further investigated having in our notation the form

$$Z = \tilde{A} = (\mathrm{k}I + X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y \tag{43}$$

where $\tilde{A}$ is the estimate of A and k is a positive number. This estimator has been introduced because it was found that it lead to increased accuracy of estimation. Note that this estimator would be unbiased if the data matrix Y would have zero mean. In such case the ridge regression would be the case Aa4.

6. CONCLUSION

It has been shown that the formulae of the minimum penalty estimator is general enough to show the connections between seemingly different estimators, to introduce a classification of a number of estimators and to point out some new estimators.

(Received October 4, 1977.)

REFERENCES

[1] P. Swerling: Modern estimation methods from the viewpoint of the method of least squares. IEEE Trans. on AC, *AC-16*, (1971) No. 6.

[2] P. A. V. B. Swammy: Statistical inference in random coefficient regression models. (Lecture notes in operations research and mathematical systems, vol. 55.) Springer-Verlag, Berlin—Heidelberg—New York 1971.

[3] P. Kovanic: Minimum penalty estimate, Kybernetika *8*, (1972), 5, 367—383.

[4] P. Kovanic: Generalized linear estimate of functions of random matrix arguments, Kybernetika *10* (1974), 4, 303—316.

[5] C. R. Hallum, T. O. Lewis, T. L. Boullion; Estimation in the restricted general linear model with a positive semidefinite covariance matrix. Communications in statistics, *1* (2), (1973), 157—166.

[6] T. O. Lewis, P. L. Odell: A generalization of the Gauss-Markov theorem. J. Am. Stat. Assoc. *61* (1966), 1063—1066.

[7] M. Blum: An extension of the minimum mean square prediction theory for sampled input signals. IRE Trans., *IT*-2 (1956), 176.

[8] P. Kovanic: Generalized discrete analogy of the Zadeh-Ragazzini problem. Automation and Telemechanics (In Russian) *XXVII* (1966), 2, 37.

[9] I. D. Krutko: Statistical dynamics of impulse systems (In Russian). Sovetskoe radio, Moskva 1963.

[10] A. F. Goodman: Extended iterative weighted least squares: Estimation of a linear model in the presence of complications. Naval research logistics quarterly *18* (1971), 2, 243—276.

[11] E. D. Farmer: A method of prediction for nonstationary processes and its application to the problem of load estimation. Transactions of IFAC 1963.

[12] K. S. Banerjee, R. N. Carr: A comment on ridge regression. Biased estimation for nonorthogonal problems. Technometrics *13* (1971), No. 4.

Ing. Pavel Kovanic, CSc., Ústav teorie informace a automatizace ČSAV (Institute of Information Theory and Automation — Czechoslovak Academy of Sciences), Pod vodárenskou věží 4, 182 08 Praha 8. Czechoslovakia.