9 PREDICTABILITY AND COMPLEXITY IN MACROECONOMICS. THE CASE OF GROSS FIXED CAPITAL FORMATION IN THE ROMANIAN ECONOMY[1]

Cornelia SCUTARU[*]
Corina SÂMAN[**]
Cristian STĂNICĂ[***]

Abstract

There is a relationship between predictability and complexity. The problem of evaluating the complexity of the macroeconomic phenomenon can be reduced to decomposition into its principal components (which may have, in their turn, a certain degree of complexity) and to identify its common sources of evolution that are predictable. In this paper, we evaluate the predictability of economic indicators and continue with its decomposition until the simplest sources allowed by available statistical data are obtained, then use this predictable sources to construct a forecasting model.

Keywords: predictability, complexity, principal components, consumption, investment, foreign trade, time series forecasting
JEL Classification: E20, E27, C22, C51

1. Introduction

The starting idea of this paper is that there is a relationship between **predictability** and **complexity** [9]. There are many ways of measuring the complexity on the basis of entropy (Kolmogorov complexity) or of interdependence between the components parts [6, 10, 7, 8]. A simple and intuitive measure of complexity of a process can be seen as the inverse of its temporal predictability. If some variables can be forecasted by their

[1] Prepared under the Programme: "Modelling and assessing the impact of direct investments on national and international labour market and macroeconomic evolutions of Romania", Contract MEC 91-052/ 10 September 2007; presented at the 5th European Conference on Complex Systems, Jerusalem, September 14-19 2008.

* *Senior researcher, Ph.D., Institute for Economic Forecasting, Romanian Academy, Bucharest;*
e-mail: corneliascutaru@yahoo.com .
** *Researcher, Ph.D. Student, Institute for Economic Forecasting, Romanian Academy, Bucharest; e-mail: csaman@ines.ro .*
*** *Researcher, Ph.D., Institute for Economic Forecasting, Romanian Academy, Bucharest;*
e-mail: stanica_cris@yahoo.com .

previous evolutions they have a low complexity and if they are practically impossible to predict they have a high complexity.

Every complex phenomenon can be decomposed into principal components, statistically uncorrelated, or more generally to independent components, which are coming from different sources that may have some common underlying factors, which might reveal some driving mechanism that otherwise remain hidden.

We consider an economy characterized by aggregated macroeconomic indicators such as GDP and its components, which are determined by other variables carefully chosen to entail the evolution of the process. But are these variables predictable? Can we make a forecasting model to evaluate the state of the economy? Or do they have underlying factors, not measurable, which are sources of process evolution and at the same time predictable? This is the problem we consider in this paper.

Thus, a country's GDP evaluated by the method of expenditure includes: Gross capital formation (including the formation of gross fixed capital - investments), consumption (including household consumption and government consumption), export, import, each of them depending on different aspects of the economic phenomena, which can be themselves sources of complexity.

We construct an empirical model for the case of the Gross Fixed Capital Formation, considering the group of macroeconomic indicators that influence it, and compute the principal components, which are decomposed into some factors which are common sources of evolution of the entire group and also predictable. Evaluating the principal components is not new for the Romanian economy: in [3] and [4] the GDP was evaluated by expenditure and production, in order to estimate a predictive model, using the principal component analysis, respectively; and in [5] the impact of oil price on GDP is studied using the same techniques.

Each processing step evaluates the vector of predictable latent components s_i as a linear combination of uncorrelated variables x_j (using the weights w_{ij}).

For each decomposition $\mathbf{s} = \mathbf{w}^T\mathbf{x}$, w is found to maximize **s** predictability and minimize its complexity. A measure F of s predictability is defined [9] as a ratio of total variance of $\mathbf{s}=(s_i)$ by the measure of it's "smoothness" calculated as the moving average.

In economy, the data series for a specific phenomenon can have such predictability properties only if they are not influenced by economic shocks. But no country is immune to unexpected shocks, which cause fluctuations, and these shocks can be determined by random phenomena (natural calamities, oil price crisis or financial crisis) or by a mix of inadequate macroeconomic policies (such as in pre-election periods). The impact of economic shocks can be evaluated if a maximal **predictability** of indicators can be achieved, and so a minimization of multivariate data **complexity**. This can be obtained by studying the response of endogenous variables to sudden variations of the exogenous ones, meaning a sudden increase in external complexity.

