Analysing Time Series Structure with Hidden Markov Models

Mehdi Azzouzi Ian T. Nabney
Neural Computing Research Group
Aston University, Birmingham B4 7ET, UK
azzouzim@aston.ac.uk i.t.nabney@aston.ac.uk

Abstract

This paper considers the problem of extracting the relationships between two time series in a non-linear non-stationary environment with Hidden Markov Models (HMMs). We describe an algorithm which is capable of identifying associations between variables. The method is applied both to synthetic data and real data. We show that HMMs are capable of modelling the oil drilling process and that they outperform existing methods.

1 Introduction

A key part of multivariate time series analysis is identifying the lags or delays between different variables. This differs from characterising the order or degree of freedom of a single time series, where the goal is to estimate the intrinsic dimensionality of the data in order to determine the window of past samples needed to map the deterministic component of the data generator. Under the assumption of stationarity, cross-correlation is a powerful tool for estimating linear relationships between variables [5]. The association measures are then used further in linear model identification procedures. They are also often used as the basis for identifying the order of nonlinear models as they are fast to apply. Unfortunately, in many real-world applications such assumptions of linear dependencies and stationarity are not valid.

In this paper, we consider the problem of modelling processes which manifest a sequentially changing behaviour: the properties of the process are usually held pretty steady, except for minor fluctuations, and then, at certain times, change to another set of properties. There is an implicit assumption that the system being monitored is not amenable to standard linear modelling.

Consider, for instance, the oil well drilling process. It exhibits complex time relationships between variables and a highly non-stationary behaviour. Amongst the problems to be addressed, ensuring that the drilling cuttings

0-7803-5060-X/98/$10.00 © 1998 IEEE

are effectively removed from the bore is a topic of considerable practical importance [4] [7]. A drilling fluid called 'mud' carries the cuttings up the hole to the surface. The time it takes for the cuttings to come up to the surface is called the lag for return and is a crucial parameter for modelling the process. This time-dependent parameter, depends not only on the depth of the hole and the pressure of the drilling fluid, but also on the geology of the surrounding rock formation and other parameters.

In this paper, the problem of computing the lag for return is approached using Hidden Markov Models (HMMs). Given two time series x_t and $y_{t+\delta}$, delayed by a lag δ and generated from a non-stationary underlying process which exhibits different regimes, we show how HMMs can be useful to estimate the value of δ. As an illustration, we consider a synthetic problem involving two time series generated by a continuous HMM and show the failure of traditional cross-correlations. We then analyse drilling data and compare our results with cross-correlations and numerical models based on fluid mechanics for computing the lag for return.

2 Hidden Markov Models for modelling Time Series

Suppose that we are monitoring a dynamic system for which observable measurements are available at discrete time intervals. Denote the observable m-dimensional random variable as $\mathcal{V}$ where $v \in \mathcal{R}^m$ is a particular realization of $\mathcal{V}$. For simplicity, assume that $m = 2$ so that at time T we have seen a sequence of such measurements $V = [v_0, \ldots, v_t, \ldots, v_{T-1}, v_T]$ where V represents all the observed data up to time T and $v_t = (x_t, y_t)$ is the observed data at time t.

Let $\mathcal{S}$ be a discrete random variable taking values in the set $\{s_1, \ldots, s_N\}$ and assume that the system at any time t is in one and only one of the N states $s_1, \ldots, s_N$.

Now consider the observable $\mathcal{V}$ in relation to the states: the random variable v_t can be considered to be a probabilistic function of the underlying states, i.e. v_t is an observed measurement from the system but the underlying states are not themselves directly observable. Assuming that the state variable $S(t)$ is a stationary discrete-time first-order Markov process, the resulting model is a doubly stochastic process and is called a first-order Hidden Markov Model (HMM). In general, the parameters of a specific model are referred as $\Lambda = \{A, B, \Pi\}$, where A denotes the state transition matrix, B the observation probability distribution in each state and Π the initial state distribution. For time series modelling, the probability distribution B is often chosen to be a finite mixture of Gaussians, as it can approximate, arbitrarily closely, any finite, continuous density function, provided that enough components are used[1].

