METROLOGY AND MEASUREMENT SYSTEMS
VOL. XIV, NUMBER 2 (2007)

MAGDALENA BOĆKOWSKA, ADAM ŻUCHOWSKI

Szczecin University of Technology
Poland
e-mail: Magdalena.Bockowska@ps.pl

AN OPTIMAL DEGREE OF COMPLEXITY OF A SIMPLIFIED MODEL

A simplified model can be created using one of the already known mathematical methods (series expansion of a function, truncating continued fraction expansion [1] and so forth) if the analytic form of the full model is known or using regression to determine the optimal parameters of the model with arbitrary assumed structure. If the measured accuracy of input variables can be obtained, the optimal degree of complexity of the model can be determined.

Keywords: simplified model, model complexity

1. INTRODUCTION

A phenomenon model which links the input variables x_i for $i = 1, 2, \ldots, n$ with an output variable y can be given in the form $y(x_i)$, too complicated for practical applications. In this case a simplified model can be created in the form of different functions $y_{m1}(x_i)$, $y_{m2}(x_i)$, ..., $y_{mr}(x_i)$. The degree of model complexity and its accuracy increase with the growth of the degree expansion r. When a phenomenon is described by a set of experimental results y_j for $j = 1, 2, \ldots, m$ obtained for a range of input variables $(x_1, x_2, \ldots, x_n)_j$ then the model structure can be assumed arbitrary as a function $y_m(x_i, p_1, p_2, \ldots, p_r)$, where p_k for $k = 1, 2, \ldots, r$ are the model parameters obtained by regression basing on the measured results [2, 3]. The mean square error of the simplified model defined by:

$$D^2 = \int \int \int \{y_{mr}(x_i) - y(x_i)\}^2 dx_1 dx_2 \ldots dx_i, \tag{1}$$

decreases with an increasing r and a well chosen structure of the simplified model.

Taking into consideration that input variables x_i are given with the determined accuracy Δx_i the error is expressed by equation:

$$D^2 = \int \int \ldots \int \{y_{mr}(x_i, \Delta x_i) - y(x_i)\}^2 dx_1 dx_2 \ldots dx_i. \tag{2}$$

Assuming that:

$$y_{mr}(x_i, \Delta x_i) = y_{mr}(x_i) + dy_{mr}(x_i) \tag{3}$$

and

$$dy_{mr}(x_i) = \sum_{i=1}^{n} \Delta x_i \frac{\partial y_{mr}(x_i)}{\partial x_i}, \tag{4}$$

omitting the higher derivatives one obtains:

$$\begin{aligned} D^2 = & \int \int \dots \int \{y_{mr}(x_i) - y(x_i)\}^2 dx_1 dx_2 \ \dots \ dx_i + \\ & 2 \sum_{i=1}^{n} \Delta x_i \int \int \dots \int \{y_{mr}(x_i) - y(x_i)\} \frac{\partial y_{mr}(x_i)}{\partial x_i} dx_1 dx_2 \ \dots \ dx_i + \\ & \sum_{i=1}^{n} \Delta x_i^2 \int \int \dots \int \left\{ \frac{\partial y_{mr}(x_i)}{\partial x_i} \right\}^2 dx_1 dx_2 \ \dots \ dx_i. \end{aligned} \tag{5}$$

The first component of the above formula corresponds to the relation (1). If the error increases or stays on the same level in the event of the rise of the parameter's number from r to $r+1$ for a given accuracy of the measurement, then the change of the model structure is not expedient and the value r is the optimal degree of the simplified model complexity.

2. BASIC ASPECTS OF THE PROBLEM

As mentioned earlier, the simplified model can be created either by a simplification of the function form $y(x_i)$, too complicated for practical applications, or by a regression for an arbitrary chosen structure $y_m(x_i, p_1, p_2, \dots, p_r)$. Moreover, the deviations Δx_i can be $|\Delta x_i| = \text{const}$ for all of variables x_i, or proportional to their own value x_i, i.e. $|\Delta x_i / x_i| = \text{const}$. Thus, four basic cases can be considered.

The first one assumes that the simplified model is created by a simplification of the function form $y(x_i)$ without the free parameters $p_1, p_2, \dots p_r$. If $|\Delta x_i| = \text{const}$, then the Eq. (5) should be used taking the absolute value of the second integral – determining the maximum value of the error which can occur.