What must be stated from the very beginning is that we are estimating the complexity and predictability of an economic phenomenon strictly limited by the series of data taken into consideration and not considering the entire economy. These data series were chosen after several tests, in accordance with the influence they can have on the components of GDP (gross capital formation, consumption, foreign trade). The principal components method allows for the separation of different influences on GDP: influences related to prices, foreign exchange and monetary policies, the evolution of the real economy (reflected by the index volume of certain industries).

This paper is organized as follows. Section 2 describes the data and its preprocessing. Section 3 provides an introduction to complexity pursuit. Section 4 presents experimental results on using the complexity pursuit algorithm on GDP components on domestic consumption. Finally, some conclusions are drawn in Section 5.

2. Principal Component Analysis

We employ a vector space model for representing the GDP components[1]. Each group is formalized as data matrix X containing the component vectors in columns and is of size TxM, where M is the number of components. We will write X referring to the whole set of data vectors and x when referring to one of them; thus $X = (x_i)$, i=1,…,M.

As a preprocessing step, we compute the principal components of the data matrix X, that is the vectors z_i and matrix Z:

$$\mathbf{Z} = B^{-1}X$$

The new data matrix Z and its columns z_i, i=1,…,M are the inputs for the algorithm that will be described in Section 3. The time-structure of the GDP components, or the minimum complexity projections can be found by projecting Z onto the directions $W = (w_1,\dots,w_M)$ given by the complexity pursuit algorithm described in the next section.

3. The Complexity Pursuit Algorithm

Complexity pursuit [1, 2] is a recently developed, computationally simple algorithm for separating interesting components from time series. It is an extension of projection pursuit to time series data and also closely related to Independent Component Analysis (ICA). Projection pursuit seeks for directions in which the data have an interesting, structured distribution neglecting any time dependency information that may exist in the data. ICA, on the other hand, finds statistically independent directions. Complexity pursuit combines these criteria in a principled way by employing the information theoretical concept of Kolmogorov complexity and developing a simple approximation of it.

We use a method of separating predictable sources in time series recently presented by Stone [9], which minimizes a measure of Kolmogorov complexity. In his approach, it is assumed that any mixture of source signals is more complex than the simplest of them. This conjecture is the basis for separating mixtures into their sources by seeking the least complex signal obtained from the mixture [9].

The complexity is measured in terms of temporal predictability, so that lower complexity corresponds to higher predictability.

---

[1] What we present in this paper is part of a larger study which explores training models of GDP components (population consumption, goverment consumption, gross capital formation, export, import) from assessment of their predictability / complexity, using the same methodology illustrated here for gross fixed capital formation.

The data model assumes that the observations x(t) are linear mixtures of some latent components:

$$X = A\mathbf{x}S \tag{1}$$

Where: $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_M)$ is the vector of observed random variables, $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_M)$ is the vector of predictable latent components, and A is an unknown constant mixing matrix.

A separate moving average model is assumed to model each component $\mathbf{s}_i = \mathbf{w}_i^T \mathbf{X}$; as an exponentially weighted sum of past values:

$$\tilde{\mathbf{s}}_i(t) = \sum_{k=0}^{t-1} \lambda^{n-k-1}(1-\lambda)\tilde{\mathbf{s}}_i(k) + \lambda^t \tilde{\mathbf{s}}_i(0) \qquad (2),$$

for some λ, $\quad 0 \le \lambda \le 1.$

That is, the predicted value $\tilde{s}_i(t)$ of $s_i(t)$ is a sum of signal values measured up to time (t-1), such that recent values have a larger weighting than those in the distant past:

$$\tilde{\mathbf{s}}_i(t) = \lambda\, \tilde{\mathbf{s}}_i(t-1) + (1-\lambda)\, \mathbf{s}_i(t-1)\,. \tag{3}$$

For each decomposition s = w^T x, w is found to maximize s predictability and minimize its complexity. A measure F of **s** predictability is defined [9] as a ratio of total variance of **s**=(s(t)), t=1…T by the measure of its "smoothness" calculated as moving average. F can be maximized if two conditions are satisfied: s has a high variance and has no high variation with time, meaning that the values of source signals show little changes from the ones predicted by past values.

