INDEPENDENDT VARIABLE SELECTION: AN APPLICATION OF INDEPENDENT COMPONENT ANALYSIS TO FORECASTING A STOCK INDEX

By

Andrzej Cichocki
Laboratory for Advanced Signal Processing
Brain Science Institute, RIKEN
2-1 Hirosawa, Wako-shi, Saitama, 351-0198, Japan
cia@brain.riken.go.jp

and

Zbigniew Leonowicz
Laboratory for Advanced Signal Processing
Brain Science Institute, RIKEN
2-1 Hirosawa, Wako-shi, Saitama, 351-0198, Japan
leon@bsp.brain.riken.go.jp

and

Stanley R. Stansell
Robert Dillard Teer Distinguished Professor of Business
School of Business
East Carolina University
Greenville, N. C. 27858
252-328-6636
stansells@mail.ecu.edu

and

James Buck
Associate Professor of Finance
School of Business
East Carolina University
Greenville, N. C. 27858
252-328-6625
buckj@mail.ecu.edu

KEYWORDS: INDEPENDENT COMPONENT ANALYSIS, NEURAL NETWORKS, FORECASTING

ABSTRACT

Forecasts of financial time series requires the use of a possibly large set of input (explanatory) variables drawn form a very large set of potential inputs. Selection of a meaningful and useful subset of input variables is a formidable task. How to find a reasonable transformation for a large set of multivariate data is a very common problem in many areas of science. We propose to use a technique called Independent Component Analysis (ICA) to extract the independent components (ICs) from monthly time series on a wide range of economic variables. This procedure will reduce the number of explanatory variables by reducing the set of financial and economic information to a much smaller subset of ICs which hopefully will capture most of the useful information. Removal of the random elements in each of the sets of economic data should make it much easier to identify relationships between the ICs and the stock indexes. Properly estimated ICs are independent of each other. We will then use the ICs from the explanatory variable data sets to perform and test forecasts of the S & P 500 stock index using neural network procedures. Numerous studies such as have shown neural networks to be very useful in nonlinear forecasting.

IC analysis has been employed in relatively few applications to finance. Kiviluoto and Oja (1998) use IC analysis in an application to parallel cash flow time series. Black and Weigand (1997) use IC analysis to extract estimates of the structure from a set of common stock returns. We feel that this research will contribute to the identification and understanding of the major economic factors affecting stock prices.

INTRODUCTION

Forecasts of financial time series requires the use of a possibly large set of input (explanatory) variables. Selection of a meaningful and useful subset of input variables is a difficult task. Procedures such as factor analysis and principal components analysis have been widely used in this area. Both have significant limitations. In our research we use a technique called Independent Component Analysis (ICA) to extract the independent components (ICs) from a set of monthly time series on a wide range of economic variables. ICA is a process which statistically reduces a possibly very complex data set into sub components which are statistically independent. This procedure reduces the number of explanatory variables by condensing the set of financial and economic information to a much smaller set of ICs. These ICs are expected to capture most of the useful information regarding the underlying economic events that form the

basis for the indexes. As Common (1994) notes, ICA is a way of finding a linear transformation of the data that minimizes the statistical dependence between the components extracted from the data. Removal of the random elements in the set of economic data makes it much easier to identify real relationships between the ICs and the dependent variable. Since properly estimated ICs are statistically independent from each other, we use them to create a new set of explanatory variables that can be used to produce forecasts of the S & P 500 stock index.

This paper produces and evaluates forecasting models of a stock index using Independent Component Analysis (ICA) as a preprocessing algorithm. See Back(1997), Back, *et al* (1994), Back and Weigand (1997), Cardoso (1998), Kiviluoto and Oja (1998) and Cichocki (2002) for applications in finance. ICA has been used also by Moody and Wu (1997) to separate observational noise from the theoretical prices found within foreign exchange rate time series.

METHODOLOGY

ICA is a process which extracts from a signal vector a new set of statistically independent components. These components represent estimates of the original data source. This process assumes a time series matrix that has an imbedded mixing process that can be demixed, and, when we assume there are as many observed signals as sources, separation occurs and we can find the best possible matrix defining the co-movement of the predictor variables. The fact that two random variables are uncorrelated does not also imply that they are independent. This fact is lost in using other methods such as Principal Components Analysis (PCA). The ICA approach seeks to find such independent directions through contrast functions. Such functions may be extracted using various methods, including neural networks. Assumptions required for the ICA process include the following:

The signals are stationary – a part of most models

The sources are statistically independent – a likely event in economic time series

No more than one course has a Gaussian distribution – not a problem in financial data

These assumptions form the basis for finding the demixing matrix which produces the optimally weighted independent components.