If $|\Delta x_i/x_i| = \text{const}$, then the formula (5) should be written in the form:

$$\begin{aligned} D^2 &= \int\int \ldots \int \{y_{mr}(x_i) - y(x_i)\}^2 \, dx_1 dx_2 \ \ldots \ dx_i + \\ &2\sum_{i=1}^{n} \frac{\Delta x_i}{x_i} \int\int \ldots \int x_i \{y_{mr}(x_i) - y(x_i)\} \frac{\partial y_{mr}(x_i)}{\partial x_i} dx_1 dx_2 \ \ldots \ dx_i + \\ &\sum_{i=1}^{n} \left(\frac{\Delta x_i}{x_i}\right)^2 \int\int \ldots \int \left\{x_i \frac{\partial y_{mr}(x_i)}{\partial x_i}\right\}^2 dx_1 dx_2 \ \ldots \ dx_i. \end{aligned} \tag{6}$$

In case of determination of the simplified model by regression one should obtain the optimal values of the simplified model parameters $p_1, p_2, \ldots, p_r$ minimizing the functional:

$$D^2(p_1, p_2, \ldots, p_r) = \int\int \ldots \int \{y_{mr}(x_i, p_1, p_2, \ \ldots, p_r) - y(x_i)\}^2 \, dx_1 dx_2 \ \ldots \ dx_i, \tag{7}$$

satisfying the following conditions:

$$\begin{aligned} &\frac{\partial D^2(p_1, p_2, \ \ldots, p_r)}{\partial p_i} = 0 = \\ &2\int\int \ldots \int \{y_{mr}(x_i, p_1, p_2, \ \ldots, p_r) - y(x_i)\} \frac{\partial y_{mr}(x_i, p_1, p_2, \ \ldots, p_r)}{\partial p_i} dx_1 dx_2 \ \ldots \ dx_i, \end{aligned} \tag{8}$$

for $i = 1, 2, \ldots, r$. The system of equations allows one to obtain the parameters of the examined simplified model. Its error can be determined on the basis of the Eq. (5) or (6) depending on the values of deviations Δx_i.

In the event of the simplified model which satisfies the conditions:

$$\frac{\partial y_{mr}(x_i, p_i)}{\partial p_i} = \frac{\partial y_{mr}(x_i, p_i)}{\partial x_i} c_i, \quad c_i = const, \tag{9}$$

the second integral in the formula (5) can be omitted. If the simplified model fulfils the following conditions:

$$\frac{\partial y_{mr}(x_i, p_i)}{\partial p_i} = x_i \frac{\partial y_{mr}(x_i, p_i)}{\partial x_i} c_i, \tag{10}$$

the second integral in the formula (6) can be omitted, what significantly facilitates the analysis. The same situation occurs in case of applying regression using the simplified model with a polynomial form:

$$y_{mr}(x, p_i) = p_0 + p_1 x + p_2 x^2 + \ldots + p_r x^r. \tag{11}$$

The model parameters are obtained by solving the system of equations:

$$\int_{x_1}^{x_2} \{y_{mr}(x, p_i) - y(x)\} x^i dx = 0, \quad i = 0,\ 1,\ 2,\ ...,\ r. \tag{12}$$

The second component of the formula (5) has the form:

$$2\Delta x \int_{x_1}^{x_2} \{y_{mr}(x, p_i) - y(x)\} \left(p_1 + 2p_2 x + ... + r p_r x^{r-1}\right) dx \tag{13}$$

and for the relation (6):

$$2\left(\frac{\Delta x}{x}\right) \int_{x_1}^{x_2} \{y_{mr}(x, p_i) - y(x)\} \left(p_1 x + 2p_2 x^2 + ... + r p_r x^r\right) dx, \tag{14}$$

because

$$\frac{\partial y_{mr}(x, p_i)}{\partial x} = p_1 + 2p_2 x + ... + r p_r x^{r-1}. \tag{15}$$

Taking into account that the system (12) is fulfilled, both integrals are equal to zero. An identical effect occurs for the simplified model, which is a result of multiplication of polynomials for every variable x_i.

The integrals for the measured data should be substituted by equivalent sums. The determined simplified models can be compared with each other if all of them were obtained for the same set of the experimental data.

3. SIMPLE EXAMPLES

At first it has been assumed that the ideal model of the relation $y(x)$ has the form $y(x) = \exp(x)$. The range of the input variable x was (0, 1) and its accuracy was Δx or $\Delta x/x$.

The simplified model was searched using the expansion of the function $\exp(x)$ in Taylor's series:

$$y_{m1}(x) = 1,\ y_{m2}(x) = 1 + x,\ y_{m3}(x) = 1 + x + x^2, \ldots \tag{16}$$

and by the application of the regression method, which allows to determine the optimal values of the parameters p_i for the assumed structures:

$$y_{m1}(x) = p_0,\ y_{m2}(x) = p_0 + p_1 x,\ y_{m3}(x) = p_0 + p_1 x + p_2 x^2, \ldots \tag{17}$$

The results of the calculations are shown in Table 1. As shown in the case of the simplified models, for the same complexity degree, by the expansion in Taylor's series the errors are much bigger as compared to those obtained by the regression method. For $\Delta x \geq 0.19$ or $\Delta x/x \geq 0.12$ the complexity degree of the model $y_{m2}(x)$ created by regression can be admitted as the optimal one.

Table 1. Forms of the simplified models and their errors for the first example.

