

Some Results of the Synthesis of GMDH and Factor Analysis for Inductive Modelling

Yuriy V. Dzyadyk¹

¹International Center of Information Technologies and Systems of the NAS and MES of Ukraine,
40, Academician Glushkov avenue, 03680 CPO, Kyiv, Ukraine

iurius@i.com.ua

Abstract. It is known that GMDH generally requires building and comparison of 2^k linear models and choice of the best among them (above $k \lesssim 2^l$ is a number of all linear arguments selected among basic functions $1, t_i, t_i t_j, \dots, \sqrt{t_i}, e^{-t_i}$ etc, $i, j, \dots = 1..l$). Since arguments are correlated, a little alteration of input data often results in a models with absolutely different suites of arguments. We propose two steps for application of factor analysis to GMDH. The first step simply consists in using GMDH in the orthogonal basis of factors. On the second, heuristic step we preliminarily obliterate so called unstable and inessential factors. Some versions of this method are realized by means of Java. It was successfully used for modelling and forecasting of extremely unstable molybdenum prices: monthly prices in 2004–05, and annual prices in 1975–98.

Keywords

inductive modelling, IWIM 2007, factor analysis, price forecasting, market forecasting, MAS forecasting

1 Introduction

In building forecast models, we often meet a typical situation, when obtained model has not any sense beyond some narrow neighbourhood of its construction domain, e.g. beyond neighbourhood of its statistical sample.

Let we build a model $y = f(t_1, t_2, \dots, t_l)$. We shall assume this model is linear with regard to some set (x_1, x_2, \dots, x_k) of basic functions, where each $x_i = \varphi_i(t_1, t_2, \dots, t_l)$. Usually we pick up these basic functions among $\binom{m+l}{m}$ monomials of all degrees $s \leq m$:

$$\prod_{j=1}^s t_{i_j}, \quad s = 0..m, \quad \forall j : i_j \in \{1..l\}.$$

The authors of [1] mention as basic for support functions harmonic and logistic functions

$$\sin(t_i), \quad \cos(t_i), \quad \frac{1}{1 + e^{-t_i}}.$$

In other works as basic functions are using roots $\sqrt[q]{t_i}$ or fractional degrees $t_i^{p/q}$ ($p, q \in \mathbb{N}$), logarithms $\log(t_i)$ etc.

Let n denote the dimension of statistics. Thus in ideal case we have obtained k vectors $(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_k)$ in a n -dimensional space. Now let \mathbf{X} denote the matrix

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

If an amount of linear arguments (i.e. basic functions) k exceeds the dimension of statistical sample n , we can build an exact model, with the exception of the algebraic variety

$$\det(\mathbf{X}\mathbf{X}^*) = 0 \quad \text{i.e.} \quad \text{rank}(\mathbf{X}) < n$$

of measure null. But it is well-known fact (see e.g. [2]) that this model is not satisfactory for extrapolation and forecasting.

How to avoid instability of obtained models and their inanity, insignificance beyond the neighbourhood of their construction domain?

2 Method

The way proposed below consists in the integration of academician Ivakhnenko's Group Method of Data Handling (GMDH) and factor analysis.

What is factor analysis? Let reduce the Gramian matrix [3] $\mathbf{X}^*\mathbf{X}$ by orthogonal transformation \mathbf{S} to diagonal form: $\mathbf{S}^*\mathbf{X}^*\mathbf{X}\mathbf{S} = \mathbf{D}$, so as to $d_{11} \geq d_{22} \geq \dots \geq d_{kk}$. 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 $L = L(\mathbf{x}_1 - \mu_1, \mathbf{x}_2 - \mu_2, \dots, \mathbf{x}_k - \mu_k)$, where $p = \dim L \leq \min(k, n)$. 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

$$\hat{y} = y_0 + y_1 z_1 + y_2 z_2 + \dots + y_p z_p, \quad (2)$$

note, that

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

Now we can apply GMDH in the orthogonal basis of factors, using all advantages of orthogonality. Further, on the heuristic step we preliminarily obliterate so called unstable and inessential factors.