The definition of predictability is in [9]:

$$F(\mathbf{w}_i, \mathbf{x}) = \ln \frac{\sum_t (\overline{\mathbf{s}_i} - \mathbf{s}_i(t))^2}{\sum_t (\tilde{\mathbf{s}}_i - \mathbf{s}_i(t))^2} = \ln \frac{V_i}{U_i} \tag{4}$$

In our application, the latent time-components, s_i, will model the predictable sources for our GDP components groups. To find the maximum of (4), the data are first whitened by PCA as described in the previous section. We denote by z(t) this preprocessed data, and **b** now corresponds to an estimate of a row of the inverse of the mixing matrix for whitened data.

$$\mathbf{X} = \mathbf{B} \times \mathbf{Z} \tag{5}$$

Considering the scalar data z_i formed by means of a weight vector w_i from a set of M sources S = $(s_1, \ldots, s_M)$, equation (4) can be rewritten as in [9]:

$$F(\mathbf{w}_i, \mathbf{z}) = \ln \frac{\mathbf{w}_i \overline{\mathbf{C}} \mathbf{w}_i^T}{\mathbf{w}_i \tilde{\mathbf{C}} \mathbf{w}_i^T} \tag{6}$$

where: $\overline{\mathbf{C}}$ and $\tilde{\mathbf{C}}$ are MxM matrix of covariance:

$$\overline{C}_{ij} = \sum_t (z_{ti} - \overline{z}_{ti})(z_{tj} - \overline{z}_{tj})$$
$$\tilde{C}_{ij} = \sum_t (z_{ti} - \tilde{z}_{ti})(z_{tj} - \tilde{z}_{tj}) \quad (7)$$

Then the gradient update of w that minimizes (6) is the following [9]:

$$\mathbf{w}_i = \mathbf{w}_i + \eta \nabla_{\mathbf{w}_i} F$$
$$\nabla_{\mathbf{w}_i} F = \frac{2\mathbf{w}_i}{V_i}\overline{\mathbf{C}} - \frac{2\mathbf{w}_i}{U_i}\tilde{\mathbf{C}} \quad (8)$$

where: η is a small constant, the step uphill along the direction of the gradient.

To estimate several projections w_i one can either use a deflation scheme or estimate all projections simultaneously. In the deflationary approach, after estimating p projections, one runs the algorithm for w_{p+1} on the modified mixtures after subtracting from each mixture z_i the projections of the previously estimated p source signals using Gram-Schmidt orthogonalization.

Thus, a matrix W = (w_1, …, w_M) is found, which verifies:

$$\mathbf{S} = \mathbf{Z} \times \mathbf{W} \quad (9).$$

The model can be used to forecast one-step-ahead:

1. equation (3) applied to time t+1 gives the sources values for time T+1:
2. $\tilde{\mathbf{s}}_i(T+1) = \lambda \tilde{\mathbf{s}}_i(T) + (1-\lambda)\mathbf{s}_i(T)$
3. calculate **Z** = $\mathbf{W}^{-1}$x**S**, where **S**=($\mathbf{s}_1$, …, $\mathbf{s}_M$) and $\mathbf{s}_i$=(s_i (t)), t=1…T+1.

For forecasting n+1 times ahead, on apply the complex pursuit algorithm to new matrix Z containing the component vectors of size (T+1)xM, (T+2)xM, …, (T+n)xM obtained by adding the one row with forecasted data calculated in previous step.

4. Experimental Results

Data used in our statistics are compiled by the National Institute of Statistics, from the first quarter of 2000 until the second quarter of 2007; we were working with the index chain of variables expressed in real terms. GDP components group were created by the expenditure decomposition method, for each component we built a group of variables that expressed, in our opinion, the most powerful influences on the component.

The G1INV Group (investments) includes variables selected for the assessment of the principal components corresponding to gross fixed capital formation (GFCF): ratio between income and general consolidated budget expenditure (RSB), the volume of construction activity (ICONSTR); the volume of machine production activity (IUTIL), the volume of industrial production (IPI), the degree of coverage of imports by exports (GXM), exchange rate against the euro (ERE); broad money in real terms (M2R), real gross wage in the economy (SBREC) .