[1] Although the continuous density HMMs are applicable to a large number of problems, autoregressive HMMs, where the observation vectors are drawn from a state-dependent

HMMs have been successfully applied in speech recognition [9] [6] [8], cryptography and more recently in other areas such protein classification and sequence alignment [1].

Consider now the following problem: a sequence of observations $V = (x_0, y_0), (x_1, y_1), \ldots, (x_T, y_T)$ is being generated by a sequence of hidden states $S = s_0, s_1, \ldots, s_T$. Unfortunately, we do not see the true sequence V but a delayed version of it: $V^\delta = (x_0, y_{-\delta}), (x_1, y_{1-\delta}), \ldots, (x_T, y_{T-\delta})$. Our task is to estimate the value of δ. Given a two dimensional time series vector $V^\delta = (x_t, y_{t-\delta})_{t=0\ldots T}$, we say that y_t leads x_t by an unknown lag δ. We focus on the problem of computing the value of δ with HMMs.

Under all the above assumptions, we propose the following procedure:

---

for $d = 0$ to $d = D$ **do**

Generate a sequence $V^d = (x_t, y_{t-d})_{t=0\ldots T}$ by delaying the y_t time series by d steps.

Train a HMM, denoted by Λ^d, with the corresponding sequence V^d with the Baum-Welch algorithm, which is the relevant version of the EM algorithm [2].

Compute the log-likelihood $\mathcal{L}^d = \log P(V^d | \Lambda^d)$ of each observation sequence V^d given the model Λ^d using the forward procedure of the Baum-Welch algorithm.

end for

Return $\delta = \max_d \mathcal{L}^d$

This approach is motivated by the fact that the observation sequence V^δ corresponds to a specific sequence of hidden states, which represents the evolution of the system. At each time t, when the data $v_t^d = (x_t, y_{t-d})$ is presented to Λ^d, for a particular value of d, the model adjusts its parameters in order to maximise $\mathcal{L}^d = P(V^d \mid \Lambda^d)$. Intuitively, $\mathcal{L}^d$ will always be less then $\mathcal{L}^\delta$ (except for $d = \delta$).

In order to illustrate our approach, we consider the following problem: we generate a synthetic two dimensional sequence of observations $V^* = (x_t, y_t)_{t=0\ldots T}$ from a two state continuous HMM denoted by $\Lambda^* = \{A^*, B^*, \Pi^*\}$ (in this case $\delta = 0$) and train two state continuous HMMs Λ^d with delayed sequences $V^d = (x_t, y_{t+d})_{t=0\ldots T}, -D \leq d \leq D$. We choose a transition matrix A^* which allows balanced transitions from one state to another, i.e. $a_{ij}^* \approx a_{ji}^*$. The parameters of the Gaussian mixture associated to each state are less important as they do not affect the simulations significantly.

The results are shown in Figure 1. It can be seen that the cross-correlogram is not capable of detecting any relationship between the two time series, even though they are generated by the same HMM. On the other hand, plotting the log-likelihood of each model Λ^d against d shows a significant peak for $d = 0$, which corresponds to Λ^*, i.e. the true model which has been used to generate the time series.

---

autoregressive process, are also interesting for time series analysis [3].

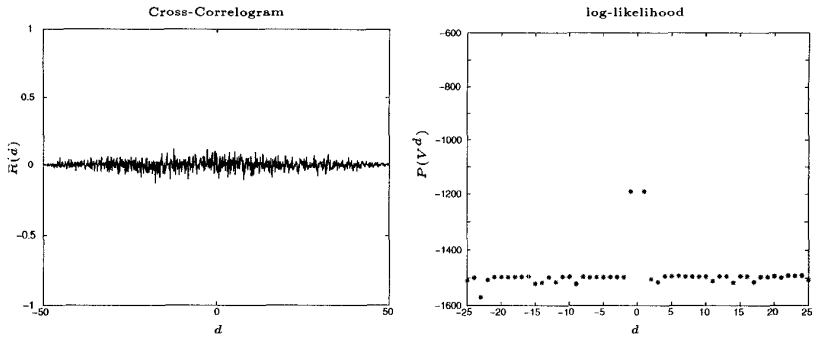

Figure 1: The left hand plot shows the cross-correlogram of the two time series generated by the 2 state continuous HMM Λ^*. The right hand plot shows the log-likelihood $\mathcal{L}^d$ against the lag d.