Let's call factor z_i as *unstable* in statistics \mathbf{X} (with the threshold β), if

$$\frac{d_{ii}}{\text{trace}(\mathbf{D})} = \frac{d_{ii}}{d_{11} + d_{22} + \dots + d_{pp}} < \beta. \quad (4)$$

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 z_i as *inessential* for model \hat{y} (with the threshold γ), if correlation

$$\text{corr}(\mathbf{y}, \mathbf{z}_i) = \cos(\mathbf{y} - y_0, \mathbf{z}_i) = \frac{(\mathbf{y} - y_0, \mathbf{z}_i)}{|\mathbf{y} - y_0| |\mathbf{z}_i|} < \gamma. \quad (5)$$

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

$$\hat{y}^s = y_0 + y_1^s z_1 + y_2^s z_2 + \dots + y_k^s z_k, \quad (6)$$

where $\forall i > 0, y_i^s = y_i^s(\beta, \gamma) = 0$, if factor z_i is unstable or inessential, and $y_i^s = y_i$ for all other factors. In other words, stabilization is the obliteration of all unstable and inessential factors.

Some versions of this model are realized by means of Java, with using, when it is expedient, MS Excel.

3 Results

This method was successfully used for modelling and forecasting of molybdenum and ferromolybdenum (Mo, FeMo) prices: extremely unstable monthly prices in 2004–05, and annual prices in 1975–98.

Experiment 3 (last in time, but the most interesting). For monthly moving forecasting of molybdenum prices in 2005-07, in Internet were selected 9 activities t_1, t_2, \dots, t_9 : 6 steel prices [5] of Experiment 2 (see Tab. 2), t_7 – cuprum price [4], and t_8, t_9 – molybdenum export and import prices [4].

We put $x_i = t_i, i \leq 9, x_{10} = t_8(m - 1), x_{11} = t_9(m - 1), x_{12} = t_9(m - 2)$, where m is month. We put $y = t_9(m + 1)$.

Assuming $\beta = 0,001, \gamma = 0$, author was very surprised with some excellent results of forecasting (see Fig. 1) in the most complicated situations, see emphasis text in Tab. 1.

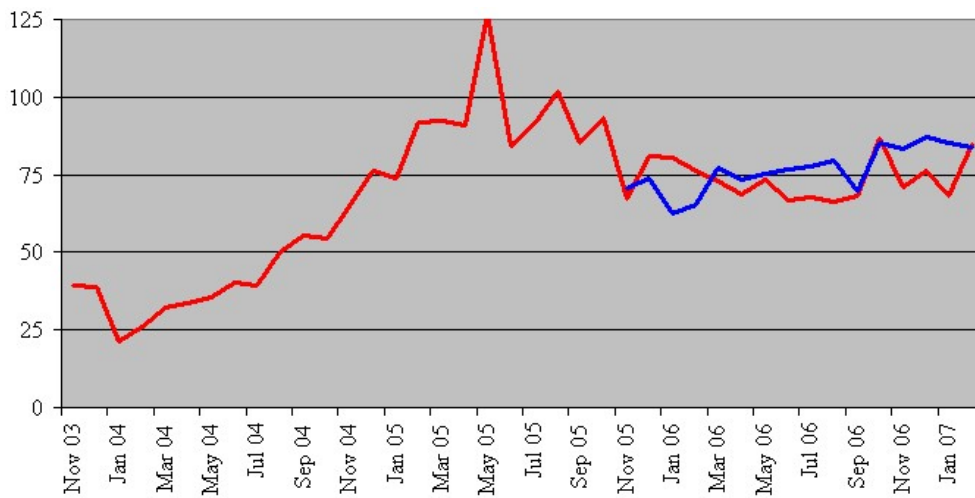


Fig. 1. Graphics of the actual values of molybdenum prices (red line), calculated as on Fig. 2, and forecasted values (blue line)

Experiment 2. For monthly simulation of molybdenum prices in the period 2004-2005 years from Internet were chosen 7 activities – world stainless steel prices of different grades [5] (see Tab. 2), and t_7 – molybdenum price [4]. Of course, $y = t_7(m + 1)$.

