Introduction to multilayer perceptron and hybrid hidden Markov, multilayer perceptron, models.

Joseph Rynkiewicz*

27 juin 2002

1 Introduction.

A time series is a succession of measures that constitute equidistant measures in time. For example, it can be made up of CAC 40 index prices, taken on a daily basis, or else a country's GDP, measured on a year-by-year basis. It is quite obvious that it would be a major advantage if we could guess how these series are going to behave. Statisticians have therefore developed tools to model this behaviour and to try to optimise their predictions of the future values of the process being observed. In the present paper we will be studying the contributions that neural networks, and more specifically multilayer perceptrons (MLP), have made to time series. In the first section, we will mostly be looking at the MLPs' selection of architecture. This will allow us to avoid any over- parameterisation of the model, something that would cause over-learning. In the second section, we will be focusing on an example of a piecewise stationary time series which requires that we use several regression models simultaneously and choose, in a probabilistic manner and at any given moment in time, which of these models makes the most relevant prediction. Lastly, we will be applying these methods to the modelling of a pollutant emission series relating to ozone levels in Paris.

2 Auto-regressive models

The present paper looks at the parametric modelling of time series. More specifically, we will be studying those models that use multilayer perceptrons (MLP) as a regression function.

Consider the following times series model :

$$Y_{t+1} = F_{W_0}(Y_t) + \varepsilon_{t+1}$$

where

- $Y_t \in \mathbb{R}$ is the observation at time "t" of the time series.
- ε_t are independent and identically distributed (i.i.d.) random variables with null expectation and a constant variance σ^2, for example a variable $\mathcal{N}(0, \sigma^2)$ independent from the series' past.

*SAMOS/MATISSE, University of ParisI, 90 rue de Tolbiac, Paris, France, rynkiewi@univ-paris1.fr

– F_{W_0} is a function represented by a MLP whose parameters (weighting) is the vector $W_0 \in \mathbb{R}^D$.

To simplify the notation, we will only be considering self-regressive models of the first order. However, it would be easy to generalise at a higher order. As such, the phenomenon we observe (Y_t) is the combination of a deterministic function out of the process's past and of a random event. If we are familiar with the underlying deterministic function F_{W_0} we can optimise the predictions we make. Given that this function is entirely determined by its vector parameter, the statistician's job consists of estimating the parameter, W_0 using a finite number of observations $(y_0, \cdots, y_T)$.

To do this, we minimise in W a functional such as

$$S_T(W) = \frac{1}{T} \sum_{t=1}^{T} \left(Y_t - F_W\left(Y_{t-1}\right)\right)^2$$

the average of the residual squares and we note

$$\hat{W}_T = \arg\min_W S_T(W)$$

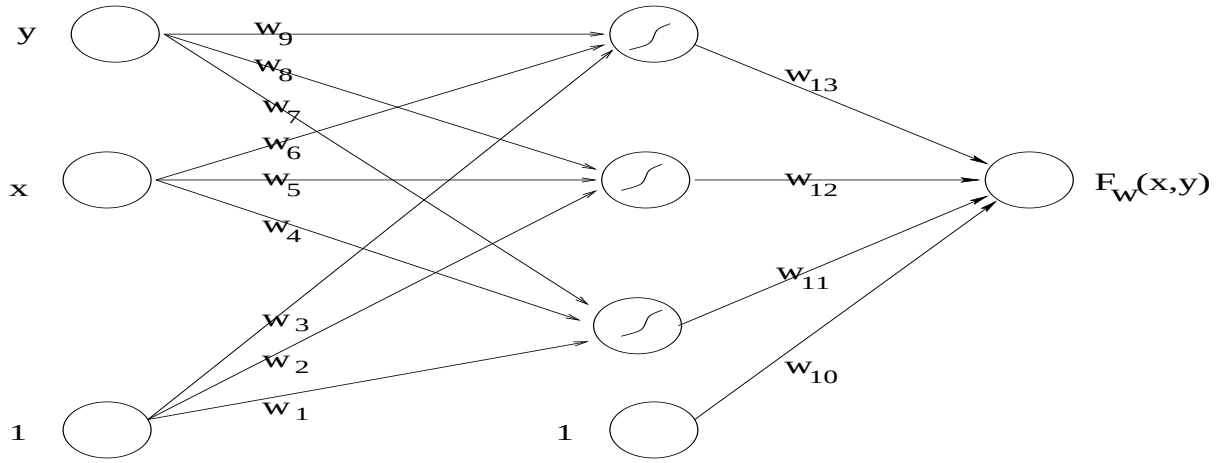
the least squares estimator of W_0.

2.1 Theoretical findings of an MLP with a hidden layer)

2.1.1 The MLP model

A function can be represented by an MLP in the following manner :

FIG. 1 – Multilayer perceptron



Here the MLP is represented by a function of $\mathbb{R}^2 \to \mathbb{R}$ which at (x, y) associates $F_W(x, y)$ with

$$F_W(x, y) = w_{10} + w_{11}\phi\left(w_1 + x \times w_4 + y \times w_7\right) + w_{12}\phi\left(w_2 + x \times w_5 + y \times w_8\right) + w_{13}\phi\left(w_3 + x \times w_6 + y \times w_9\right)$$

The activation function ϕ of the hidden layer is generally a sigmoid function that we can consider (without any loss of generality) as being equal to the hyperbolic tangent. As such, here the vector parameter is $W = (w_1, \cdots, w_{13})$.

