

Some Fundamental Topics of Linear Modelling

Yuriy V. Dzyadyk

International Center of Information Technologies and Systems, Acad. Glushkov av. 40, Kyiv, 03680 Ukraine

iurius@i.com.ua

Abstract. *This paper consider the next various topics. (1) Stabilization Principle in Linear Modelling. (2) Factor Analysis and Stabilization. Method of Two Thresholds (MTT), or (β, γ) -Method. (3) The essence of GMDH. (4) Economic Criterium (5) Active Agent Models. Cycles of Activization and Stabilization. On some examples we demonstrate an advantage of the (β, γ) -method over some well-known methods.*

Keywords

Inductive modelling, ICIM 2008, stabilization principle, factor analysis, (β, γ) -method, economic criterium, multiagent systems, active models

1 Introduction

Let we build an inductive model $y = f(t_1, t_2, \dots, t_i)$, which is linear relative to basic functions $\{x_1, x_2, \dots, x_k\}$, where $x_i = \varphi_i(t_1, t_2, \dots, t_i)$. Let n be a dimension of statistics. We obtain vector \mathbf{y} and k vectors $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k\}$ of a real n -dimensional space. Let us denote

$$\mathbf{X} = (\mathbf{x}_1 - \mu_1, \mathbf{x}_2 - \mu_2, \dots, \mathbf{x}_k - \mu_k), \text{ where } \mu_i = \mu(\mathbf{x}_i) = \frac{1}{n} \sum_{t=1}^n x_{it}. \quad (1)$$

2 Problem

Let we search for a form \mathbf{A} of the linear dependence $\mathbf{y} = \mathbf{X}\mathbf{A}$, where vector \mathbf{y} and matrix \mathbf{X} are known, matrix \mathbf{A} is sought. As well known, the heuristic (or symbolic) way

$$\mathbf{y} = \mathbf{X}\mathbf{A} \Rightarrow \mathbf{X}^T \mathbf{y} = \mathbf{X}^T \mathbf{X} \mathbf{A} \Rightarrow (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \mathbf{A} \quad (2)$$

leads to the same result that least-squares method (LSM):

$$\mathbf{A} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}. \quad (3)$$

Denote $\mathbf{X}^T \mathbf{X} = \mathbf{Z}$. Let us investigate the expression (3). It has no sense if $\det(\mathbf{Z}) = 0$, i.e. condition number of matrix \mathbf{Z} , $\text{cond}(\mathbf{Z}) = \|\mathbf{Z}\| \cdot \|\mathbf{Z}^{-1}\| = \infty$. Moreover, if $\text{cond}(\mathbf{Z})$ is near to infinity, the solution $\mathbf{y} = \mathbf{X}\mathbf{A}$ is worthless for extrapolation or forecasting. The obvious example: the exact polynomial interpolation model has no sense beyond of interval of interpolation. Note that for any reasonable norm $\|\cdot\|$ we have:

$$\text{cond}(\mathbf{Z}) = |\lambda_1 \lambda_n^{-1}|, \det(\mathbf{Z}) = \lambda_1 \lambda_2 \dots \lambda_n \quad (4)$$

where $\{\lambda_1, \lambda_2, \dots, \lambda_n\}$ is the set of all eigenvalues of \mathbf{Z} which are numbered so that $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_n|$. Obviously, for any constant k , $\text{cond}(k\mathbf{Z}) = \text{cond}(\mathbf{Z})$. Note that all λ_i for Gram matrix $\mathbf{Z} = \mathbf{X}^T \mathbf{X}$ are real nonnegative: $\forall i \lambda_i \geq 0$.

Define the *measure of stability* of the matrix \mathbf{X} as

$$\text{stab}(\mathbf{X}) = \lambda_n \text{Tr}(\mathbf{X}^T \mathbf{X})^{-1} = \lambda_n (\lambda_1 + \lambda_2 + \dots + \lambda_n)^{-1}. \quad (5)$$

Obviously, for any constant k , $\text{stab}(k\mathbf{X}) = \text{stab}(\mathbf{X})$. We have $\text{stab}(\mathbf{X}) \text{cond}(\mathbf{X}^T \mathbf{X}) = \lambda_1 (\lambda_1 + \lambda_2 + \dots + \lambda_n)^{-1}$, and $\lambda_1 \leq \lambda_1 + \lambda_2 + \dots + \lambda_n \leq n\lambda_1$, so

$$1/n \leq \text{stab}(\mathbf{X}) \text{cond}(\mathbf{X}^T \mathbf{X}) \leq 1, \text{ or } \text{stab}(\mathbf{X}) \cong [\text{cond}(\mathbf{X}^T \mathbf{X})]^{-1}. \quad (6)$$

The main problem is – how to avoid inanity, insignificance of the model $\mathbf{y} = \mathbf{X}\mathbf{A}$ beyond the neighbourhood of its construction domain?