Factor analysis gives the next values of

$$\frac{d_{ii}}{\text{trace}(\mathbf{D})} : \quad 0.98095 \quad 0.01740 \quad 0.00093 \quad 0.00049 \quad 0.00017 \quad 0.00004 \quad 0.00002;$$

$$\text{corr}(\mathbf{y}, \mathbf{z}_i) : \quad 0.9208 \quad -0.0029 \quad -0.0542 \quad 0.1236 \quad -0.0574 \quad -0.0756 \quad 0.1556.$$

Assuming $\beta = 0,001$, we see that for stabilization of model it is enough to reserve three main factors. The rest, especially 6-th and 7-th, are destabilizing.

On the diagram Fig. 2 are represented graphics of the real values of molybdenum prices, which were calculated from data given in tables [4], columns *Imports*, row *Molybdenum*, and simulated values.

Experiment 1 (the first in time). For modelling of prices on concentrated molybdenum in 1975-98 in Internet were selected 16 annual indicators, among them: world molybdenum production, production and consumption of molybdenum in USA, consumption of molybdenum for production of steel, cast irons, superalloys, other alloys, mill products, chemical and ceramic uses et alia, plus refiner acquisition cost of imported crude oil (IRAC). An accuracy of forecast was very successful, but practical interest to annual forecasts is much less than to monthly or weekly ones.

Tab. 1. Results of forecasting of Molybdenum price (\$ US/kg)

Month	Price		Diffe- rence	Change	
	Fact	Forecast		Fact	Forecast
Oct 05	93.51				
Nov 05	67.45	70.36	2.91	-26.05	-23.15
Dec 05	81.06	73.88	-7.18	13.61	6.43
Jan 06	80.58	62.40	-18.18	-0.48	-18.66
Feb 06	76.19	65.50	-10.69	-4.39	-15.08
Mar 06	73.15	77.08	3.94	-3.04	0.90
Apr 06	68.49	73.37	4.88	-4.66	0.22
May 06	73.38	75.50	2.12	4.89	7.01
Jun 06	66.83	76.55	9.72	-6.55	3.17
Jul 06	67.55	77.46	9.91	0.72	10.63
Aug 06	66.29	79.73	13.45	-1.26	12.18
Sep 06	68.37	69.60	1.23	2.08	3.32
Oct 06	86.76	85.23	-1.53	18.39	16.86
Nov 06	71.01	83.53	12.52	-15.75	-3.23
Dec 06	76.35	87.05	10.70	5.33	16.03
Jan 07	68.18	85.02	16.83	-8.16	8.67
Feb 07	84.67	84.02	-0.65	16.48	15.84

Tab. 2. MEPS – World Stainless Steel Product Prices (\$ US/tonne).

Date	Hot Rolled Plate		Cold Rolled Coil		Drawn Bar	
	Grade 304	Grade 316	Grade 304	Grade 316	Grade 304	Grade 316
Jan 04	2117	2915	2137	2919	2376	3111
Feb 04	2367	3206	2372	3222	2556	3349
Mar 04	2468	3389	2484	3399	2639	3480
Apr 04	2540	3451	2509	3409	2697	3546
...
Sep 05	2768	4933	2487	4714	2968	5225
Oct 05	2714	4746	2434	4534	2895	4973
Nov 05	2665	4750	2389	4531	2803	4925
Dec 05	2506	4578	2217	4341	2607	4705

4 Conclusion, Discussion et alia

The results of simulation (modelling) and forecasting sufficiently represent a price dynamics on the world markets.

