# Local Linear Correlation Analysis with the SOM

Antonio Piras, Alain Germond

Swiss Federal Institute of Technology
Electrical Power Systems Laboratory
1015 Lausanne, Switzerland

**Abstract**

The purpose of this paper is to illustrate methods which can be used to select relevant input variables for non-linear regression. Relevant variables are ones which are though to explain in some extent the behaviour of the process we want model. We propose a method based on the self-organised map which allows to compute the local linear correlation between variables in neighbour subspaces. The graphical ordered plot of the obtained local correlations allows to study the non-linear dependencies of variables and suggests relevant ones.

# 1 Introduction

One of the most important task in building a mathematical model of a process for prediction purposes is the selection of relevant input variables. In order to obtain a good set of representative variables we have to assume that we effectively can measure or observe all variable that we need, and that the data is not contaminated by outliers. There is not a systematic way which can be followed, but some statistical methods like auto-correlation, partial-auto-correlation, and cross-correlation have been used successfully.
Basically, admitting that we can evaluate with a certain criterion when a model is better than another, we have the following computational choices to find a good subset of variable from the M original ones:

- all possible regression, which consists in trying all $2^M$ combinations. This is a method restricted to the case where we have a small number of candidate variable M, and the estimation procedure run sufficiently fast,
- forward selection, which consists in the repetitively inclusion of a new variable, until a stop criterion is satisfied, where at any step the added variable is the one yielding the biggest increase in accuracy for the model,
- backward elimination, which consists in eliminating one variable at a time, starting with a model containing all variables, the eliminated variable being the one changing the least the accuracy of the model,
- combined selection and elimination, where at each step we can consider the addition and the deletion of a variable simultaneously.
- genetic selection, as an emerging methodology which consists in trying to optimise the inputs set with genetic algorithms.

The selection would be simple if the regressors were orthogonal, that is possessing a diagonal covariance matrix, because in this case we could compute how much of the predictor variance is due to each separated regressors. Generally, the input variables are to some extents collinear, that means they measure the same thing and are correlated with the output variables in the same way. In the linear case it would mean that the covariance matrix of the estimated parameters will be large, causing a loss of confidence in prediction quality outside the estimation set.

In the next chapters we propose a method based on the self-organised map which allows to compute and analyse the local linear correlation between variables in neighbour subspaces. The method suggests variables which could be candidates to be included in the input subset but does not say which one have to be finally considered in the model. In chapter 2 we review the linear correlation concept with its limitations, in chapter 3 we formulate the local linear correlation method and in chapter 4 we apply it to the problem of choosing relevant variables for electrical load forecasting.

# 2 Linear Correlation

Let consider two variables *x* and *y* on the same graph as function of the time sampled at t= {1,..,N}. We can say that they are related if they move in the same direction, or in the opposite direction simultaneously. We can measure this belief with the correlation factor, yielding us an index varying between [-1,1], which indicates how much two variables are related. A value equal to 0 indicates no correlation, a value equal to 1 a positive correlation (variables move in the same direction) and -1 a negative correlation (variables move in different directions). However this will not say us that two variables are dependent with a cause/effect relation, just only that they are linear correlated and when the index is equal to zero that are not correlated at all.
To compute the correlation factor we first subtract the variables by their means, then compute the variance with

$$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N} x_i^2 \tag{1}$$

and the covariance with

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N} x_i y_i \tag{2}$$

and the linear correlation, or cross-correlation coefficient $r_{xy}$ independent from the units in which variables are measured:

$$r = \frac{1}{N-1}\sum_{i=1}^{N} \frac{x_i y_i}{S_x S_y} = \frac{S_{xy}}{S_x S_y} \tag{3}$$

As example the relation $y = x$ has linear correlation equal to one, instead $y = x^2$ ( $x = [-k, k]$ where *k* is a constant) has linear correlation equal zero.
It is important to retain that these statistics converge to the population statistics for large *N*, thus attention must be paid when computing and interpreting correlation with small *N* (< 100).
Linear correlations can be used to detect a linear relation between two variables or two time series at different time lags, which is known as *correlogram*. The linear correlogram is however unable to tell us if the correlation changes over the definition space. Next, we propose to analyse the correlation between variables by mean of a locally computed correlation coefficient.