From here on in, we will be assuming that our model is identifiable, meaning that there can only be one representative vector parameter for any function that can be represented by a given MLP. To obtain this property, all of the parameters will have to be restrained in a suitable fashion (cf Sussmann [18]).

2.1.2 Statistical properties

We are interested in the behaviour of the estimator $\hat{W}_T$ when T tends towards infinity. It is useful to have two fundamental properties at our disposal :

- Consistency, meaning that $\hat{W}_T \overset{T\to\infty}{\longrightarrow} W_0$
- The asymptotic normality ensures that the preceding convergence will take place at a rate of $\sqrt{T}$, and makes it possible to obtain a limit law of $\hat{W}_T$. For example, we can find in Yao [19] a demonstration of the following theorem :

Theorem 1 *Consistency and asymptotic normality of the estimator $\hat{W}_T$ avec $\phi(x) = \tanh(x)$, we assume that :*

1. *$(\varepsilon_t)_{t\in\mathbb{N}^*}$ is a series i.i.d. such that $E\varepsilon_t^6 < \infty$,*
2. *W_0 belongs inside of a compact sub-set of Euclidian space $\mathbb{R}^D$.*
3. *If μ_0 is the stationary measurement of (Y_t), a matrix with a dimension of $m \times m$*

$$\Sigma_0 = \int_{\mathbb{R}} \left[\frac{\partial}{\partial w_i} F_{W_0}(y) \frac{\partial}{\partial w_j} F_W(y) \right]_{1\le i,j\le m} \mu_0(dy)$$

is definite positive.

In which case :

- *The estimator $\hat{W}_T$ will almost surely converge towards W_0 when T tends towards $+\infty$.*
- *The term $\sqrt{T}\left[\hat{W}_T - W_0\right]$ converges by law towards a multidimensional Gaussian distribution $\mathcal{N}\left(0, \Sigma_0^{-1}\right)$.*

2.1.3 Identification of the model

One of the main problems encountered when using increasingly complex functions to estimate processes statistically is that the models can be overfitted. In actual fact, if we use an overly complex model in an area where too little data exists, we end up with a modelling of the noise that had generated the data. By so doing, we introduce a bias that strongly undermines the model's ability to make predictions using new and as yet unobserved data on the same process. An efficient statistical principle for fighting against the bias introduced by a complexification of models is the use of a penalty term that is itself a function of the number of parameters being applied (cf Akaike [1]).

Let us assume that an upper limit M exists for all of the possible dimensions of the model. This can be indicated as $(F_W)_{W\in\mathbb{R}^M}$ a family of dominant models with a dimension of M, so that a true parameter W_0, with a dimension of B, can be expressed as a vector of this family

with $M - B$ null components. Note $\hat{W}_T^L$ a least squares estimator with a dimension of L . The parsimony principle would then consist of choosing the estimator that minimises the new penalised cost function :

$$CP\left(W^L\right) = \frac{S_T\left(W^L\right)}{T} + \frac{c\left(T\right)}{T} \times L$$

or else

$$CP^*\left(W^L\right) = \frac{\ln\left(S_T\left(W^L\right)\right)}{T} + \frac{c\left(T\right)}{T} \times L$$

where $c\left(T\right)$ is the rate of penalisation. If $c\left(T\right) = 2$ the penalised contrast CP^* is then equal to Akaike's AIC criterion, if $c\left(T\right) = 2\ln\left(T\right)$, CP^* is equal to Schwarz's BIC criterion. [17]. Based on these definitions, we display the finding whose proofs can be found in Yao [19] or else in Rynkiewicz [15] :

Theorem 2 *We assume that the conditions of the theorem 1 have been verified. We assume also that the penalisation rate* $c\left(T\right)$ *is such that*

$$\lim_T \frac{c\left(T\right)}{T} = 0, \quad et \quad \liminf_T \frac{c\left(T\right)}{2\ln\ln T} > \sigma^2 \frac{\Lambda}{\lambda}$$

where Λ *(resp.* λ*) is the largest (resp. the smallest) proper value of the matrix* Σ_0*. In this case, the couple* $\left(L, W_T^L\right)$ *will almost surely converge towards the true value and the true dimension* $\left(L_0, W_0^{L_0}\right)$ *of the parameter when* T *will tend towards* ∞*.*

Based on this theorem, we can suggest a methodology that enables an almost certain identification, something which allows us to determine the true model.