The executed calculation show that in 2004-05 price forecast only on 1 month ahead (fig. 2) would permit to economize up to 32,3% of funds in comparison with monthly purchasing of Mo. Forecasting on 6 monthes should preserve about 60% of funds. Total negative profits owing to lack of forecast in this period amount to 68,11%.

However, having analysed oil price chronology in 1970-2005 [6], it is necessary to mention that *it is impossible to increase reliability of forecasting of metals and fuel prices without inclusion of discrete factors in a model*, such as coming to power of political forces, armed conflicts, decision making on national level about liberalization or, on the contrary, about state support of national interests etc.

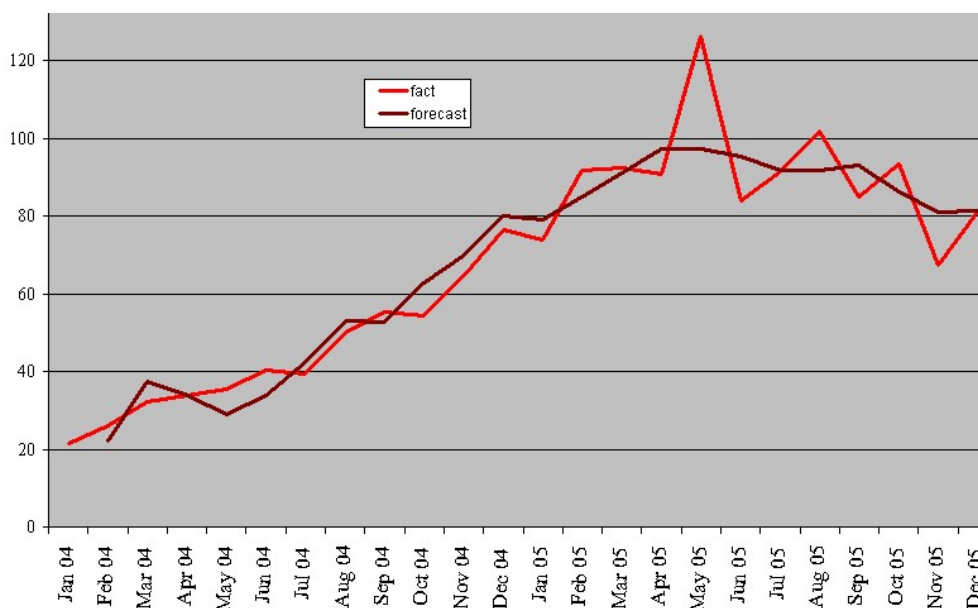


Fig. 2. Graphics of the actual values of molybdenum prices, which were calculated from data given in tables [4], columns *Imports*, row *Molybdenum*, and simulated values.

References

- [1] Ivakhnenko A.G., Ivakhnenko G.A.: [The Review of Problems Solvable by Algorithms of the Group Method of Data Handling](#) (pdf)
- [2] Koppa Yu.V., Stepashko V.S.: [A Comparison of the Forecasting Properties of Regression Types Models versus GMDH](#) (pdf) (*in Russian*)
- [3] [Gramian matrix.](#)
Jørgen Pedersen Gram.
- [4] [Metals Statistics, U.S. Metals Trade](#)
- [5] MEPS (International) Ltd. – Independent Steel Industry Analysts, Consultants, Steel Prices, Reports and Publications. [World Stainless Steel Product Prices](#)
- [6] Energy Information Administration. Official Energy Statistics from the U.S. Government. [World Nominal Oil Price Chronology: 1970-2005.](#)
- [7] Tymashova L.A., Dzyadyk Yu.V., Leshchenko V.A., Bondar L.A.: [An Intelligent System for Price Forecasting](#) (*in Ukrainian*) // *Problemy vprovadzhennya informacijnyh tehnologij v ekonomici. Tezy dopovidej VI mizhnarodnoyi naukovopraktychnoyi konferenciyi. Irpin', 2007*