Historical attempts to use Principal Component Analysis (PCA) to extract factors affecting stock index time series have had mixed results. PCA is assumed to produce orthogonal factors that improve prediction. However, it is often observed that PCA does not use both the covariance matrices of the data and changes in the data simultaneously – and these models avoid higher linear and nonlinear correlations. ICA may be found to be a more appropriate technique in the analysis and forecasting since it uses both the covariance matrix of the data and changes in the data simultaneously, with the possibility of adding higher order linear and non-linear correlations. ICA produces uncorrelated factors that approximate independence but may be non-orthogonal. Thus the contrasting goals of PCA and ICA are uncorrelated components vs. statistically independent components. In this paper, we apply the ICA model using the TICA algorithm described by Cichocki and Amari (2002). We use software written by Cichocki called ICALAB which is available free on the net.

Applications of ICA to financial time series data can be separated into two groups: (1) block-based approaches and (2) on-line adaptive models. A block-based method takes all data in at one time and produces output ICs, whereas on-line adaptive models process data points in continuous space. Given the tuning requirements and within-process decisions that must be made, it is important to pre-define time series relationships in order to pre-determine within process parameters. As a practical matter, the goal is to generate stable systems that can make use of consistent decision parameters within certain economic relationships. Another goal of this work is to define such a parameter set for the relationship between economic data and stock indices.

NEURAL NETWORKS

We use Neural Network (NN) models to generate the best forecasts of S&P index values in future months. NN applications have become widely used within the financial community due to their ability to quickly develop successful prediction and classification models that can be used within dynamic processes. A recent study by Hill, *et a*l (1994) discusses the use of NN models in forecasting. See also Gorr (1994). Traditional backward-propagation neural networks are used to minimize estimation error by propagating the network estimation error back through a network using the first derivative of each nodes transfer function, where the transfer function is a monotonic differentiable function such as the sigmoid or hyperbolic tangent function. Training becomes an iterative gradient decent process, with exemplars presented to the network and fed forward to produce the desired result. As errors are recognized between the initial result and the desired outcome, this difference is back-propagated to adjust the weights in the network so as to reduce the error. Thousands of iterations are often required to minimize total error in favor of the desired outcome of the overall network (e.g. highest possible correlation between inputs and output).

We employ NeuralWorks *Predict* software (an add-in to Microsoft Excel) to estimate the NN models. This tool is best suited for rapid development of prediction algorithms within a traditional neural network (NN) system. This system combines genetic algorithms, statistics and fuzzy logic to automatically find optimal solutions for economic time series analysis. Such development environments are user friendly, with many parameters of the models found automatically by the system itself. Analyses of data sets are performed to identify appropriate transforms and data partitions for both training and test sets. Relevant input variables are identified and tailored to match the goal of the user. After initial setup of the system, the *Predict* software is designed to automatically perform all of the actions necessary to build a successful prediction model. We anticipate the use of this step in the analysis of ICs to result in improved predictive success relative to the use of raw data alone.

DATA

This paper examines a preprocessing and predictive process for stock indexes. Using initial data from the period 1982.11 through 2002.12, we apply ICA to monthly time series on: interest rates, money supply, production, employment, consumer confidence and a wide range of other economic variables. These variables become simultaneous input values to the ICA model. A complete list of the economic input variables is provided in Table 1. We extract and condense ICA components that are used as input values to neural networks for the prediction of the S & P 500 stock index. We extract the ICs, determine which have forecasting power, estimate deflated or reconstructed series, and apply the neural network models to predict the S & P 500 index several months in advance.

PROCEDURES

Three different approaches can be used in forecasting a stock index. First, we could simply use lagged values of the original economic input series and ignore the entire topic of ICA. Second, we could use the ICs estimated from the economic series listed in Table 1 as inputs into the NN model. Third, we could use the ICs estimated from the economic series listed in Table 1 to reconstruct (deflate) the economic series and then use these reconstructed series as inputs in the forecasting process. We test the second and third approaches in this paper.

The first step was to select a set of economic variables which could reasonably be expected to have some impact on a broad stock index. These variables are listed in Table 1. We then used the TICA algorithm (described in Cichocki and Amari (2002) and the ICALAB software to estimate the ICs for these 28 input economic time series.