Our working hypothesis is – to increase *stability*, or, the same, to decrease condition number.

3 Factor Analysis and Stabilization. Method of Two Thresholds (MTT), or (β, γ) -Method

Let reduce the Gramian matrix $\mathbf{X}^T \mathbf{X}$ by orthogonal transformation \mathbf{S} to diagonal form: $\mathbf{S}^T \mathbf{X}^T \mathbf{X} \mathbf{S} = \mathbf{D}$, so as to $d_{11} \geq d_{11} \geq \dots \geq d_{kk}$. Then vectors (columns) of the matrix $\mathbf{X}\mathbf{S} = \mathbf{Z} = (\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k)$ are named as factors. Note, that $\forall i: \mu(\mathbf{z}_i) = 0$.

By construction, first non-zero factors $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_p\}$ form an orthogonal basis of the linear enveloping space

$$L = L(\mathbf{x}_1 - \mu_1, \mathbf{x}_2 - \mu_2, \dots, \mathbf{x}_k - \mu_k), \text{ where } p = \dim L \leq \min(k, n). \quad (7)$$

Thus, an arbitrary linear model $\hat{y}(x_1, x_2, \dots, x_k)$ by transformation \mathbf{S} may be represented in the form of

$$\hat{y} = y_0 + y_1 z_1 + y_2 z_2 + \dots + y_p z_p; \quad (3)$$

note, that

$$\forall i, 0 < i \leq p, y_i = \frac{\langle \mathbf{y} - y_0, \mathbf{z}_i \rangle}{\langle \mathbf{z}_i, \mathbf{z}_i \rangle} = \frac{\langle \mathbf{y}, \mathbf{z}_i \rangle}{|\mathbf{z}_i|^2} = \frac{\langle \mathbf{y}, \mathbf{z}_i \rangle}{d_{ii}}. \quad (3)$$

Further, we obliterate so called *unstable* and *inessential* factors. Let's call factor \mathbf{z}_i as *unstable* in statistics \mathbf{X} (with the threshold β), if

$$\frac{|\mathbf{z}_i|^2}{\text{trace } \mathbf{D}} = \frac{d_{ii}}{\text{trace } \mathbf{D}} = \frac{d_{ii}}{d_{11} + d_{22} + \dots + d_{pp}} < \beta. \quad (3)$$

Obviously, there exists such $j \in \mathbb{N}$, that all *stable* factors form the set $\{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_j\}$.

Let's call factor \mathbf{z}_i as *inessential* for model \hat{y} (with the threshold γ), if correlation

$$\text{corr}(\mathbf{y}, \mathbf{z}_i) = \cos(\mathbf{y}, \mathbf{z}_i) = \frac{\langle \mathbf{y} - y_0, \mathbf{z}_i \rangle}{|\mathbf{y} - y_0| \cdot |\mathbf{z}_i|} < \gamma. \quad (3)$$

By stabilization with the threshold (β, γ) of a model \hat{y} we shall call a model

$$\hat{y}^s = y_0 + s_1 y_1 z_1 + s_2 y_2 z_2 + \dots + s_p y_p z_p, \quad (6)$$

where $\forall i > 0, s_i = s_i(\beta, \gamma) = 0$, if factor \mathbf{z}_i is unstable or inessential, and $s_i = 1$ for all other factors. In other words, stabilization is the obliteration of all unstable and inessential factors.

4 Stabilization Principle in Linear Modelling

Now, let $L = L(\mathbf{X})$ is the linear envelope of \mathbf{X} , V is any subspace of L , and \mathbf{P} is corresponding projection matrix, so that $\mathbf{U} = \mathbf{X}\mathbf{P}$ is the projection of \mathbf{X} to V .