2.2 Practical research on the true model

2.2.1 The search for a dominant model

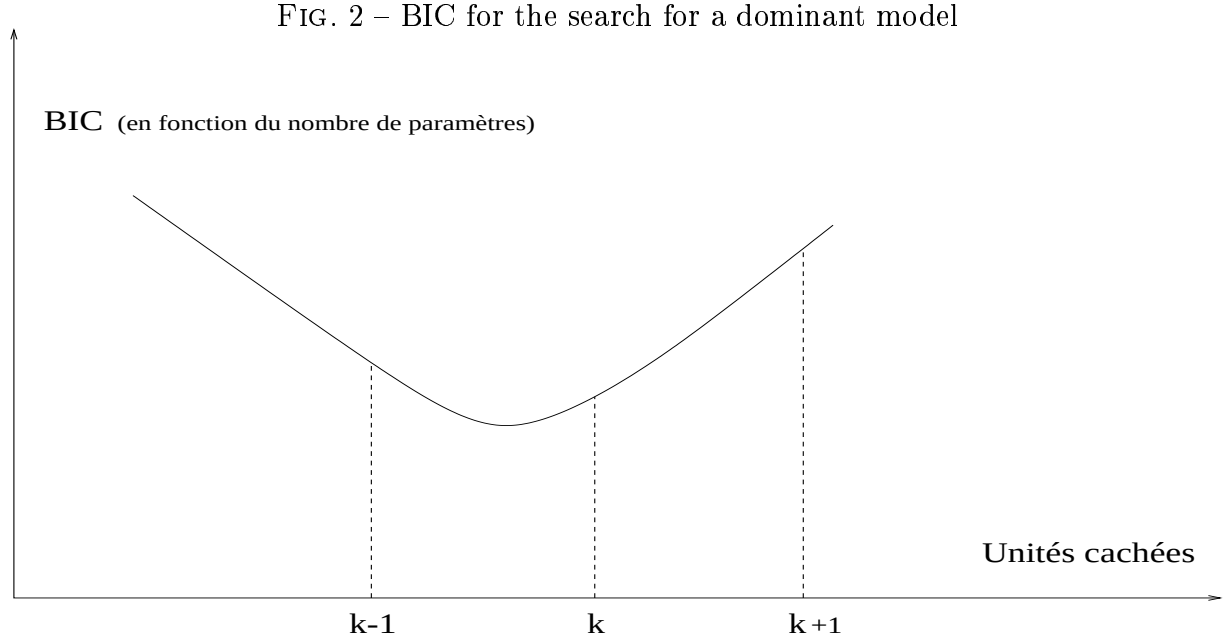
Using the findings from the preceding section, we can propose the following method for determining the true model. We launch the architecture by taking all of the relevant entries (as we would get them from a linear AR model) plus a single hidden unit. Then we progressively add units into the hidden layer, calculating the BIC criterion at each step. We continue this process as long as the BIC value drops. When the addition of another hidden unit causes a renewed upswing in the BIC, we stop the model search and construe this latest MLP as the dominant model. Schematically, this search can be represented by the following figure, which translates into a dominant model with $k + 1$ hidden units.

2.2.2 Determination of the true model

Once this dominant model has been obtained, we get the real model by successfully pruning away extra weight.

Remember that $W^M = (w_1, \cdots, w_M)$ is the vector parameter that is associated with the dominant model. In principle, to estimate the true model we should be comprehensively exploring the finite family of all of the submodels. However, this number is an exponentially large one, and this is the reason why to guide our research we are proposing, as is done in linear regression, a

FIG. 2 – BIC for the search for a dominant model



Statistical Stepwise Method (SSM). This kind of strategy is based on the asymptotic normality of the estimator $\hat{W}_T$ (cf Cottrell et al. [9]) that uses Student statistics as an aid for exploring subfamilies of the dominant model. To decide whether or not the weightings w_l should be eliminated, we compare the BIC values of the model F and those of the model F without the weightings w_l : F_l. As F_l is a submodel of F it suffices that the BIC criterion diminishes for us to move somewhat closer to the true model, which minimises the BIC criterion. By so doing, we obtain a series of MLPs with fewer and fewer parameters corresponding to a decreasing BIC trajectory. The criterion for ceasing our pruning exercise is quite straightforward, in that we will refuse to eliminate a weighting if this causes the BIC to rise. We keep the final MLP that maintained this weighting.

In short, the procedure involved in searching for a true model is as follows :

1. Determine F_{max} a dominant model
2. thanks to Student statistics specifying the weighting of l which is a candidate for elimination.
3. Accepting the elimination of this weighting if and only if the BIC criterion diminishes, otherwise keeping the last MLP model before this pruning took place.

This procedure has been tested with a number of examples and provides excellent results as long as the volume of data is sufficiently large, generally more than 500 observations.

3 Hybrid Models

3.1 Introduction

By modelling time series with the help of neural networks, we can account for any non-linearities that the model may contain. At the same time, this is based on a restrictive hypothesis

as regards the model's stationariness. One simple generalisation we could make would be to account for the piecewise stationary time series For instance, Hamilton [12] studied these kinds of models in an effort to model time series that are subject to discrete changes of regime. He used this as a way of analysing GNP series in the United States. We can therefore use such models for series featuring a particular regime for periods of economic growth, for example, and another one for periods of recession.

Although this model is more general than its predecessor, we still have need of a number of restrictive hypotheses. First of all, there has to be a finite number of possible regimes. Secondly, even though regime changes might take place, we will assume that such changes occur in a stationary manner, something that will ultimately enable us to make use of the law of large numbers and thus engage in statistical analysis.

3.2 The model

The theory of hidden Markov chains and their first applications in voice recognition are more than 30 years old. The basic theory was published in a series of articles written by Baum et al. ([4, 5, 3, 2]) in the late 1960s. Hidden Markov chains have subsequently been applied in a number of different fields, such as genetics, biology, economics, etc.