We next employed the *Predict* neural network program developed by NeuralWare to model and test which of the ICs possessed any predictive ability. The *Predict* neural network program has a set of internal algorithms which tries different transformations of all the input and output variables and selects those which have some predictive power. The transformations are shown in Table 2.

Separate tests were run using lags of 1 through 6 months on the S & P 500 index. Extensive test forecasts were run over both the entire data sample and over the most recent 12 months. The NN models indicated that ICs number 1 – 17 and 20 had some predictive power at various lags.

RESULTS AND CONCLUSIONS

Successive forecasts were performed over both the entire data sample and the most recent 12 months for time horizons ranging from 1 to 6 months. The process was repeated using both the ICs directly as input variables and using the reconstructed economic input variables. As noted, the detailed results are reported in Table 3. For illustrative purposes we have included Chart 1 (full sample, 1 month forecast horizon) and Chart 2 (most recent 12 months, 1 month forecast horizon). The forecasts fit the actual data quite well.

The summary statistics on forecast errors shown in Table 3 indicate that the forecasts performed with deflated or reconstructed variables based on ICs yield the best forecasts. Using the ICs as inputs does not work well. In Table 3 column one should be compared to column three and column two compared to column four. (One and three deal with the full data sample; two and four with the most recent twelve months of the data.) The range and standard deviation of the forecast errors using reconstructed variables of the most recent 12 month forecasts are consistently lower. The mean of the errors is usually lower. It appears that the best performance can be found when variable selection is based on the ICs and the series are reconstructed (deflated) using only these ICs.

Another way of comparing the results is to examine the compounded returns from both a naïve investment approach and from using the best forecasts. First we convert the full data sample into monthly returns. A naïve approach would be to reinvest each month and earn the actual return (positive or negative) for that month, resulting in an original investment of one dollar accumulating to $6.08. If we instead rely on our one month ahead forecasts, we would invest in the S & P 500 if our forecast for the coming month is positive and earn the actual return (which could be positive or negative) or, if the forecast is negative, we

could invest in the money market and earn an annualized return of 2%. This procedure would result in an accumulated value of $10.82 from an original investment of one dollar.

We view this paper as being in the preliminary stages. Further work is needed to substantiate the results obtained herein. The evidence suggests that the use of independent component analysis as an input variable selection procedure has great promise.

REFERENCES

Back, Andrew D. and Weigand, Andreas S., 1997. “A First Application of Independent Component Analysis to Extracting Structure from Stock Returns,” International Journal of Neural Systems, 8, No. 4, 473-484.

Back, A., G. Ooosterom, K. Sere and M. van Wezel. 1994. “A comparative study of neural networks in bankruptcy prediction” in *Multiple Paradigms for Artificial Intelligence*, Helsinki, Finland.

Cardoso, J. F..1998. “Blind signal Separation: Statistical Principles”, Proceedings of the IEEE 86, 1998.

Cichocki, A. and Shun-ichi Amari, 2002. *Adaptive Blind Signal and Image Processing*. John Wiley and Sons, LTD. 2002.

Common, P.,1994, ”Independent Component Analysis, A New Concept?”, *Signal Processing, 36 ,287-314.*

Gorr, Wilpen L., 1994, “ Neural Networks in forecasting : Special section”, *International Journal of Forecasting*, 10, 1-4.

Hill, T., Marquez, L., O’Connor, M. and Remus, W., 1994, “Artificial Neural Network Models for Forecasting and Decision Making”, *International Journal of Forecasting*, 10, 5-15.

Kiviluoto, Kimmo and Oja, Erkki, 1998, “Independent Component Analysis for Parallel Financial Time Series”, Helsinki University of Technology, *Laboratory of Computer and Information Science*, 895-898.

Moody, J. and L. Wu. 1996. “What is the True Print? – State Space Models for High Frequency FX Data”. *Decision Technologies for financial engineering* – proceedings of the NNCM ’96 vol 7 of Progress in Neural Processing, Pages 346,358. World Scientific, 1997.

Smari, S., A. Cichocki and H. H. Yang. 1996. “A new learning algorithm for blind source separation,” in *Advances in Neural Information Processing* 8, Cambridge, MA. .

Tong, L., R. Liu, V. Soon and Y. Huang 1991. “Indeterminancy and Identifiability of Blind Identification”, in IEEE Trans. Circuits, Syst 38(5), 499-509.

