Improving predictability of time series using maximum entropy methods

G. Chliamovitch\(^1,2\)(a), A. Dupuis\(^1,3\), A. Golub\(^1,3\) and B. Chopard\(^1\)

1 Department of Computer Science, University of Geneva - Geneva 1227, Switzerland
2 Department of Theoretical Physics, University of Geneva - Geneva 1211, Switzerland
3 Olsen Ltd. - Zurich 8008, Switzerland

received 9 December 2014; accepted in final form 25 March 2015
published online 15 April 2015

PACS 02.50.Ey – Stochastic processes
PACS 02.50.Tt – Inference methods
PACS 89.65.Gh – Economics; econophysics, financial markets, business and management

Abstract – We discuss how maximum entropy methods may be applied to the reconstruction of Markov processes underlying empirical time series and compare this approach to usual frequency sampling. It is shown that, in low dimension, there exists a subset of the space of stochastic matrices for which the MaxEnt method is more efficient than sampling, in the sense that shorter historical samples have to be considered to reach the same accuracy. Considering short samples is of particular interest when modelling smoothly non-stationary processes, which provides, under some conditions, a powerful forecasting tool. The method is illustrated for a discretized empirical series of exchange rates.

Copyright © EPLA, 2015

Introduction. – A prominent issue of modern science is that experimental data are collected at ever higher frequencies, which increasingly challenges our ability to shape sensible pictures from such data flows and makes it necessary to develop statistical tools providing a live comprehension of phenomena. We intend here to adress a simple instance of this question: assuming we observe a time series driven by some unknown Markov dynam-ics, what is the most efficient way to rebuild this underlying dynamics? In a second time, we shall re-consider this question when the time series is not generated by a Markov process but only modelled by it, and extend it to the non-stationary case.

The naïve way to tackle this problem is to consider a historical window from which the transition probabilities are deduced by counting the transitions which occurred. We shall instead resort to a class of statistical methods known as maximum entropy (MaxEnt) methods. This approach was introduced to physicists by Jaynes [1–3] as an attempt to free statistical mechanics from physical assumptions going beyond first principles (thereby we mean any ergodic-like hypothesis) and to disentangle its physical and statistical aspects. The purpose of the method is to build generic probability distributions on the basis of partial knowledge, using the criterion that the appropriate distribution is the one which has the largest entropy while satisfying a set of constraints. A classical example is Gibbs’ canonical distribution, which is the maximum entropy distribution corresponding to constrained mean energy. Much more recently, this method has been applied successfully to other kinds of systems, from neuroscience to linguistics [4–8]. Usually the focus is put on the unbiased approximation of distributions too large to be sampled, but MaxEnt distributions have interesting properties even in low dimension [4], so that in this letter we do not try to put particular emphasis on large state spaces.

Stationary processes. – Though many studies have attempted to generalize the MaxEnt principle to dynamical situations, some of them in the context of complex systems [9–11], we focus here on an approach promoted in ref. [12], which showed that the MaxEnt method may be transferred to the realm of Markov stochastic processes. Adaptations are however necessary, since considering the joint entropy of some given subsequence of a process would appear arbitrary; one should instead maximize the entropy rate (or entropy per sign) \( h = - \sum_{ij} \mu_i \ln W_{ij} \), where \( W_{ij} \) denotes the probability of transitioning from state \( i \) to state \( j \), and \( \mu \) the stationary distribution.

Unfortunately, \( \mu \) itself depends on \( W \) which makes the
calculations considerably more involved. A way out [12]
is to maximize instead
\[ \eta := - \sum_{ij} p_i W_{ij} \ln W_{ij}, \]
where \( p \) is some distribution which is kept independent.
One ensures that \( p \) eventually is the stationary distribution by imposing the detailed balance condition \( p_i W_{ij} = p_j W_{ji} \) \( \forall i, j \). In other words, we end up with the following three structural constraints:
\[ \sum_i p_i = 1, \quad \sum_j W_{ij} = 1, \quad p_i W_{ij} = p_j W_{ji}. \]
A simple way one is likely to measure order to get a non-trivial process is the autocorrelation of the time series, and it happens that considering one-step autocorrelation \( A \) makes calculations especially tractable. We therefore constrain the one-step autocorrelation, namely\(^1\)
\[ \sum_{ij} x_i x_j p(i, t; j, t + 1) = \sum_{ij} x_i x_j p_i W_{ij} = A, \]
where \( x_k \) denotes the value assumed by the process in state \( k \).
For constraints given by (2) and (3), the MaxEnt method requires maximizing
\[ \eta + \alpha \sum_i p_i + \sum_j \beta_j W_{ij} + \sum_{ij} \gamma_{ij} (p_i W_{ij} - p_j W_{ji}) + \lambda \sum_{ij} x_i x_j p_i W_{ij}, \]
where \( \alpha, \beta_j, \gamma_{ij} \) and \( \lambda \) are the Lagrange multipliers associated with the respective constraints.