Model $y_{mr}(x)$	$D^2(\Delta x)$	$D^2\left(\frac{\Delta x}{x}\right)$
1	0.785	0.785
$1 + x$	$0.0913 + 0.437\,\lvert\Delta x\rvert + \Delta x^2$	$0.0913 + 0.667\left\lvert\frac{\Delta x}{x}\right\rvert + 0.333\left(\frac{\Delta x}{x}\right)^2$
$1 + x + 0.5x^2$	$0.0063 + 0.187\,\lvert\Delta x\rvert + 2.33\Delta x^2$	$0.0063 + 0.153\left\lvert\frac{\Delta x}{x}\right\rvert + 0.583\left(\frac{\Delta x}{x}\right)^2$
$\exp(x)$	$3.195\Delta x^2$	$0.718\left(\frac{\Delta x}{x}\right)^2$
Regression method		
1.718	0.242	0.242
$0.873 + 1.690x$	$0.0038 + 2.857\Delta x^2$	$0.0038 + 0.952\left(\frac{\Delta x}{x}\right)^2$
$1.013 + 0.851x + 0.839x^2$	$0.000029 + 3.092\Delta x^2$	$0.000029 + 1.52\left(\frac{\Delta x}{x}\right)^2$

The second example concerns the function $y(x) = (x+1)!$. The models were created by regression using the auxiliary Stirling formula:

$$(x+1)! = (x+1)^{(x+1)} \exp[-(x+1)] \quad \sqrt{2\pi(x+1)} \exp\left(\frac{1}{12(x+1)}\right). \tag{18}$$

The way of expansion in Taylor's series is here too complicated, hence its results are not presented. Table 2 contains a set of the optimal parameters for four simplified models:

$$\begin{aligned} &y_{m1}(x) = a_0,\ y_{m2}(x) = a_0 + a_1 x,\ y_{m3}(x) = a_0 + a_1 x + a_2 x^2, \\ &y_{m4}(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 \end{aligned} \tag{19}$$

and for the range of variable $x = (0, 2)$.
The model errors D^2 for all the models are put together in Table 3.
For the assumed accuracy the model $y_{m4}(x)$ is too complicated, but $y_{m3}(x)$ can be accepted as optimal.

Table 2. Optimal values of the simplified model parameters for the second example.

Model	Optimal values of parameters			
	a_0	a_1	a_2	a_3
$y_{m1}(x)$	2.4654			
$y_{m2}(x)$	0.2428	2.2226		
$y_{m3}(x)$	1.2150	−0.6940	1.4583	
$y_{m4}(x)$	0.9527	0.8807	−0.5108	0.6565

Table 3. Errors D^2 of the simplified models created in the second example.

Model	Error D^2	Error for $\Delta x = 0.1$ and $\Delta x/x = 0.05$
$y_{m1}(x)$	3.6949	3.6949
$y_{m2}(x)$	$0.4002 + 9.8800\Delta x^2$ $0.4002 + 13.1700(\Delta x/x)2$	0.499 0.433
$y_{m3}(x)$	$0.0209 + 15.550\Delta x^2$ $0.0209 + 39.530(\Delta x/x)^2$	0.176 0.120
$y_{m4}(x)$	$0.0008 + 18.720\Delta x^2$ $0.0008 + 54.160(\Delta x/x)^2$	0.188 0.136

4. CONCLUSIONS

The presented examples contain simple models which are functions of one variable, but nevertheless prove that the optimal complexity degree of the simplified model can be determined for a known accuracy of input variables. Its determination is important on account of the fact that the measurements are always linked with disturbances and usually – despite of the application of different regularization methods [4–8] – the accuracy of less important terms decreases. Hence, it is worth determining the errors which are introduced by any of the model terms. The terms characterized by low accuracy must be rejected from the model structure. The results of analysis depend on the manner of the determination of accuracy of input variables and how their accuracy is determined by a constant absolute or relative error or otherwise. The comparison of models with different structures and parameters is purposeful when all of them were determined for the same range of input variables. The kind of the simplification method has substantial influence on the analysis results and choice of the optimal complexity degree of the simplified model.

REFERENCES

1. Sierpiński W.: *Number theory. Mathematic monographies*. Warszawa-Wrocław. Drukarnia Uniwersytetu i Politechniki Wrocławskiej 1950. (in Polish)

2. Halawa J.: *Methods of determination of simplified transfer functions and their application in automatic and electroenergetic*. Politechnika Wrocławska, Instytut Cybernetyki Technicznej. Prace naukowe 1991. Seria Monografie Nr 21. (in Polish)
3. Layer E.: *Modelling of Simplified Dynamical Systems*. Berlin, Springer Verlag 2002.
4. De Larminant P., Thomas Y.: *Automatic – linear systems. Vol. 2. Identification*. Warszawa, WNT 1983. (in Polish)
5. Hoerl A. E.: *Application of rigde analysis to regression problems*. Chem. Eng. Progress 58, pp. 54–59, 1962.
6. Hoerl A. E., Kennard W. R.: *Rigde regression: Biased estimation for non-orthogonal problems*, Technometrics, vol. 12, pp. 55–68, 1970.
7. Marquardt D. W.: *An algorithm for least squares estimation of nonlinear parameters*. J. SIAM 11, pp. 431–441, 1963.
8. Tikhonov A. N., Arsenin Y. V.: *Solutions of ill-posed problems*. Winston, Washington, D. C., 1977.