3.2.1 Markov chains in a discrete space

Consider $(X_t)_{t\in\mathbb{Z}}$, a homogeneous Markov chain with values in the finite state space $\mathbb{E} = \{e_1, \cdots, e_N\}$, $N \in \mathbb{N}^*$. Without any loss of generality, we can identify the state space $\mathbb{E}$ with the simplex of $\mathbb{R}^N$, where e_i is a vector unit of $\mathbb{R}^N$ with 1 on the i-th component and 0 everywhere else. The chain X_t is characterised by its transition matrix $A = (a_{ij})_{1\leq i,j\leq N}$ which is such that :

$$P\left(X_{t+1} = e_i \,|X_t = e_j\right) = a_{ij}$$

If additionally we were to define : $V_{t+1} := X_{t+1} - AX_t$, we would get the following notation for this model :

$$X_{t+1} = AX_t + V_{t+1}.$$

3.2.2 Equations of the model

Assume that the time series we have observed (Y_t) verifies the following equations :

$$\begin{cases} X_{t+1} = AX_t + V_{t+1} \\ Y_{t+1} = F_{X_{t+1}}(Y_t) + \varepsilon_{t+1}\left(X_{t+1}\right) \end{cases}$$

where $\{F_{e_1}, \cdots, F_{e_N}\}$ are functions of $\mathbb{R}^P \longrightarrow \mathbb{R}$ which will be represented in our example by MLP. For every $e_i \in \mathbb{E}$, $(\varepsilon_t(e_i))$ is a succession random variables that are independent and distributed identically. This model makes it possible to use several MLPs (here we would talk about a mixture of experts) and to use the Markov chain (X_t) in order to specify at a time "t" which MLP makes the most appropriate prediction. Note that since we can only observe the series (Y_t), we will have to find a way of reverting to the rest of the states of the chain (X_t), thanks to a (Y_t) behaviour.

To adjust this sort of model to our observations, we estimate the parameters (weightings of the MLPs F_{e_i} , the variance of noise $\varepsilon(e_i)$, and the transition matrix A) thanks to the method of maximum of likelihood. A study of the theoretical properties of this estimator can be found in Ryden and Krishnamurthy [13] and Douc, Moulines, Ryden [10].

3.3 Maximum of likelihood for the hybrid models

From here on in we will consider that the density of noise $(\varepsilon(e_i))_{1\leq i\leq N}$ is Gaussian in nature. We start by studying the free parameters that are under consideration :

- The transition matrix A, this matrix is stochastic, meaning that the sum of any given column A is 1. Thus there are no more than $(N-1)\times N$ free parameters.
- The variances $(\sigma_{e_i})_{1\leq i\leq N}$, which are supposed to be strictly positive.
- The parameters of the regression functions $(F_{e_i})_{1\leq i\leq N}$, since we are using MLPs, the parameters will obviously be the weighting vectors $(W_{e_i})_{1\leq i\leq N}$ of the MLP

The parameter vector θ will therefore be :

$$\theta = \left(W_{e_1}, \cdots, W_{e_N}, \cdots, a_{11}, \cdots, a_{(N-1)N}, \sigma^2_{e_1}, \cdots, \sigma^2_{e_N}\right)$$

3.3.1 Calculation of the log-likelihood and of its derivative

Following on from this we will be assuming that the first observation y_0 as well as the initial probability of the state X_1 are known, and that the conditioning of the expressions with respect to these initial conditions will always be implied.

An initial writing The likelihood of the model for a succession of observations of the series $y := (y_0, \cdots, y_T)$ for a supposedly fulfilled path of $x := (x_1, \cdots, x_T)$ is therefore :

$$L_\theta(y,x) = \prod_{t=1}^{T}\prod_{i=1}^{N}\left[\Phi_{e_i}\left(y_t - F_{e_i}\left(y_{t-1}\right)\right)\right]^{\mathbf{1}_{\{e_i\}}(x_t)} \times \prod_{t=1}^{T}\prod_{i,j=1}^{N} a_{ij}^{\mathbf{1}_{\{e_j,e_i\}}(x_t,x_{t+1})} \times \pi_0(x_1)$$

where Φ_{e_i} is the density of the normal law $\mathcal{N}(0,\sigma_{e_i})$, $\mathbf{1}_G$ the indicative function of the set G and π_0 the probability of the initial state x_1. To get the overall likelihood of the observations, we could add up all of these likelihoods along all of the potential paths of the hidden Markov chain. We would then have

$$L_\theta(y) = \sum_x L_\theta(y,x)$$

It is well known that the complexity of this sum is exponential, something that makes it difficult to make the calculation whenever the number of observations is more than a few hundred. It might also be possible to calculate the maximum likelihood thanks to the E.M. algorithm by using Baum and Welch's forward-backward algorithm. However, here we would prefer to use a differential optimisation technique that is generally faster than the E.M. algorithm when the regression functions being used involve multilayer perceptrons.

The log-likelihood A more elegant way to write the log-likelihood is to use the predictive filter $P\left(X_t = e_i \left| y_{t-1}, \cdots y_0 \right.\right) := p_t\left(i\right)$ since the likelihood will be written

$$L_\theta\left(y_1, \cdots, y_T\right) = \prod_{t=1}^{T} L_\theta\left(y_t \left| y_{t-1}, \cdots, y_0 \right.\right) = L_\theta\left(y_t \left| y_{t-1} \cdots, y_0 \right.\right) \times \prod_{t=1}^{T-1} L_\theta\left(y_t \left| y_{t-1}, \cdots, y_0 \right.\right)$$

$$= \sum_{i=1}^{N} L(y_T \left| X_T = e_i, y_{T-1}, \cdots, y_0\right) P(X_T = e_i \left| y_{T-1}, \cdots, y_0\right) \times \prod_{t=1}^{T-1} L(y_t \left| y_{-p+1}, \cdots, y_{t-1}\right).$$

Note that

- p_t the vector whose i-th component is : $p_t(i) = P(X_t = e_i \left| y_{t-1}, \cdots, y_0\right)$
- b_t the vector whose i-th component is : $b_t(i) = L(y_t \left| X_t = e_i, y_{t-1}, \cdots, y_0\right)$, i.e., the conditional density of y_t given that $X_t = e_i$ and $(y_{t-1}, \cdots, y_0)$.
- u^T the transposed vector u.