# 3 Local Linear Correlation

The case that two variables are not linear correlated does not mean that they are not non-linear correlated. It could happen that in some regions of the definition space the variables are correlated and in some other not. It would be nice to monitor how the correlation changes between two adjacent subspaces, as example consider the relation between temperature and electrical energy consumption, which is negative in winter ( colder = higher consumption due to heating ) and positive in summer ( warmer = higher consumption due to cooling ). Over the whole

year the correlation would be zero when computed with the linear correlation factor ! However if we look to the variation of the linear correlation computed in small subspaces ( e.g. one or two months ) we will notice that the linear correlation changes smoothly, indicating non-constant correlation.
For the moment let consider a vector quantization algorithm which divide the manifold where $\mathbf{v} = (x, y)$ is defined (e.g. $\Re^2$ for the two dimensional case ) in $S$ subspaces. This is usually obtained by an iterative procedure which approximates the solution for $t \to \infty$, and the results are the centers $\mathbf{m}_i$ of the subspaces $\mathbf{v}_i, i = \{1,..,S\}$, also called code book or prototypes. It would also be possible to separate the subspaces by hand, if one would know exactly which data (and under which criterion ) belongs more to a group than an another.
The recursive equation for the computation of $\mathbf{m}$ (omitting the index *i* for simplicity) is

$$\mathbf{m}(t+1) = \mathbf{m}(t) + \varepsilon(t)(\mathbf{v}(t) - \mathbf{m}(t)). \tag{4}$$

Similar, we can derive an expression for the computation of the variances of each cluster:

$$[S_x^2(t+1), S_y^2(t+1)] = \mathbf{s}^2(t+1) = \mathbf{s}^2(t) + \varepsilon(t)\left(\left(\mathbf{v}(t) - \mathbf{m}(t)\right)^2 - \mathbf{s}^2(t)\right), \tag{5}$$

and thus also the covariance between *x* and *y*

$$S_{xy}(t+1) = S_{xy}(t) + \varepsilon(t)\Big((x(t) - m_x(t))(y(t) - m_y(t)) - S_{xy}(t)\Big), \tag{6}$$

where $\varepsilon(t)$ is the adaptation parameter, or learning rate, which is function of the time, and for the Kohonen self-organising methods is also function of a neighbourhood value and other variables, for details see e.g. [1].
At the end of the estimation phase we obtain as usual the correlation coefficient which is equal to the covariance of *x* and *y* divided by the own variances:

$$r = \frac{S_{xy}}{S_x S_y} \tag{7}$$

Note that when using topological preserving algorithms, we can tune $\varepsilon(t)$ so that the correlation coefficient overlap the subspaces. In this case *r*, can be viewed as a weighted measure of neighbouring correlations.

## 3.1 Measuring local linear correlations

We now need to sum up the local linear correlations with a single index in order to be able to discover quickly the important variables. It is probable that the correlations in some clusters will be negative and in others positive and also that the absolute values are small and big, thus the simple mean of the correlations is not satisfactory. Instead, we want to privilege big correlations or take the strongest ones.
A first index can be computed as the root of the mean squared correlations (RMSC) as following:

$$R_{RMSC}(x, y) = \sqrt{\frac{1}{S}\sum_{i=1}^{S} r_i^2} \tag{8}$$

where $S$ is the number of clusters and $r_i$ is the correlation between *x* and *y* in the cluster $\mathbf{v}_i, i = \{1,..,S\}$.

A second possible index could be the maximum of the absolute correlations (MAC) and is computed as

$$R_{MAC}(x, y) = \max_{i=\{1,..,S\}} \left| r_i \right| \tag{9}$$

Additionally to the indexes it is also important to look to the correlations in the neighbouring preserving maps, this may help, as we will see below, to understand the process under analysis.

# 4 Application to Electrical Load Forecasting

The role of an electrical utility is to supply its customers with the required electrical energy all over the year, under different possible scenarios. The optimisation of the production resources and the contractual interchanges of electrical energy with neighbour utilities is based mainly on the short term forecast of the electrical load [2]. This forecast, with an horizon ranging from one hour up to one week allows to schedule the own machines, e.g. hydroelectric or gas turbines, and in case of insufficient production to buy the additional energy somewhere else or in case of over-production to sell it.

The time series is characterised by a strong seasonal behaviour, by a weekly and a daily periodicity. The presence of a non-linear trend, depending from year to year, on economical or political factors shifts the mean up or down

## 4.1 Description of the available information sources

The Western Swiss electrical utilities are equipped with SCADA systems for automated data measurement and registration since a ten of years. Data consists of hourly integrated electrical power consumption, so that each day is composed by a vector of 24 values.