Deriving (4) with respect to \( p_i \) and \( W_{ij} \) and equating to zero results in the system
\[ W_{ij}^{\text{ME}} = \exp(\lambda x_i^2 - \lambda x_j^2), \]
\[ W_{ij}^{\text{ME}} W_{ji}^{\text{ME}} = \exp(\lambda(x_i - x_j)^2). \]
In general the \( W_{ij}^{\text{ME}} \)'s have to be found numerically which may be computationally difficult. In order to focus on the aspects of the method which are related to statistical inference, we shall restrict ourselves to 2- and 3-state cases, for which \( W^{\text{ME}} \) is easily and quickly found. These cases are enough to cover many situations of interest as many models deal with such binarized or ternarized states (for financial models like the one considered below, this would correspond to up, down or flat market moves [13]).

We now focus on the case of two states encoded as ±1. Letting \( A \) denote the autocorrelation of the process, (5) and (6) can be solved to give the MaxEnt transition matrix
\[ W^{\text{ME}} = \begin{pmatrix} \frac{1 + A}{2} & \frac{1 - A}{2} \\ \frac{1 - A}{2} & \frac{1 + A}{2} \end{pmatrix}. \]

We now prove that there exists a subset of the space of 2 × 2 stochastic matrices for which the MaxEnt method is more efficient in estimating \( W \) when we only have short samples at our disposal. We detail the calculations for the coefficient \( W_{1-} \), the other three being similar.

Since the sample autocorrelation of a well-behaved process obeys a central limit theorem [14], we make the assumption that the sample autocorrelation \( A^{(n)} \) measured from a sample of size \( n \) is distributed normally according to \( N(A, n^{-1}) \). We shall not try here to provide an estimate for the accuracy of this assumption since it turns out to be quite good even for short samples, in particular when \( A \) stays small. According to (7), it follows that the error made on the estimation of \( W_{1-} \) using the MaxEnt method is distributed as \( N\left( \frac{1 \pm A}{2} - W_{1-}, (4n)^{-1} \right) \). The absolute value of this error thus obeys a folded normal distribution, which has mean and standard deviation given by [15]
\[ \mu_{-} = \frac{1 \pm A}{2} - W_{1-}, \quad \Phi \text{ denotes the normal cumulative distribution} ^{2}\]
\[ \sigma^{(n)}(\Delta_{ME}^{(n)} W_{-1}) = \sqrt{\frac{1}{4n}} \left( \frac{\left| \Delta_{ME}^{(n)} W_{-1} \right|}{\Phi} \right)^2, \]

where \( \mu_{-} = \frac{1 \pm A}{2} - W_{1-} \) and \( \Phi \) denotes the normal cumulative distribution.

We can similarly provide an estimate of the error committed when estimating \( W_{-1} \) by frequency sampling. It can be shown that the coefficient sampled from a window of size \( n \) is distributed normally according to \( N(W_{-1}, (1 - W_{-1})(1 - W_{-1})^{-1} n^{-1}) \), where \( p_{-} \) denotes the stationary probability of being in state −1, which in the current setting is given by \( p_{-} = \frac{1}{2} - \frac{1}{2} W_{1-} \). Following the same steps as previously, the sampled absolute error on \( W_{-1} \) has mean and deviation:
\[ \langle \left| \Delta_{S}^{(n)} W_{-1} \right| \rangle^{(n)} = \sqrt{\frac{2 W_{-1} (1 - W_{-1})}{\pi n p_{-}}}, \]
\[ \sigma^{(n)}(\Delta_{S}^{(n)} W_{-1}) = \sqrt{\frac{2}{\pi} \frac{W_{-1} (1 - W_{-1})}{n p_{-}}}. \]

We are therefore led to define the accuracy gain
\[ \Delta_{(n)} := \langle \left| \Delta_{S}^{(n)} W_{-1} \right| \rangle^{(n)} - \langle \left| \Delta_{ME}^{(n)} W_{-1} \right| \rangle^{(n)} \]

\(^1\)Defining autocorrelation without centering nor normalization is not unusual in the field of signal processing.