This gives us :

$$L(y_1, \cdots, y_T) = b_T^T p_T \times \prod_{t=1}^{T-1} L(y_t \left| y_{t-1}, \cdots, y_0\right) = \prod_{t=1}^{T} b_t^T p_t.$$

From this we deduce a practical form of the log-likelihood :

$$\ln(L(y_1, \cdots, y_n)) = \sum_{t=1}^{n} \ln(b_t^T p_t). \tag{1}$$

All we have to do is to calculate p_t for $t = 1, \cdots, n$, To be able to calculate the log-likelihood, since :

$$b_t(i) = L(y_t \left| X_t = e_i, y_{t-1}, \cdots, y_0\right) := \Phi_{e_i}(y_t - F_{e_i}(y_{t-1}))$$

Calculation of p_t In light of $B_t = diag\left(b_t\right)$, the diagonal matrix whose diagonal is the vector b_t, we can easily verify (cf Rynkiewicz [16, 15]) that the predictive filter p_t verifies the recurrence :

$$p_{t+1} = \frac{AB_t p_t}{b_t^T p_t}. \tag{2}$$

We will assume that p_1 follows a uniform distribution over $\{1, \cdots, N\}$ and we will therefore be able to calculate p_t, $t = 1, \cdots, T$ by means of recurrence. The choice of an initial value of p_t is relatively unimportant thanks to the initial distribution's exponential forgetting property (cf Legland et Mevel [14]).

3.4 Derivative of the log-likelihood

Remember that we have

$$\ln(L(y_1, \cdots, y_T)) = \sum_{t=1}^{T} \ln(b_t^T p_t)$$

thus, if we write θ_j the j-th parameter of the model, we get

$$\frac{\partial \ln(L(y_1, \cdots, y_n))}{\partial \theta_j} = \sum_{t=1}^{T} \frac{\frac{\partial b_t^T p_t}{\partial \theta_j}}{b_t^T p_t}.$$

All we have to do then is calculate $\frac{\partial b_t^T p_t}{\partial \theta_j}$ in order to be able to calculate the derivative of the log-likelihood, being aware of the fact that :

$$\frac{\partial b_t^T p_t}{\partial \theta_j} = \frac{\partial b_t}{\partial \theta_j}^T p_t + b_t^T \frac{\partial p_t}{\partial \theta_j}. \tag{3}$$

We can find the details of the calculation of this derivative in Rynkiewicz [15]. Here we will be simply explaining the calculation of the derivative of the filter :

Calculation of $\frac{\partial p_t}{\partial \theta_j}$ following θ_j Since we have the recurrence

$$p_{t+1} = \frac{AB_t p_t}{b_t^T p_t}$$

By deriving this expression from the parameter θ_j , we find :

$$\frac{\partial p_{t+1}}{\partial \theta_j} = \frac{\partial AB_t p_t}{\partial \theta_j} \times \frac{1}{b_t^T p_t} + AB_t p_t \times \frac{\partial b_t^T p_t}{\partial \theta_j} \times \left(-\frac{1}{(b_t^T p_t)^2} \right)$$

or

$$\frac{\partial p_{t+1}}{\partial \theta_j} = \left(\frac{\partial AB_t}{\partial \theta_j} p_t + AB_t \frac{\partial p_t}{\partial \theta_j} \right) \times \frac{1}{b_t^T p_t} + AB_t p_t \times \left(\frac{\partial b_t^T}{\partial \theta_j} p_t + b_t^T \frac{\partial p_t}{\partial \theta_j} \right) \times \left(-\frac{1}{(b_t^T p_t)^2} \right).$$

We then have :

$$\frac{\partial p_{t+1}}{\partial \theta_j} = \frac{AB_t}{b_t^T p_t} \left[I - \frac{p_t b_t^T}{b_t^T p_t} \right] \frac{\partial p_t}{\partial \theta_j} + \left(\frac{\partial AB_t}{\partial \theta_j} \right) \frac{p_t}{b_t^T p_t} - \frac{AB_t p_t}{(b_t^T p_t)^2} \left(\frac{\partial b_t^T}{\partial \theta_j} p_t \right)$$

hence :

$$\frac{\partial p_{t+1}}{\partial \theta_j} = \frac{AB_t}{b_t^T p_t} \left[I - \frac{p_t b_t^T}{b_t^T p_t} \right] \frac{\partial p_t}{\partial \theta_j} + \left(\frac{\partial A}{\partial \theta_j} B_t + A \frac{\partial B_t}{\partial \theta_j} \right) \frac{p_t}{b_t^T p_t} - \frac{AB_t p_t}{(b_t^T p_t)^2} \left(\frac{\partial b_t^T}{\partial \theta_j} p_t \right) \tag{4}$$

with, if p_1 is the initial distribution : $\frac{\partial p_1}{\partial \theta_j} = 0$ for every j.