On the other side, the Swiss institute of meteorology disposes of meteorological stations all over the Swiss territory, allowing an on-line measurement of different weather variables. The weekly forecast of some meteorological variables is available as an additional service on request and is scheduled early in the morning of every day, whose results can be obtained by fax or by internet.

After discussion with expert engineers, working in the dispatching center, and a review of common used variables [3], only the minimal and maximal temperatures of a day were retained for further study.

## 4.2 Local Linear Correlation Analysis

We expect that the correlation between temperatures and electrical load change during the year in a ordered fashion. In order to visualise the phenomena we will use a closed Kohonen ring, and train it, so that adjacent neurons will represent neighbour periods of the year. The analysis of the correlation coefficient is done for each hour $h=\{1, ...,H\}$ by a separate networks and the input vector is composed by the following variables:

- the maximal and minimal temperatures of the forecast day and of the day before of three geographical recording points: TMIN1(-1), TMAX1(-1), TMIN2(-1), TMAX2(-1), TMIN3(-1), TMAX3(-1), TMIN1(-1), TMAX1, TMIN2, TMAX2, TMIN3, TMAX3,
- the autoregressive terms Y(-10) and Y(-1) , that is the ten days before load and one day before load at the same hour $h=\{1, ...,H\}$,
- the load of the forecast day at hour $h=\{1, ...,H\}$ , indicated by Y.

where each variable has been previously filtered from the seasonal term, so that we are working with the residual terms.

One networks with 12 units has been trained with a set composed of two year of historical data, for a total of about 700 patterns. We obtain first the mean and the standard deviation of the input

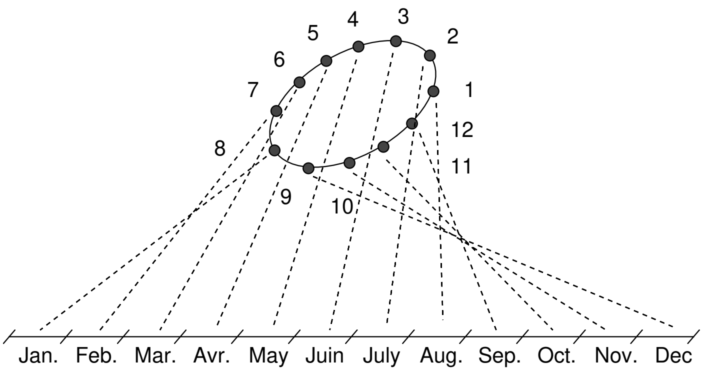


Figure 1 Organisation of the Kohonen ring after learning is terminated.

vector in each cluster, and secondly the covariance and the correlation coefficient between the load Y and the rest of variables in the input vector according to the equations (4)-(6).
Fig. 1 shows the conceptual organisation of the Kohonen ring with 12 neurons, which was deduced by observing the clustered values of input components. Approximately each neuron represents a region corresponding to a month, so that neuron N. 8 is centered around January, and going backwards (7-6-5-4-3-2-1-12-11-10-9-8 ) each month is traversed. In this way the Winter period is represented by neurons 7-8-9, the Summer by neurons 1-2 , Spring-time by 3-4-5-6 and Autumn by 10-11-12.

In Fig 2 (left) the obtained correlations for the hour 1 a.m. are plotted. First one should note that the correlation are not constant, instead the correlations with the temperatures change regularly from 0.2 to -0.8, that is from a not significant correlation to a clear negative correlation in Winter (a negative correlation means that if the temperature is lower than its seasonal average then the load correlated response is higher than the seasonal average ). Secondly one should see that the auto-regressive terms Y(-1) is always significantly positive correlated, that is, over the year the next day's deviation of the load from seasonal pattern has the same sign than the deviation of today. Instead the ten days old deviation Y(-10) is not correlated at all with the today load.