CHART 1 S & P 500 1 MONTH AHEAD FORECASTS OVER FULL SAMPLE USING DEFLATED SIGNALS AS INPUTS

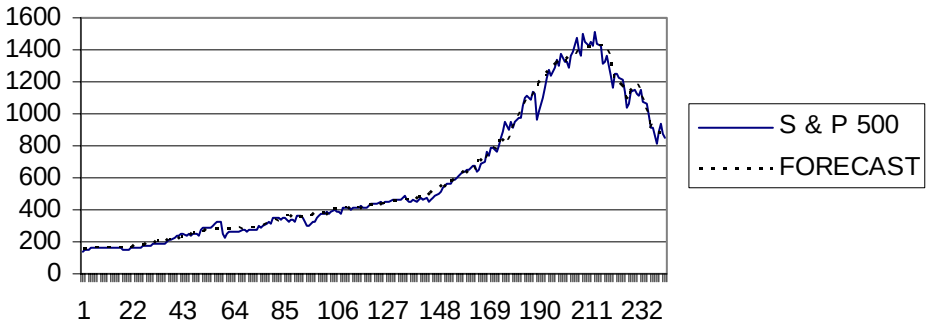

CHART 2: S & P 500 1 MONTH AHEAD FORECASTS OVER MOST RECENT 12 MONTHS USING DEFLATED SIGNALS AS INPUTS

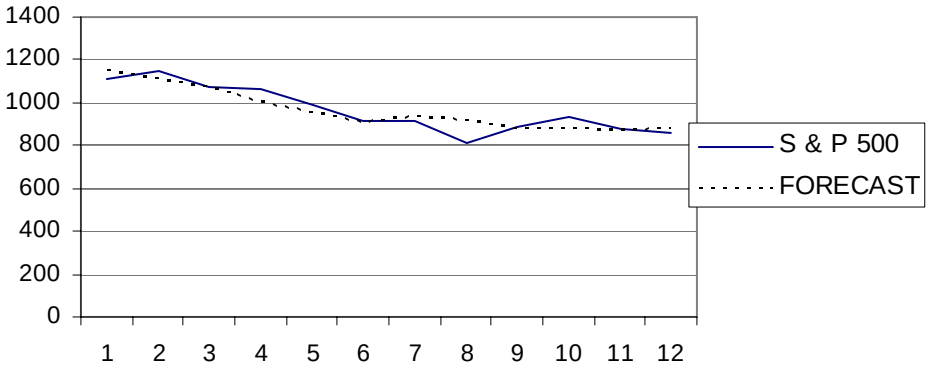

Table 1 LIST OF ECONOMIC VARIABLES

VARIABLE	DESCRIPTION
CPIAUCNS	CONSUMER PRICE INDEX
PPIACO	PRODUCER PRICE INDEX
UMICHINFL	UNIV. OF MICHIGAN INFLATION INDEX
CONSENT	CONSUMER SENTIMENT INDEX
INDPRO	INDUSTRIAL PRODUCTION INDEX
TCU	TOTAL CAPACITY UTILIZATION INDEX
NAPM	NAPM INDEX
UNRATE	UNEMPLOYMENT INDEX
HOUST	HOUSING STARTS
OILPRICE	OIL PRICES
TWEX	TRADE WEIGHTED EXCHANGE RATES
MZM	MZM BROAD MONEY SUPPLY
M1	M1 MONEY
M2	M2 MONEY
M3	M3 MONEY
LAG	LAGGING INDEX
COIN	COINCIDENT INDEX
LEAD	LEADING INDEX
PRIME	PRIME RATE
DISRATE	DISCOUNT RATE
FEDFUND	FEDERAL FUNDS RATE
3MTBILL	3 MONTH TREASURY BILL RATE
6MTBILL	6 MONTH TREASURY BILL RATE
1YRT	1 YEAR TREASURY BOND RATE
5YRT	5 YEAR TREASURY BOND RATE
7YRT	7 YEAR TREASURY BOND RATE
10YRT	10 YEAR TREASURY BOND RATE
20YRT	20 YEAR TREASURY BOND RATE

