Independent Component Analysis for ensemble predictors with small number of models

Ryszard Szupiluk[1], Tomasz Ząbkowski[2], Krzysztof Gajowniczek[2]

(1) Warsaw School of Economics

(2) Warsaw University of Life Sciences, Faculty of Applied Informatics and Mathematics

Abstract

The article presents independent component analysis (ICA) applied to the concept of ensemble predictors. The use of ICA decomposition enables to extract components with particular statistical properties that can be interpreted as destructive (noises) or constructive for the prediction. The key issue of the presented method is the identification of noise components. For this purpose, a new method for evaluating the randomness of the signals was developed.

Introduction

In this paper we develop an independent component analysis (ICA) approach for ensemble predictions. Its main idea is based on decomposition of the prediction results into underlying independent components [1]. Some of these components may be associated with the true value prediction and some of them can be treated as noise or interference. Elimination of noises, termed as destructive components, should result in prediction improvement. The process can be perceived as data filtration aimed to reveal hidden noises in a way that is typical for blind source separation techniques [2]. Standard filtering using ICA involves the components separation into source signals (separation step), the identification and elimination of noise components, and then inverse procedure with respect to separation (remixing step).

Prediction improvement

We assume, that after the learning process, each prediction result includes two types of latent components: constructive $\hat{s}_j$, associated with the target, and destructive s_i, associated with the inaccurate learning data, individual properties of models, missing data, not precise parameter estimation, distribution assumptions. The relation between observed prediction results $\mathbf{x} = [x_1, x_2, ..., x_m]^T$ and latent components $\mathbf{s} = [\hat{s}_1, \hat{s}_2, ..., \hat{s}_k, s_{k+1}, s_n]^T$ can be expressed as linear transformation and can be assumed as

(1) $$\mathbf{X} = \mathbf{AS}$$

where matrix $\mathbf{A} \in R^{n \times n}$ represents the mixing system.

Our aim is to find the latent components and reject the destructive ones (replace them with zero). Next we mix the constructive components back to obtain improved prediction results as

(2) $$\hat{\mathbf{X}} = \mathbf{A}\hat{\mathbf{S}} = \mathbf{A}[\hat{s}_1, \hat{s}_2, ..., \hat{s}_k, \mathbf{0}_{k+1}, ..., \mathbf{0}_n]^T$$

The estimation A and S can be perform by ICA method [2] and the main problem after components **S** estimation is to identify destructive ones.

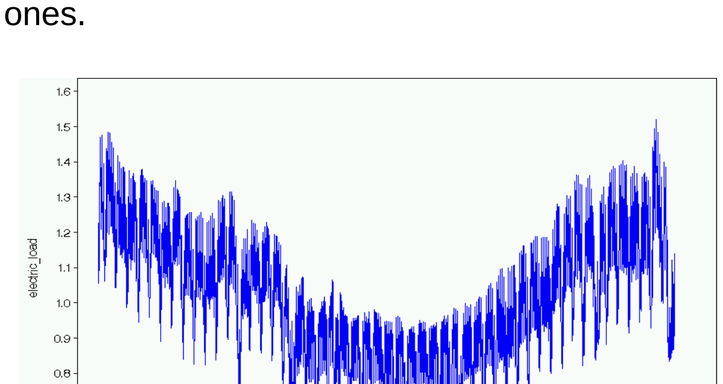

Fig.1. Hourly load observed in 1997.

Statistical analysis of destructive components

In the prediction improvement procedure, after the independent signals are identified the next step is to distinguish the informative elements from the noises and disturbances. Therefore, we need to classify latent components as constructive or destructive. In general, this can be difficult task because the obtained components might be somewhere between constructive and destructive. It means that particular component can have constructive impact on one model and destructive on the other.

Our approach is based on the assumption that the destructive components may be interpreted in terms of noise, for which we can deliver mathematical characteristics describing them. Therefore, to assess the similarity of the signal to the noise and its randomness we propose a general scheme for comparative analysis, described with the following formula:

(3) $$\phi_{1,2}(y_1, y_2) = \frac{1}{2}\frac{u(y_1 + y_2)}{u(y_1) + u(y_2)}$$

where $u(y)$ is a measure of randomness.

To explore directly the temporal characteristics of the noise signals we propose following variability measure:

(4) $$u(y) = \frac{\frac{1}{N}\sum_{k=2}^{N} |y(k) - y(k-1)|}{\rho(\max(y) - \min(y))}$$

where function $\rho(\cdot)$ was introduced to avoid dividing by zero.

The measure (4) has simple interpretation: it is maximal when the changes in each step are equal to range (maximal change), and is minimal when data are constant. The possible values are ranging from 0 to 1.

In case of multiple signals, their mutual similarity Φ $\phi_{i,j}(y_i, y_j)$ may be presented in a matrix form:

(5) $$\Phi = \begin{bmatrix} \phi_{1,1} & \cdots & \phi_{1,m} \\ \vdots & \ddots & \vdots \\ \phi_{n,1} & \cdots & \phi_{n,m} \end{bmatrix}.$$

Illustration of how the measure works:

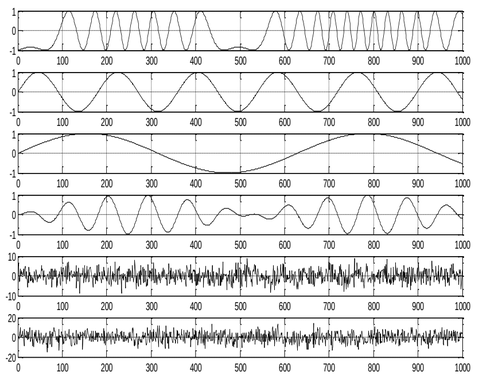

Fig.2. Source signals.