We can substitute matrix \mathbf{X} by matrix $\mathbf{U}=\mathbf{X}\mathbf{P}$ and to repeat all previous definitions: $\text{cond}(\mathbf{U}^T\mathbf{U})$, $\text{stab}(\mathbf{U})$, ..., etc.

$$\hat{y}^s = y_0 + s_1 y_1 z_1 + s_2 y_2 z_2 + \dots + s_p y_p z_p, \quad (6)$$

We call subspace V as *stable*, if all factors of V are stable and essential.

Hypothesis. The essence of GMDH is the construction of some stable subspace in V .

5 Active Agent Models. Cycles of Activization and Stabilization

6 Some Results. Comparisons with Other Methods

The method of two thresholds was successfully used for modelling and forecasting of molybdenum and ferromolybdenum (Mo, FeMo) prices. Especially interesting was forecast of extremely unstable monthly prices in 2004–07.

For monthly moving forecasting of molybdenum prices in 2005-07, in Internet were selected 9 activities (t_1, t_2, \dots, t_9): t_1, t_2 – molybdenum export and import prices calculated from [4], t_3 – cuprum price [4], and (t_4, t_5, \dots, t_9) – steel prices [5].

From activities $\{t_1, t_2, \dots, t_9\}$ we form the dependent variable $y(\tau)$ and $n = 12$ input variables $\{x_1, x_2, \dots, x_{12}\}$:

$$\begin{aligned} y(\tau) &= t_1(\tau+1); \\ x_1(\tau) &= t_1(\tau-2), \quad x_2(\tau) = t_1(\tau-1), \quad x_3(\tau) = t_1(\tau), \\ x_4(\tau) &= t_2(\tau-1), \quad x_5(\tau) = t_2(\tau), \\ x_i(\tau) &= t_{i-3}(\tau), \quad i = 6..12. \end{aligned}$$

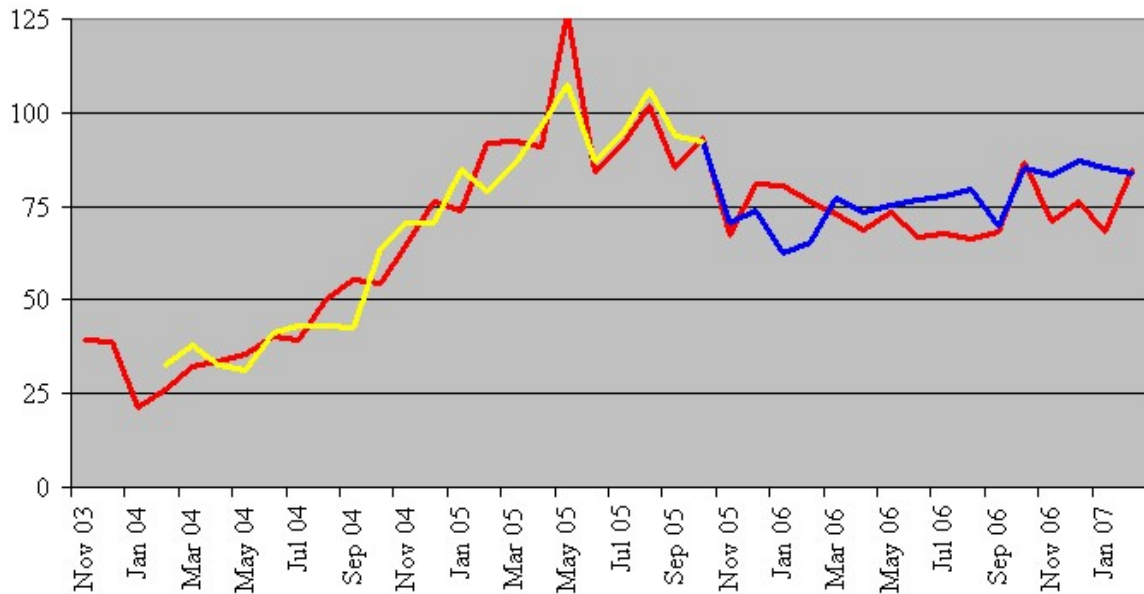


Fig. 1. Graphics of the actual values of molybdenum prices (red line), which were calculated from data given in tables [4], columns Imports, row Molybdenum, simulated (yellow line) and forecasted values (blue line).