3 Drilling Data Analysis

One significant aspect of exploration drilling is that of ensuring the drilling debris is effectively removed from the bore; this is known as the 'hole cleaning' problem. At present, no equipment exists to act as a monitor of hole cleaning status. In the case of vertical wells, an adequate velocity in the mud circulation is generally sufficient to guarantee that most debris are brought to the surface. The problem is more complicated when drilling deviated wells[2] since gravity settlement can occur. The gradual build up of low gravity solids increases the torque required to turn the drill string. In extreme cases, the drill pipe may get stuck or even fracture off. The time it takes for the cuttings to come to the surface is called the lag for return and is a crucial parameter in early stuck pipe detection and modelling the drilling process. The current algorithms used on a rig to compute the lag for return are based on fluid mechanics but are believed to have an accuracy in the order of several minutes, mainly because of the poor understanding of downhole conditions.

Recently a new device, capable of detecting fine particulate solids in drilling fluids, has been developed by Thule Rigtech Ltd. [10]. Our aim is to use this device to monitor trends in the volumes of drilled solids in order to obtain a better picture of downhole conditions with regard to drilled solids than has ever been possible before.

As all the data are collected on the surface, if δ represents the lag, then x_t, which is the amount of low gravity solids measured at time t, is effectively the amount of solids that has been generated by the bit at time $t - \delta$. Thus assuming that x_t is related to other drilling parameters $y_{t-\delta}$, it makes sense to use the procedure described in Section 2 in order to estimate the lag for

[2]Whenever possible, wells are drilled vertically, but sometimes, especially offshore, it is necessary to deviate from vertical in order to reach a wide spread of targets from a single platform.

return, as we believe that δ remains relatively constant over a 2 hour time scale. Typically y_t represents one relevant drilling parameter (although we have also considered models with more than one parameter): for instance, the pressure of the circulating fluid inside the pipe or the torque of the pipe, as the drill pipe is subjected to both torsion and tension. The total force applied on the drilling system in order to hold the drill pipe in the rig and the rate of progress[3], are other important parameters.

Figure 2 plots the rate of progress and the amount of low gravity solids for a specific event. At 08h04 the formation changed from soft to hard rock, which can be easily seen in the first plot where the rate of progress decreases suddenly. The second figure plots the amount of solids and shows a significant regime transition at 08h27: the amount of particles decreases as well. Also note the noisy signal for low gravity solids. For this day, the numerical models based on fluid mechanics suggested a value of δ of $43min$, which seems to be wrong as a simple inspection of the graphs gives a value of $23min$.

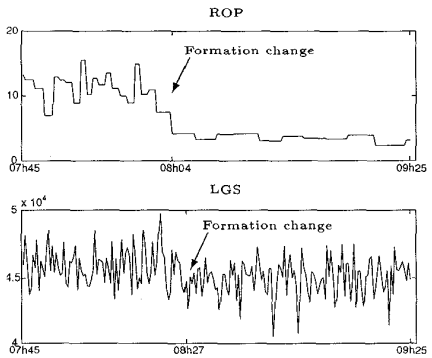

Figure 2: Evolution of the rate of progress and the amount of low gravity solids for a particular event in drilling operations. The formation changes from soft to hard rock at a certain time. The lag of return is visually identifiable in such a situation.

Figure 3 shows our results: 2 state continuous HMMs have been trained with $V^d = (x_t, y_{t-d})_{t=0...T}, 0 \leq d \leq D$ where x_t and y_t denote respectively the amount of low gravity solids and the pressure inside the drill pipe. A significant peak around $23min$ can be seen[4]. Moreover, by applying the Viterbi algorithm to the sequence $V^d = \{x_t, y_{t-d}\}$ with $d = 23min$, we can recover the most probable sequence of hidden states [8]. This algorithm shows that the HMM stays in one state before the event and jumps to the other state precisely when the formation change occurs. Such a clear sequence could not be obtained with other HMM^d trained with a different value for d, confirming the computed value for δ. The cross-correlogram of the two time series shows clearly that this approach does not suggest any value for δ.