\(^2\)Note that (9) and (11) are displayed for completeness but are not used in the forthcoming developments.
which is positive when the MaxEnt method provides a better estimation of $W_-$ than frequency sampling does for samples of size $n$. Let $n^-_c$ be the value of $n$ above which $\Delta^{(n)}$ becomes negative. Though $n^+_c$ depends on the coefficient, one may wish to define a global $n_c$ for the $W$ matrix considered. While a conservative option is to choose the minimum over all $\Delta^{(n)}_{ij}$ coefficients, we shall rather tolerate a poor estimation of one of the coefficients as long as the corresponding transitions occur scarcely and therefore define $n_c(W)$ in terms of the sum of all $\Delta^{(n)}_{ij}$'s, weighted by the stationary distribution. From our experience, the definition of $n_c(W)$ has only a marginal influence on the forthcoming results.

The quantity $n_c(W)$ is found numerically and plotted in fig. 1 over the space of $2 \times 2$ stochastic matrices parametrized by $W_{-,-}, W_{+,-}$. Note that $n_c$ is large close to the diagonal but decays when one gets remote from the diagonal, which means that a matrix which is “compatible” with structure (7) is better estimated using MaxEnt.

Denoting $M(n)$ the set of matrices such that $n_c(W) \geq n$ and $\mu(n)$ the relative size of $M(n)$ compared to $M(0)$ (the space of all $2 \times 2$ stochastic matrices), then the relevance of the MaxEnt approach for a given state space obviously will depend critically on the function $\mu(n)$. In the 2-state case, one can read from fig. 1 that $M(50)$ is concentrated in a neighbourhood around the diagonal so that $\mu(50) \approx 0.15$, which means that for samples of size smaller than $n = 50$ the MaxEnt estimate is better than the frequency sampling estimate for 15% of all possible processes. One should however note that processes on which one might want to apply the method are unlikely to be scattered randomly over $[0,1] \times [0,1]$, but will rather be processes having a large entropy, that is low predictability. This tends to focus our interest on the central area of $[0,1] \times [0,1]$ and increase the effective $\mu(n)$.

It is not obvious to set up a general argument regarding the way $\mu(n)$ changes when the dimensionality of the state space gets larger. On the one hand, since an efficient frequency sampling in high-dimensional space should require very long samples, one might intuitively expect the MaxEnt to outperform sampling in high dimension, but on the other the MaxEnt procedure (in the case of one constraint put on the process) squeezes the space of independent coefficients onto a one-dimensional submanifold. Our conjecture is therefore that the net result of these competing effects is to penalize the MaxEnt approach for large state spaces. This can nevertheless be alleviated by considering extra constraints (see below), since each extra constraint, as long as it is consistent, adds one more dimension to the MaxEnt submanifold.

To illustrate these points we consider 3-state processes taken randomly in each cumulated quintile of the process entropy rate distribution [16]. Figure 2 shows the effectiveness of our approach by highlighting that processes having large entropy rate are more suited to our approach. In addition, we observe that $\mu(50)$ seems to be below 10%, which seems to confirm our conjecture that MaxEnt performs better for low-dimensional state spaces. For comparison purpose, fig. 2 also displays the success rate of the 2-state process discussed above.

**Supplementary constraints.** Before moving on to non-stationary processes, we should pause to discuss the relevance of extra constraints put on the system, which is somewhat more subtle than might be thought at first. It follows from the MaxEnt statement that putting in extra constraints should bring us closer to the true underlying process but this conception — though correct — is misleading in the case of very short samples, for then the errors made on the observables tend to accumulate. The blue plot in fig. 2 illustrates this nicely by showing the success rate obtained by adding, besides the constraint on autocorrelation, a constraint on the variance of the process. The relative gain is marginal for short samples but tends to become quite significant when the sampling window lengthens. Multiples constraints therefore allow to boost the performance of the MaxEnt method for stationary
processes, as long as one can handle an additional source of error for short samples and extra computational issues. When dealing with non-stationary processes (see below), a major new concern arises since it might be possible that there exists no Markov transition matrix satisfying the constraints deduced from the true process. One should be extremely careful in such a case to introduce constraints which are not mutually exclusive.

**Non-stationary processes.** As long as stationary processes only are considered, the MaxEnt method is actually of limited interest since nothing precludes the use of arbitrarily long samples. Things are very different when the dynamics itself changes over time, for then a quick estimation of dynamical parameters becomes necessary. The crucial point, which follows immediately from our previous results, is as follows: if the coefficients evolve within $M(\tau)$, where $\tau$ is the typical time scale on which the dynamical parameters of the dynamics change, then MaxEnt provides a quicker estimation of the instantaneous dynamics than sampling does.

The change of perspective should be emphasized: while the maximum entropy approach is stationary in essence, we now apply it on non-stationary processes by approximating them locally in time by (hopefully) effective Markov processes.