The components are INVPC1, INVPC2, INVPC3 and INVPC4, which correspond to the eigenvalues of 3,956, 1,412, 0,946 and 0.0.642; the proportion of variance explained by these four main components is 0.87%. The analysis of the principal components indicates that the first component depends on the balance of the consolidated general budget, the volume of activity in construction, manufacturing machinery, industrial and the financial variables: the foreign trade deficit, broad money, and real gross wage in the economy. The second component mostly includes financial influences: the exchange rate and broad money, wage and other variables having much lower weights. The third includes the budgetary influences, the volume of activity in construction and industry, as well as the real gross wage in the economy. A fourth component depends on the broad money, exchange rate and the volume of equipment production activity.

The equations of the four main significant components are:

INVPC1 = -0.36*IRSB + 0.43*IICONSTR + 0.39*IUTIL + 0.41*IIPI - 0.44*IGXM + 0.22*IM2R + 0.34*ISBREC

INVPC2 = 0.15*IRSB - 0.07*IIPI + 0.25*IUTIL - 0.10*IGXM - 0.71*IERE - 0.55*IM2R + 0.37*ISBREC

INVPC3 = 0.49*IRSB + 0.35*IICONSTR + 0.16*IUTIL + 0.51*IIPI + 0.08*GXM - 0.12*IERE - 0.07*IM2R - 0.57*ISBREC

INVPC4 = -0.11*IRSB - 0.14*IICONSTR + 0.32*IUTIL + 0.08*IIPI - 0.07*IGXM + 0.60*IERE - 0.69*IM2R - 0.07*ISBREC

Our experiments showed that choosing a moving average model:

$\tilde{\mathbf{s}}_i(t) = \lambda \tilde{\mathbf{s}}_i(t-1) + (1-\lambda)\mathbf{s}_i(t-1)$ was successful and that $\lambda = 0.99$ was the most suitable.

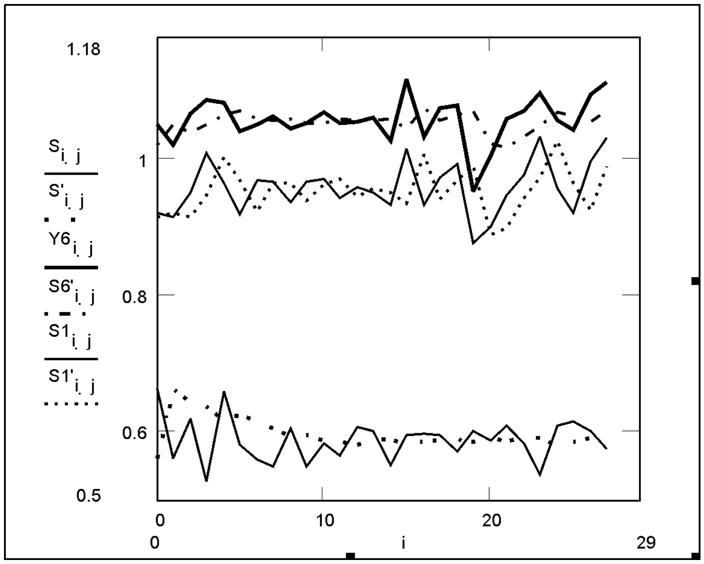
In the next graph, 3 source signals $\mathbf{s}_i(t)$ (solid line) are plotted together with model output $\tilde{\mathbf{s}}_i(t)$ (dashed) for 3 different lambdas (0.5, 0.6, 0.1) (Figure 1).

We calculated two types of measures for deviations from this model:

$$M1(\mathbf{s_i}) = \frac{\tilde{\mathbf{s}}_\mathbf{i}(T) - \mathbf{s_i}(T)}{\mathbf{s_i}(T)} * 100$$

$$M2(\mathbf{s_i}) = \frac{\tilde{\mathbf{s}}_\mathbf{i}(T) - \mathbf{s_i}(T)}{stdev(\mathbf{s_i})} \tag{10}$$

Figure 1



where M1 represents the percentage error of the model and M2 represents the number of standard deviations away from real data.

Next, we give the results for predictions on 2007:Q2 using observed data from period 2000:Q1 – 2007:Q1.

The identified predictable sources $\tilde{\mathbf{s}}_i(t)$ have a deviation of at most 0.93 standard deviation from $\mathbf{s}_i(t)$ and the forecasts have good accuracy,

For the first three principal components, which assess for 79% of the explained variance, the forecasts have high accuracy (percentage error is at most 5.36%).