[3]The rate of progress or drilling rate simply indicates how fast we are drilling in ft/hr.

[4]Other simulations with different number of states and other drilling parameters did not lead to an improvement of the shape of the plot.

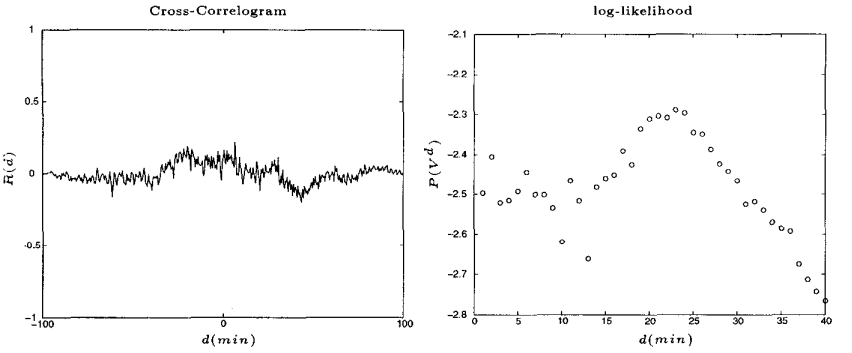

Figure 3: The left hand plot shows the cross-correlogram and the right hand plot shows the average log-likelihood of the observation sequences $V^d = (x_t, y_{t-d})_{t=0...T}, 0 \leq d \leq D$ given the model HMM^d. For each value of d, 10 models, with different seeds for random initialisation, have been trained and the plot represents the average over these 10 models.

We have successfully applied our procedure to other datasets corresponding to different drilling situations. For normal drilling activities, the peak around the estimated value of δ is sharper, suggesting a more precise value. However, we have noticed difficulties associated with the variable selection, as no characterisation of the most relevant drilling parameters is available at this stage. We plan to address this problem by using qualitative relationships given by the numerical models.

4 Conclusion

In this paper, it has been shown how relationships between variables can be identified using Hidden Markov Models. The proposed approach was tested on data from a real-world process and clearly demonstrated its ability to outperform traditional cross-correlations methods. We focussed on the estimation of an important parameter of the drilling process and obtained better results than the numerical models based on fluids mechanics used in the oil industry.

In the future, we want to restate the problem by adopting a novel Mutual Information Estimation approach. We want also to provide an on-line estimate of the lag for return in order to track the amount of solids coming up to the surface for the purposes of early stuck pipe detection.

5 Acknowledgement

The authors would like to thank Mike Affleck of Thule Rigtech Ltd. for providing the data. This work is funded by EPSRC grant GR/L08632.

References

[1] P Baldi, Y Chauvin, T Hunkapiller, and M A McClure. Hidden Markov Models in molecular biology: New algorithms and application. In J D Cowan S J Hanson and C L Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 11–18. Morgan Kaufmann, San Mateo, 1993.

[2] A P Dempster, N M Laird, and D B Rubin. Maximum Likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.*, 39(1):1–38, 1977.

[3] A M Fraser and A Dimitriadis. Time Series Prediction: Forecasting the Future and Understanding the Past. In A S Weigend and N A Gershenfeld, editors, *Forecasting Probability Densities by Using Hidden Markov Models with Mixed States*, pages 264–281. Addison-Wesley, 1994.

[4] G J Guild, I M Wallace, and M J Wassenborg. Hole Cleaning Program for Extended Reach Wells. *Society of Petroleum Engineers*, pages 425–433, 1995. In SPE/IADC Drilling Conference.

[5] M Kendall and J K Ord. *Time Series.* Edward Arnold, 1990.

[6] A B Poritz. Hidden Markov Models: A guided Tour. In *IEEE International Conference on Acoustic Speech Signal Proceedings*, volume 198-8, 1988.

[7] H Rabia. *OilWell Drilling Engineering: Principles and Practice.* Graham & Trotman, 1985.

[8] L R Rabiner. A Tutorial on Hidden Markov Models and Selected Application in Speech Recognition. In *Proceedings of the IEEE*, volume 77-2, pages 257–286, 1989.

[9] L R Rabiner and B H Juang. An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, pages pp. 4–16, January 1986.

[10] Thule Rigtech. Personal communication, 1995.