The rest of the calculation of the derivative should not cause any problems. As such, we can calculate, for a reasonable calculation cost, the log-likelihood and its derivative. This enables us to apply a wide array of differential optimisation techniques in order to get closer to the maximum (or to the local one, at least) of the log-likelihood.

4 Application : Study of pollution, defined in level of ozone

The purpose of this study is to predict the maximum daily pollution rate, defined in the level of ozone, between the months of April and September, included. Towards this end, we will be using the previous day's maximum of the pollution rate (i.e., of the ozone level) as our regressor - plus the following meteorological observations :

- Total radiation
- Average daily wind speed
- Maximum daily temperature
- The temperature gradient over the course of the day

Statistical modelling of ozone levels (and in particular of models of regression) have been studied many times over. Linear models do not seem to capture all of the complexity of this phenomenon, hence our need to use richer models (cf Chen, Islam and Biswas [7] or Gardner and Dorling [11]). Amongst these models, the MLPs seem to come up with better results that the linear models, even though they often require a much greater effort for just a slight improvement in the prediction. (cf Comrie [8]).

Here we will be showing how such predictions can be further improved thanks to the hybrid model HMM/MLP. In addition, and asides from the improvement in predictions, this modelling offer additional and precious information that allows us to predict peaks in pollution.

For the present study, we have at our disposal meteorological observations and ozone pollution rates for 1994 through 1997. We will be using the 1994-1996 data to estimate our models (data "in sample"), and comparing these different models with the data from 1997 (data "out of sample).

4.1 Comparison between the MLP and the linear model

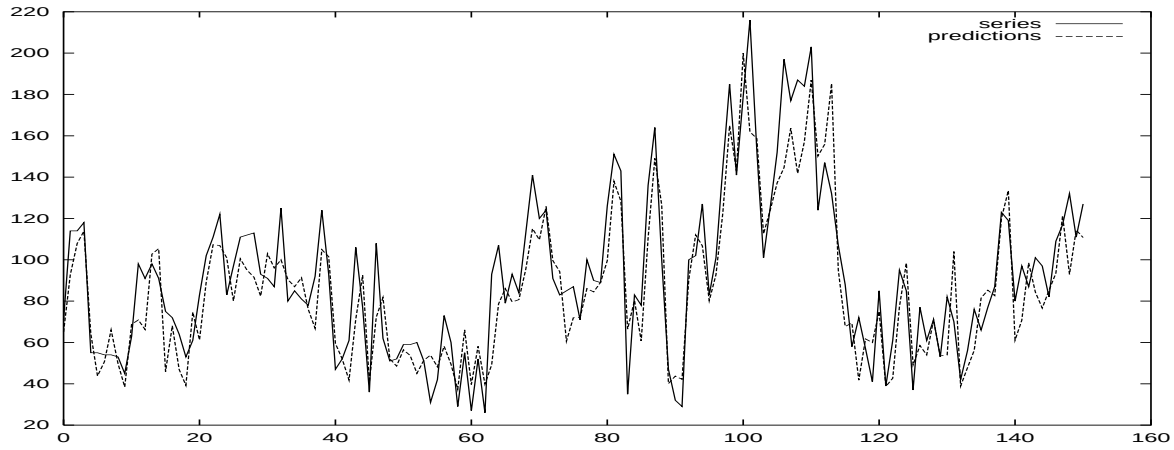
This preliminary study will allow us to detect what it is that the MLP contributes to the simple linear model. The architecture of the MLP model was determined thanks to the SSM method that was described in the first section of the present paper. Here our performance criterion is the square root of the average quadratic error (RMSE). It is expressed in an microgrammes of ozone per cubic meter ($\mu g/m^3$). The table 1 summarises the findings obtained :

Table 1: Comparison MLP, linear model

Années	1994-1996 (in sample)	1997 (out of sample)
RMSE MLP	17.49 $\mu g/m^3$	17.98 $\mu g/m^3$
RMSE LINEAR	20.97 $\mu g/m^3$	19.70 $\mu g/m^3$

Note first of all that the MLP leads to a marked improvement in the performances of the linear model, whether in terms of the "in sample" or "out of sample" data. Although there is relatively little "in sample" data to help us to estimate the model (550), the SSM method makes it possible to avoid over-fitting. This is achieved in an entirely satisfactory manner, given that the RMSE difference between 1994-1996 and 1997 is relatively small. Note in addition that these finding are entirely coherent with preceding studies on atmospheric pollution in Paris (cf Bel et al. [6]).

FIG. 3 – MLP prediction on "out of sample" data



The 3 figure compares the true value of the ozone rate with its MLP prediction based on the "out of sample" series. Note that the prediction for the average values is a particularly good one, but that peaks are generally underestimated. This behaviour is all the more troubling since it is the higher values that the State authorities are more interested in. We are therefore going to use a hybrid HMM/MLP model, hoping that an expert will specialise in the prediction of average and lower values, whilst another will be trying to ascertain the dynamic of the high ones.

4.2 Performance of the hybrid model for an ozone series

Since the linear models are capable of correctly modelling the low values of ozone pollution, we have chosen to use for our hybrid model a linear regression model - as well as a MLP model, whose architecture is the one that is determined on the basis of what we have studied in the preceding section of the present paper. This choice is guided by the desire to maintain the most reasonable number of parameters.