TABLE 2 SIGNIFICANT INDEPENDENT COMPONENTS AND THEIR TRANSFORMATIONS

LEAD 1 MONTH	LEAD 2 MONTHS	LEAD 3 MONTHS	LEAD 4 MONTHS	LEAD 5 MONTHS	LEAD 6 MONTHS
1 Linear	1 Linear; Exp.	1 Linear; Tan h	1 Linear	1 Linear; Exp.	1 Linear
2 Tan h	2 Powr 4	2 Linear	3 Linear; Tan h	2 Powr 4	2 Linear
3 Linear; Powr 4	3 Linear; Tan h	3 Linear	4 Powr 4	3 Linear	3 Linear
4 Powr 4; Tan h	4 Linear; Powr 4	4 Tan h	5 Linear	4 Linear; Tan h	4 Linear; Powr 4; Tan h
5 Linear	5 Linear	5 Linear; Powr 4	6 Linear	5 Linear	5 Tan h
6 Linear	7 Linear	6 Powr 4	12 Linear	6 Tan h	6 Tan h
11 Tan h	12 Linear	7 Powr 4	15 Powr2	7 Powr 4	8 Linear
13 Linear	20 Powr 4	9 Powr 4	20 Linear; Tan h	8 Tan h	10 Fzrt
20 Linear		10 Powr 4		9 Powr 4	13 Tan h
		12 Linear		10 Powr 4	14 Linear; Powr 4
		13 powr 4		12 Powr 4	15 Fzlft
		16 Exp.		13 Powr 4	20 Linear; Powr 4
		20 Linear		17 Linear	
				20 Linear	

Table 3 Statistics On Errors: Forecast-Actual

Panel A: 1 Month Ahead Forecast

	Based On Independent Components		Based On Deflated Signals	
	Full Sample	Last 12 Obs.	Full Sample	Last 12 Obs.
Mean	-8.66	-33.36	1.73	-4.67
Std. Dev.	105.19	133.09	34.52	45.09
Minimum	-341.95	-341.95	-105.95	-68.18
Maximum	521.89	191.47	195.69	95.38
No.of Obs.	242	12	242	12

Panel B: 2 Month Ahead Forecast

	Based On Independent Components		Based on Deflated Signals	
	Full Sample	Last 12 Obs.	Full Sample	Last 12 Obs.
Mean	-10.14	Obs.	1.67	-2.18
Std. Dev.	127.65	20.47	37.55	51
Range	1267.86	75.68	310.63	197.41
Minimum	-830.41	-4.89	-122.96	-78.64
Maximum	437.45	70.79	187.67	118.78
No.of Obs.	242	12	242	12

Panel C: 3 Month Ahead Forecast

	Based on Independent Components		Based on Deflated Signals	
	Full Sample	Last 12 Obs.	Full Sample	Last 12 Obs.
Mean	-9.01	7.82	-0.88	-11.51
Std. Dev.	117.33	229.16	31.34	42.61
Range	879.19	730.79	306.93	137.78
Minimum	-529.66	-381.26	-112.29	-66.83
Maximum	349.53	349.53	194.64	70.94
No.of Obs.	242	12	242	12

Panel D:4 Month Ahead Forecast

	Based on Independent Components		Based On Deflated Signals	
	Full Sample	Last 12 Obs.	Full Sample	Last 12 Obs.
Mean	-11.96	-68.63	-1.4	23.77
Std. Dev.	129.79	204.68	47.53	52.41
Range	818.25	655.82	410.67	190.50
Minimum	-487.73	-325.29	-184.24	-76.19
Maximum	330.52	330.52	226.43	114.31
No.of Obs.	242	12	242	12

Panel E: 5 Month Ahead Forecast

	Based On Independent Components		Based On Deflated Signals	
	Full Sample	Last 12 Obs.	Full Sample	Last 12 Obs.
Mean	-9.37	-20.32	0.75	42.25
Std. Dev.	115.14	157.25	51.97	51.76
Range	951.87	474.13	482.50	157.89
Minimum	-462.79	-235.09	-210.76	-41.72
Maximum	489.07	239.04	271.74	116.16
No.of Obs.	242	12	242	12

Panel F: 6 Month Ahead Forecast

	Based On Independent Components		Based On Deflated Signals	
	Full Sample	Last 12 Obs.	Full Sample	Last 12 Obs.
Mean	-4.65	12.93	-1.37	5.04
Std. Dev.	88.60	106.67	47.14	42.88
Range	666.69	371.43	346.68	128.67
Minimum	-382.69	-191.79	-169.44	-57.16
Maximum	283.99	179.63	177.25	71.01
No.of Obs.	242	12	242	12