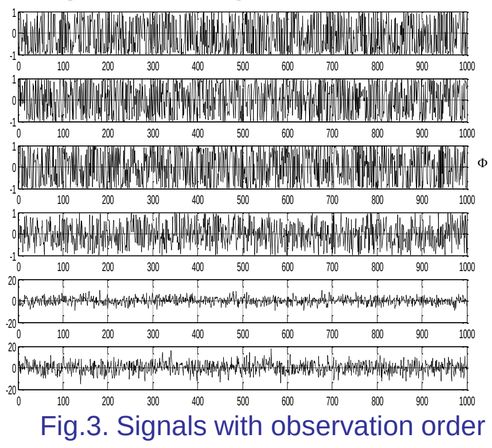

Fig.3. Signals with observation order randomly assigned.

According the matrix Φ_1, the last two noise signals (values in bold) are identified as these which have the highest similarity. Also, the value of similarity did not changed substantially after mixing the signals, as presented in matrix Φ_2 (in bold).

Numerical experiment – electricity demand forecasting

An important and utilitarian dimension of our work concerns the application for practical business problem which was electricity load forecasting on Polish market using hourly data from 1988 – 1998 (Fig 1). Forecasting electricity demand is an important issue from an economic point of view, due to the fact that direct financial incentives are related to the energy market. Any mismatch between the size of demand and supply results in tangible losses. Over estimation, due the storage problems, causes its irretrievable loss, while under estimating leads to urgent purchase on balancing market on which the prices are higher.

We trained six MLP neural networks (models M1 – M6) with one hidden layer (with 12, 18, 24, 27, 30, 33 neurons respectively) to create the **24 hours** ahead forecast based on historical energy demand and calendar variables such as month, day of the month, weekday, and holiday indicator.

Table 1. The results of BSS aggregation.

MAPE ×10^{-3}		M1	M2	M3	M4	M5	M6
Primary models		23.9431	**23.5021**	23.6750	23.9850	24.1374	23.5776
ICA	FPICA	23.8712	**23.0011**	23.5819	23.6068	24.0282	23.5715
	SANG	24.1180	**22.3682**	23.8860	24.0345	24.2017	23.7412
	JADE	24.1081	23.5996	23.7717	23.7653	22.5332	**22.3239**

The other part of the experiment was proposed to illustrate how the prediction improvement depends on the number of models used for BSS aggregation.

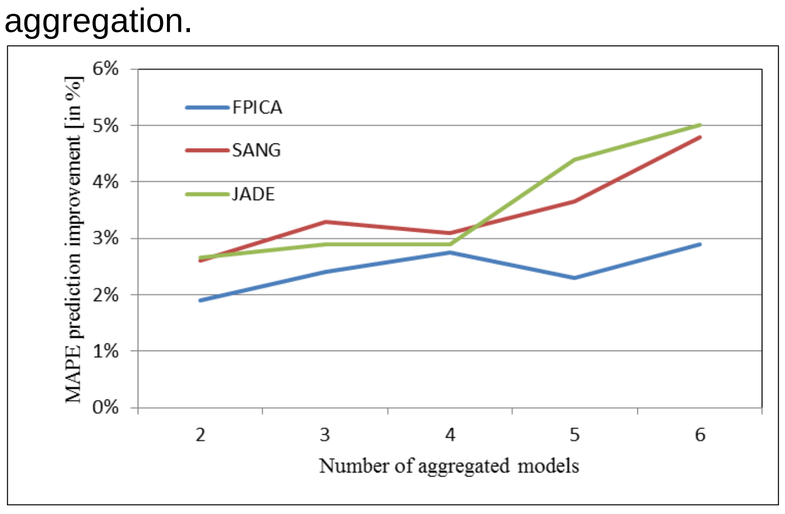

Fig. 4. Prediction improvement observed for different number of aggregated models (with MAPE error).

In general, we observed that with increasing number of aggregated models the efficiency of aggregation was improved.

Conclusions

Based on the electricity load data we showed that presented approach is effective for ensemble prediction taking into account MAPE error criteria and different number of aggregated models. As a result we could benefit of about 5% of MAPE reduction (best primary model vs. best model after decomposition) which may be regarded as significant improvement for power sector entities to maintain high efficiency in terms of balancing the market better.

It should be noted that the proposed method is adequate for a small number of aggregated models which meets typical requirement for ensemble methods.

References

[1] R. Szupiluk, P. Wojewnik, T. Zabkowski, *Lecture Notes in Computer Science* **3070**, 1199 (2004).

[2] A. Cichocki, S. Amari, *Adaptive Blind Signal and Image Processing*, John Wiley, Chichester 2002.

Authors: dr inż. Ryszard Szupiluk, e-mail: rszupi@wp.pl; dr inż. Tomasz Ząbkowski, e-mail: tomasz_zabkowski@sggw.pl; mgr Krzysztof Gajowniczek, e-mail: krzysztof_gajowniczek@sggw.pl.