After estimating the parameters, we obtain the following transition matrix for the hidden Markov chain :

$$\hat{A} = \begin{pmatrix} 0.97 & 0.02 \\ 0.03 & 0.98 \end{pmatrix}$$

Note that the diagonal terms, which represent the probability of remaining in the same state, are very close to the maximal probability 1. This means that for long periods of time the model stays in the same state, a sign that it has indeed identified two distinct regimes. The standard deviations for the linear expert σ_0 and for the MLP σ_1 are as follows :

$$\begin{cases} \sigma_0 = 0.11 \\ \sigma_1 = 0.20 \end{cases}$$

This result is intuitively coherent, since as we will see the linear model has specialised in the easier part of the series (i.e., in the average or low values, thus allowing us to come up with good

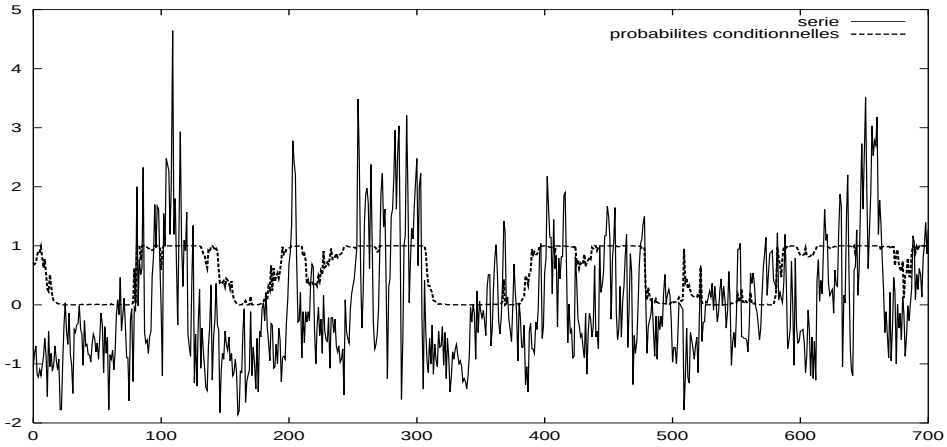
predictions), whereas the MLP is specialised in the more difficult part of the series, i.e., the high values.

Lastly, prediction error results are as follows :

Années	1994-1996	1997
RMSE	$16.51 \mu g/m^3$	$16.75\ \mu g/m^3$

Note first of all a significant improvement in forecasting errors versus the simple MLP. In addition, the hybrid model provides information that is ever richer than is the case with the simple regression model. We can in fact obtain a segmentation of the series that depends on the conditional probability of the two regimes (cf figure 4). Note that the high probabilities of the regime that is associated with the MLP tend to apply to the strong values. In addition, there is never a peak in pollution when this probability is low.

Figure 4: Centred and normalised series and probability of the state that is associated with the MLP



We can also break the model's prediction down into two sections : the linear expert ; and the MLP. This gives the user a lot of flexibility. After all, if s/he is only interested in the strong values, s/he will only take those MLP predictions that are much more relevant into consideration. Figures such as **??** and 6 show the 'scatterplot' nature of the two experts' predictions with respect to the true values for all of the data (in and out of sample)

Note with these figures that the linear expert is better than the MLP for the low values, whereas for the high one its predictions are much less far-reaching than are the true values. The MLP expert is the other way around, inasmuch as it overestimates low values but is much better at estimating the high ones. If we were to use this MLP alone to make predictions, the average quadratic error would be worse than with a single auto-regressive model, but better for the part of the series that we are interested in : pollution peaks as defined by ozone levels.

5 Conclusion

The present paper has introduced two important models that can be used with time series. The first section was devoted to the difficult problem of the model's dimensions. We have offered

Figure 5: Predictions of the linear expert, as a function of the true values

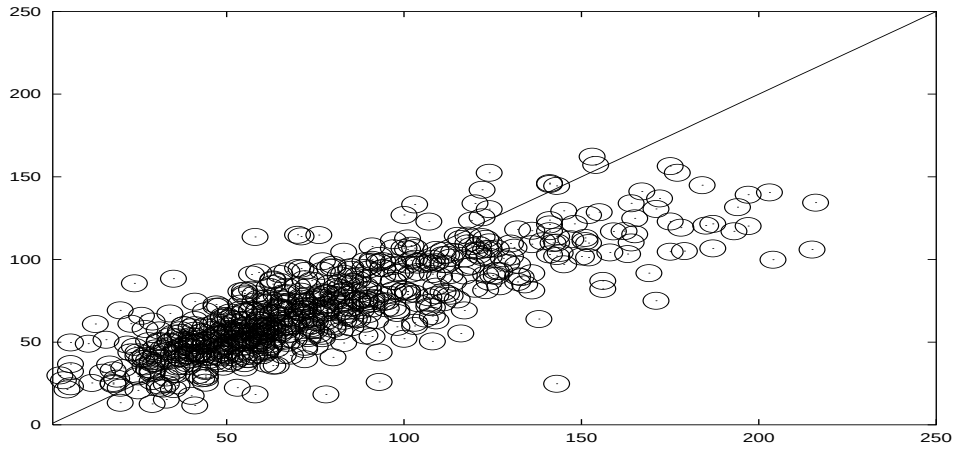
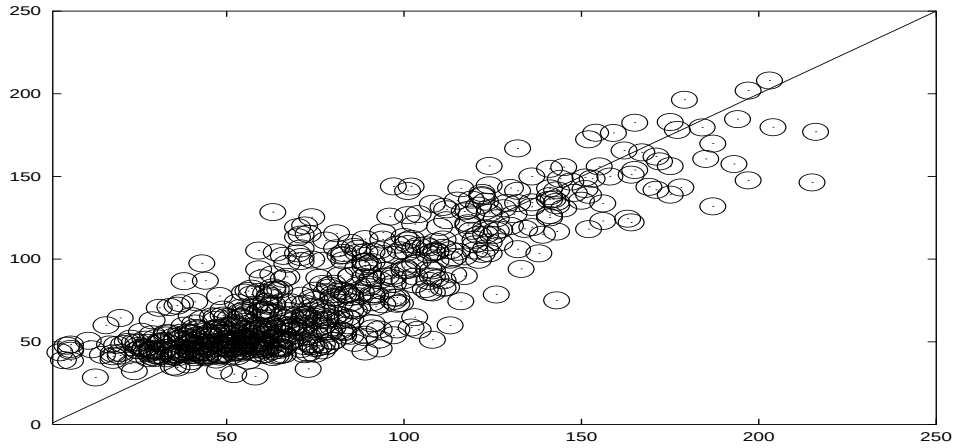