$M1(s_i)$	$M2(s_i)$	$M1(z_i)$	$M2(z_i)$
2.85	0.89	-4.5	0.2
5.64	-1.19	5.35	0.37
1.4	0.54	-5.36	0.37
1.83	0.6	-28.56	-0.35
-0.1	0.02	-2.3	-0.11
-13.54	0.43	3.83	0.93
19.58	0.23	-6.82	-0.34
-41.72	-1.12	-3.12	0.14

The criterion adopted for analyzing the quality of prediction is presented in the table below:

Table 1

M1 Criterion for the Evaluation of a Forecasting Model

M1 - Absolute value	Forecasts Classification
<10	high accuracy
10-20	good accuracy
20-50	reasonable accuracy
>50	unreliable

Thus, we obtain high accuracy for all principal components, except for the fourth, which gives reasonable accuracy (28.56%).

The values of percentage error M1 are given in the tables below.

Table 2

Values of M1 for Principal Components from G1INV group over the period 2006:Q2 to 2007:Q2

	INVPC1	INVPC2	INVPC3	INVPC4	INVPC5	INVPC6	INVPC7	INVPC8
2006:Q2	-3.04	-0.79	-4.88	-31.24	-2.11	1.95	5.32	-0.77
2006:Q3	25.02	-1.98	9.86	-24.80	-12.12	0.22	17.81	18.97
2006:Q4	4.37	-1.70	49.14	-5.68	-8.42	1.22	96.49	30.97
2007:Q1	45.63	16.25	24.47	81.91	-40.58	-3.22	77.68	129.37
2007:Q2	-4.50	5.35	-5.36	-28.56	-2.30	3.83	-6.82	-3.12

From the analysis of the error values and also based on the criteria presented in Table 1, it may be said that the models successfully produced reasonable accurate forecasts for the first 6 principal components for this period, except for 2007:Q1.

Table 3

Values of M1 for variables from G1INV group over the period 2006:Q2 to 2007:Q2

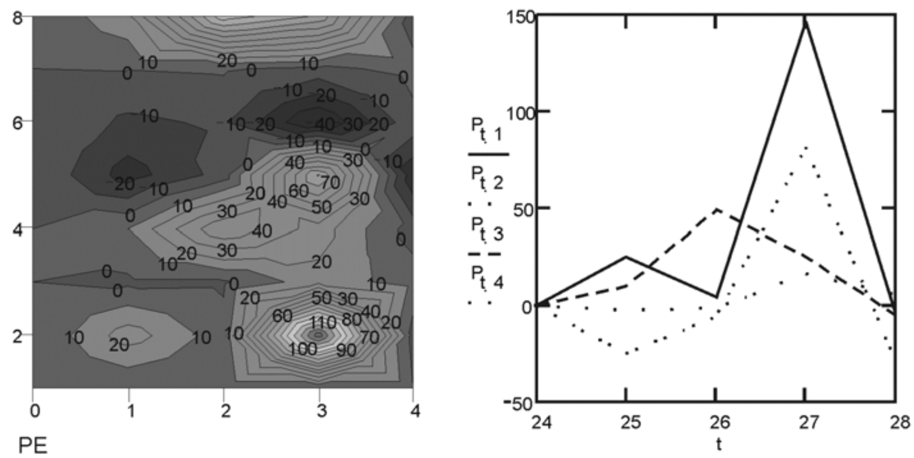
	IRSB	IICONSTR	IIUTIL	IIPI	IGXM	IERE	IM2R	ISBREC
2006:Q2	-0.48	-1.72	-3.37	-2.05	2.64	4.25	-5.84	4.13
2006:Q3	-6.60	25.73	8.96	6.07	-9.11	1.99	-4.45	5.64
2006:Q4	42.33	36.96	-2.87	2.58	-2.62	4.70	-9.74	-4.70
2007:Q1	-29.23	146.77	3.45	8.55	-20.05	5.56	-0.50	7.78
2007:Q2	-4.93	-4.46	-2.99	0.61	10.80	5.78	-2.12	1.02

It may be said that the models successfully produced highly accurate forecasts for IIUTIL, IIPI, IERE, IM2R, ISBREC, since the M1 has values lower than 10%; good accurate forecasts for IGXM and reasonable forecasts for IRSB. The exception is IICONSTR, which has reasonable accuracy, except for 2007:Q1.