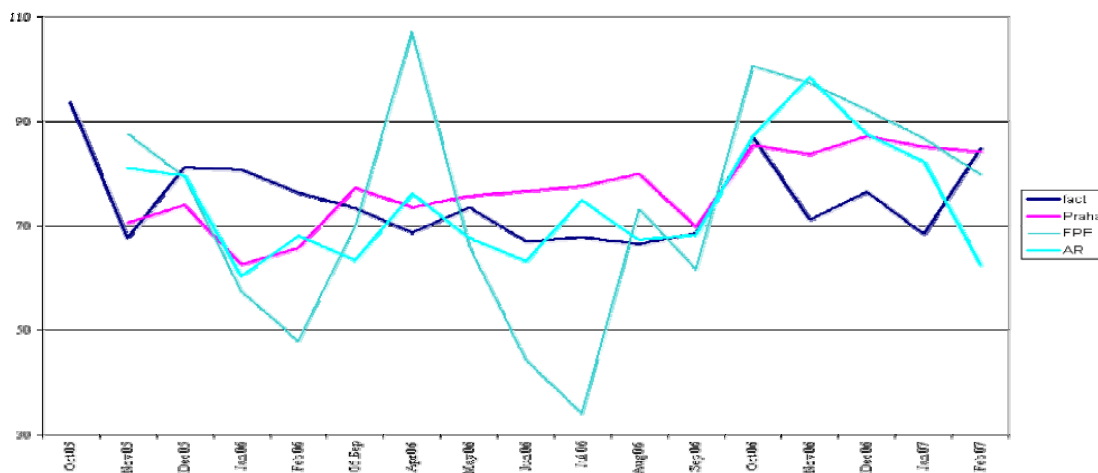


Fig. 2. Graphics of the actual values of molybdenum prices (line fact), and forecast values, computed by (β, γ) -method (line Praha) and GMDH (lines FPE and AR).

Tab.1. Comparison of the (β, γ) -method (MTT) with GMDH and GAME [6] on examples of molybdenum price forecasting (row Mo) in 2005-07 and forecast of CO₂ in brain [6] (row CO₂): RMS error of prediction

example\method	MTT	GMDH, FPE	GMDH, AR	GMDH, AC	GAME
CO ₂	0,0380	—	—	0,0704	0,0386
Mo	9,62	20,22	12,51	—	—

7 Economic Criteria

Let $y = f(\dots)$ be a price for some good. We can buy this good month by month, then expenses equal to Z_0 . On the end of period we can see the least price. If we buy all good by the least price, expenses equal to Z_{\min} . Of course, we are not omniscient. But if we have a forecast by method M , we can use it in some way W . Then we can compute corresponding expenses $Z_{M,W}$. It gives some types of economic criteria. E.g., relative criterion

$$(Z_0 - Z_{M,W}) / (Z_0 - Z_{\min}) \quad ()$$

for Mo price forecast by MTT equals 69%. Absolute economy for simulated plant which consumes 20 tons Mo per month equals to 1,93 mln \$.

References

- [1] Ivakhnenko A.G., Ivakhnenko G.A. The Review of Problems Solvable by Algorithms of the Group Method of Data Handling. *Pattern Recognition and Image Analysis*, 1995, vol.5, no.4, pp.527–535 – <http://www.gmdh.net/articles/review/algorithm.pdf>, <http://www.gmdh.net/articles/review/algorithm.zip/algorithm.doc>
- [2] Koppa Yu.V., Stepashko V.S.: A Comparison of the Forecasting Properties of Regression Types Models versus GMDH (pdf) (in Russian) – <http://www.gmdh.net/articles/rus/compare.pdf>
- [3] Gramian matrix – <http://www.answers.com/topic/gramian-matrix>
- [4] Metals Statistics, U.S. Metals Trade – <http://www.ita.doc.gov/td/metals/statindx.html>
- [5] MEPS (International) Ltd. – Independent Steel Industry Analysts, Consultants, Steel Prices, Reports and Publications. World Stainless Steel Product Prices – <http://www.meps.co.uk/Stainless Prices.htm>
- [6] Josef Bouška, Pavel Kordik. Time Series Prediction by means of GMDH Analogues Complexing and GAME – http://www.gmdh.net/articles/iwim/IWIM_39.pdf