Figure 6: Predictions of the MLP, as a function of the true values



a methodology that is based on the asymptotic properties of the parametric regression models. In our experience, this provides good results as long as there is a sufficient number of observations (at least 500).

In the second section, we have generalised our problem and studied the example of a piecewise stationary time series. The greater complexity of the dynamics underlying such series cannot be captured in a simple regression model. We therefore use, for a particular series, a number of regression functions that are interconnected via a hidden Markov chain. All of these auto-regressive functions make a simultaneous prediction of this series, with the hidden Markov chain's role now being to weight the predictions that these regression models are making by the various regimes' conditional probabilities. By applying this model to pollution data on ozone levels in Paris, we have shown that it can be a relatively promising way of predicting the sort of phenomena that are likely to be associated with regime changes such as pollution peaks.

We should note however that no statistical tools exist yet enabling us to choose the number of regimes for a given series. In fact, if we overestimate the number of regimes, we will lose the

model's identifiability, and traditional statistical tools such as those that were used during the first section of the present paper to determine the MLPs' architecture are no longer theoretically justifiable. Furthermore, this problem constitutes a very active area of research, one that uses highly complex statistical and mathematical tools that go well beyond the purview of the present paper.

Références

[1] H. Akaike. A new look at the statistical model identification. *Transactions on automatic Control*, 19 :716–723, 1974.

[2] L.E. Baum. An inequality and associated maximization technique in statistical estimation for probalistic functions of Markov processes. *Inequalities*, 3 :1–8, 1972.

[3] L.E. Baum and A. Egon. An inequality with applications to statistical estimation for probalistic functions of a Markov process and to a model for ecology. *Bull. Amer. Meteorol. Soc.*, 73 :360–363, 1967.

[4] L.E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite Markov chains. *Annals of Mathematical statistics*, 37 :1559–1563, 1966.

[5] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximisation technique occuring in the statistical estimation of probabilistic functions of Markov processes. *Annals of Mathematical statistics*, 41 :1 :164–171, 1970.

[6] L. Bel et. al. Elément de comparaison de prévisions statistiques des pics d'ozone. *Revue de Statistique appliquée*, 47 :3 :7–25, 1972.

[7] J.L. Chen, S. Islam, and P. Biswas. Nonlinear dynamics of hourly ozone concentrations : nonparametric short-term prediction. *Atmospheric Environment*, 32 :1839–1848, 1998.

[8] A.C. Comrie. Comparing neural network and regression models for ozone forecasting. *Journal of the Air and Waste Management Association*, 47 :653–663, 1997.

[9] M. Cottrell, et al. Neural modeling for time series : a statistical stepwise method for weight elimination. *IEEE Transaction on Neural Networks*, 6 :1355–1364, 1995.

[10] R. Douc, E. Moulines, and T. Rydén. Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regimes. Technical reports 9, University of Lund, 2001.

[11] M.W. Gardner and S.R. Dorling. Statistical surface ozone models : an improved methodology to account for non-lionear behaviour. *Atmospheric environment*, 34 :21–34, 2000.

[12] J. D. Hamilton. A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, 57 :357–384, 1989.

[13] V. Krishnamurthy and T. Rydén. Consistent estimation of linear and non-linear autoregressive models with Markov regime. *Journal of time series analysis*, 19 :3 :291–307, 1998.

[14] F. LeGLand and L. Mevel. Exponential forgetting and geometric ergodicity in Hidden Markov Models. *Mathematics of control, Signal, and Systems, à paraître*, 1999.

[15] J. Rynkiewicz. Modèles hybrides intégrant des réseaux de neurones artificiels à des modèles de chaînes de Markov cachées : applications à la prediction de séries temporelles. PhD thesis, Université de Paris 1, 2000.

[16] J. Rynkiewicz. Estimation of Hybrid HMM/MLP models. In *ESANN'2001*, 2001.

[17] G. Schwarz. estimating the dimension of a model. *The Annals of Statistics*, 6 :2 :461–464, 1978.

[18] H.J. Sussmann. Uniqueness of the weights for minimal feedforward nets with a given input-output Map. *Neural Networks*, 5 :589–593, 1992.

[19] J. Yao. On least square estimation for stable nonlinear AR processes. *The Annals of Institut of Mathematical Statistics*, 52 :316–331, 2000.