In Fig. 2 (right) the results of the correlations for the hour 11 a.m. are plotted. We can derive the same conclusions as for hour 1 a.m., with the exception that if for hour 1 a.m. the minimal temperature was more important than the maximal one, for hour 11 a.m. it is just the opposite ( which is logically correct, because night hours are sensible to minimal temperatures and daylight hours are sensible to maximal hours).
In order to quantify the importance of each separate variables we compute the RMSC and the MAC for the networks with different number of neurons. As an example consider the Table 1 indicating the MAC statistics for hour 11 a.m. and for the networks with 1 and with 12 neurons respectively.
We can see that the locally computed correlation (MAC(12)) is bigger than the average global correlation (MAC(1)), up to values of about 0.8, which is a significant linear correlation which otherwise would not have been considered. ( We knew that the temperature is an important explanatory variable, but this fact can not be shown with the simple linear correlation coefficient computed over a whole year of observations ).

# Conclusions

In this paper we presented a method to study the linear correlation between variables in neighbour subspaces based on the self-organising map. The visualisation of the local linear correlation computed at each units of the self-organising map ( or ring ) allows to understand the varying dependency between variables over the whole definition manifold.

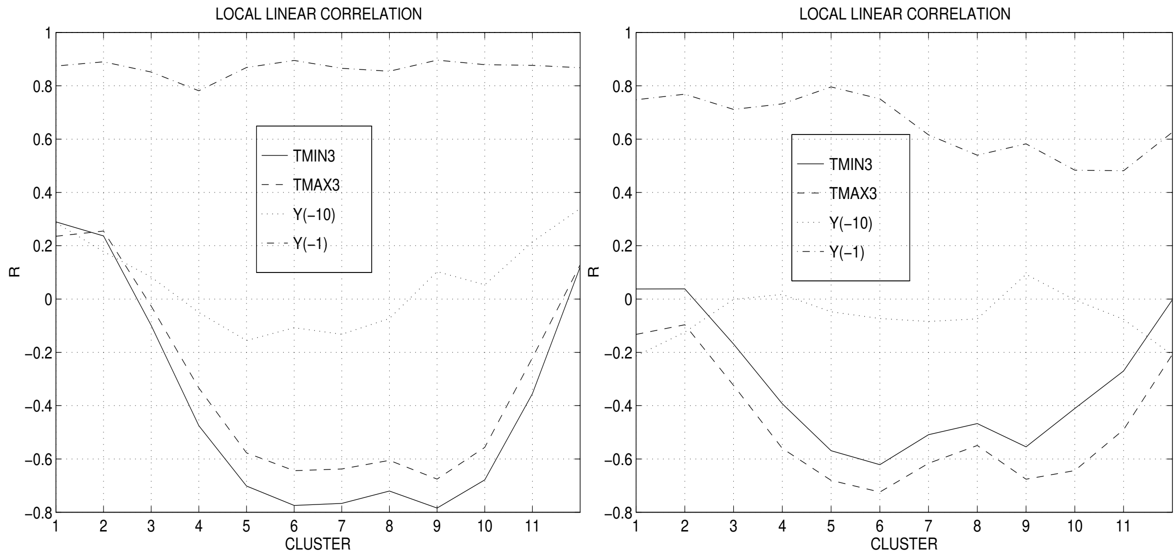


Figure 2 Local cross-correlations of the load at 1 a.m. (left)
and 11 a.m. (right) with other 4 variables.

Table 1: MAC Statistics

| VARIABLE | MAC(1) | MAC(12) |
|---|---|---|
| TMIN1 | 0.17 | 0.49 |
| TMAX1 | 0.42 | 0.74 |
| TMIN2 | 0.09 | 0.40 |
| TMAX2 | 0.42 | 0.73 |
| TMIN3 | 0.25 | 0.62 |
| TMAX3 | 0.40 | 0.72 |
| Y(-10) | 0.10 | 0.21 |
| Y(-1) | 0.43 | 0.79 |
| Y | 1.00 | 1.00 |

# Acknowledgements

This work was funded by ABB Network Partner Ltd., Switzerland. We are grateful to the Swiss Institute of Meteorology for the availability of climate related data, to l'Energie de l'Ouest Suisse (EOS) for the electrical load data.

# References

[1] T. Kohonen, *Self-Organizing Maps*. Berlin: Springer-Verlag, 1995.

[2] A. Piras, A. Germond, B. Buchenel, et al., "Heterogeneous Artificial Neural Networks for Short Term Load Forecasting," *IEEE Transactions on Power Systems*, vol. 11, pp. 397-402, 1996.

[3] T. Czernichow, A. Piras, K. Imhof, et al., "Short Term Electrical Load Forecasting with Artificial Neural Networks," *Eng. Int. Syst.*, vol. 2, pp. 85-89, 1996.