We applied the model for forecasting on a horizon of four periods, using observed data on the period 2000:Q1 – 2006:Q1 as input.

Let be $PE_{t,i} = M1(z_i(t))$, where $z_i(t)$ is the data vector corresponding to i-th principal component and M1 is the measure define before. The next graphs (Figure 2) present the M1 percentage error for the first four principal components on this forecasting horizon.

Figure 2



For the first three forecasting periods the percentage errors are good, but after 3 steps big errors appear.

5. Conclusions

In this paper, we have shown experimental results on how maximum predictable projections (or, equivalently, minimum complexity projections) of GDP components can be used to predict future values. As an example of such applications we used variables selected for the assessment of the main components corresponding to gross capital formation. The method we used for finding latent variables as predictable sources, complexity pursuit [2, 9], is a generalization of the projection pursuit to times series and consists in estimating projections of data whose complexity measure is minimized. In our experiment, the complexity pursuit algorithm was used to find a forecast model for data series. Our result suggests that this method could serve in predicting future times series values for groups of variables which have common sources of variations.

In particular, one could use this algorithm to study indicators specific to each component of GDP, then to evaluate their predictability in order to create a model to evaluate the answer of endogenous variables to shocks.

References

[1] Hyvärinen, A. and Oja, E. (2000), "Independent Component Analysis: Algorithms and Applications", *Neural Networks*, 13(4-5):411-430

[2] Hyvärinen, A., (2001), “Complexity pursuit: separating interesting components from time-series”, *Neural Computation*, 13(4):883–898.

[3] Klein, Lawrence R.and Andrei Roudoi, Vladimir Eskin, Lucian-Liviu Albu, Cristian Stanica, (2004a), “Quaterly GDP Data Correction Using Principal Components Analysis. The Case of the Romanian Economy – GDP Expenditures Side”, *Romanian Journal of Economic Forecasting, Supplement.*

[4] Klein, Lawrence R.and Andrei Roudoi, Vladimir Eskin, Lucian-Liviu Albu, Cristian Stanica, Mariana Nicolae, Mihaela Nona Chilian, (2004b), “Principal Components Model of the Romanian Economy GDP – Production Side”, *Romanian Journal of Economic Forecasting, Supplemet*

[5] Klein, Lawrence R. and Andrei Roudoi, Vladimir Eskin, Mariana Nicolae, (2004c), “Principal Components Model of the Romanian Economy. Study of the Oil Price Impact upon GDP” *Romanian Journal of Economic Forecasting, Supplement*

[6] Lopes, J. Carlos, Dias, J., Fereira do Amaral, J., (2007), “Economic complexity as Input-Output Interrelatedness: A Comparison of Different Measurement Methodes”, paper submitted at the: *16th International Input-Output Conference*, Istambul, Turkey, 2-6 July 2007.

[7] Shnerb, N., Bettelheim, Y., Louzoum, Y., Agam, O., Solomon, S., (2001), “Adaptation of Autocatalytic Fluctuatios to Diffusive Noise”, *Phys. Rev. E* 63(2)*,* **http://xxx.lanl.gov.abs/cond-mat/007097**

[8] Solomon, S. (1998), “Stochastic Lotka Volterra Systems of Competing Auto-catalytic Agents lead Generically to Truncated Pareto Power Wealth Distribution of Market Returns, Clustered Volatility, Booms and Crashes”, *Decision Technologies for Computational Finance,* edited by A.-P. Refenes, A. N. Burgess and J. E. Moody (Kluver Academic Publishers. **http://xxx.lanl.gov/abs/cond-mat/9803367**

[9] Stone, James V., (2004), “*Independent Component Analysis. A tutorial Introduction*“, The MIT Press Cambridge, Massachusetts, London, England.

[10] Yaari, G., (2006), “The AB Model, the GLV, and a Data set from Poland”, Presentation of the PhD thesis, under the supervision of Prof. Sorin Solomon, Racah Institute of Physics in the Hebrew University of Jerusalem, Israel; Multi-Agent Systems Division. ISI Torino, Italy.