Figure 3 conveys a qualitative illustration of this approach for a 2-state process generated by the time-varying transition matrix

$$W(t) = \begin{pmatrix}
0.6 + 0.1\sin\left(\frac{2\pi t}{T}\right) & 0.4 - 0.1\sin\left(\frac{2\pi t}{T}\right) \\
0.4 - 0.1\sin\left(\frac{2\pi t}{1.2T}\right) & 0.6 + 0.1\sin\left(\frac{2\pi t}{1.2T}\right)
\end{pmatrix}, \quad (13)
$$

where $T = 500$. Due to the relatively short period of oscillation, considering samples more than a few dozens of time units long would give meaningless over-averaged results. The figure shows that for a sliding window of size $n = 50$, the MaxEnt estimate (shown in blue) provides a better match of the coefficient $W_{-\cdot}(t)$ (red). In particular, it avoids the large deviations shown by the sampling estimate (yellow).

We now turn our attention to an empirical application aiming at quantifying the risk of a financial asset and thus at estimating its return tail distributions. For convenience, we choose to consider the EUR/USD price series over the period from January 1 2009 up to January 1 2012. A series of returns is established by defining $r_t = (p_t - p_{t-1})/p_{t-1}$, where $p_t$ is the price at time $t$ and where the data is picked up every 15 minutes. A 3-state time series $x_t$ is built by discretizing returns according to

$$x_t = \begin{cases}
-1, & \text{if } r_t < -0.01%, \\
+1, & \text{if } r_t > 0.01%, \\
0, & \text{otherwise.} \quad (14)
\end{cases}
$$

Modelling the discretized time series with a 3-states Markov chain, we estimate the transition matrix by the MaxEnt approach. This allows to compute the distribution $s$ steps ahead of current time $t$ as $q_{s+t}(k|0) = \sum w_{i_1,\ldots,i_s} W_{i_0,i_1} \cdots W_{i_{s-1},i_s}$ with $k = i_1 + \cdots + i_s$ and where $i_r \in \{-1,0,+1\}$ is the state at time $t + r$.

In order to assess the quality of the tails of the distribution $q$, we define its first and last ten symmetrized\(^3\) centiles $\pi_k$, $k = 1,\ldots,10$, and compare them to the fraction $\hat{s}_k$ of realizations falling actually into the centile. We then compute the average error $\Delta = \sum_{k=1}^{10} |\pi_k - \hat{s}_k|/\pi_k$. Figure 4 shows, as a function of the sample length $n$, the average error $\Delta$ obtained using the MaxEnt (squares) against the sampling (circles) approach. According to our expectations, we observe that MaxEnt outperforms sampling for sample length shorter than $n = 40$ (i.e. 10 hours). Interestingly, we note that for longer samples, sampling errors seem to grow and depart from MaxEnt errors, which we believe to be an over-average effect as discussed above. The non-informed guess, assuming all transitions equiprobable, is also displayed in fig. 4 (triangles), and is always outperformed by the MaxEnt estimate.

\(^{3}\)That is we aggregate the first and the last centile, and so on.
This highlights that a financial process, exhibiting low correlations, appears to be suited to our procedure.

It is remarkable that the MaxEnt method turns out to be relevant in this context, since contrarily to our previous toy dynamics, it is likely that the underlying dynamics does not change smoothly nor evolves in a subspace of the space of stochastic matrices which is suitable to the MaxEnt procedure.

Conclusions. – Starting from an adaptation of the maximum entropy method suitable to the estimation of Markov processes, we have shown that this method gives more accurate estimates of the transition parameters when short samples only are available for inference. This is however true only when processes to estimate fit to the structure imposed by the MaxEnt procedure, but we argue that many processes of interest do satisfy this property. The full strength of the method appears when dealing with processes changing over time; then, under the same conditions as in the stationary case, MaxEnt provides a quicker on-the-fly estimation of the dynamical parameters. We illustrated the relevance of this approach on empirical data, for which we got a considerably better estimate of the tails of the forecasted distribution of trajectories compared to results obtained using naive sampling methods. The non-stationary case however raises interesting new problems related to the use of multiple constraints. Addressing these issues, as well as the associated computational complexity, in a satisfying way could therefore lead to further improvement of the method.

***

The authors would like to acknowledge funding from the European Union Seventh Framework Programme, under grant agreement 317534 (Sophocles).

REFERENCES

[1] Jaynes E. T., Phys. Rev., 106 (1957) 620.
[2] Jaynes E. T., Phys. Rev., 108 (1957) 171.
[3] Cover T. and Thomas J., Elements of Information Theory (Wiley-Interscience, New York) 2006.
[4] Schneidman E., Still S., Berry M. J. and Bialek W., Phys. Rev. Lett., 91 (2003) 238701.
[5] Schneidman E., Berry M. J., Segev R. and Bialek W., Nature, 440 (2006) 1007.
[6] Stephens G. J. and Bialek W., Phys. Rev. E, 81 (2010) 066119.
[7] Bialek W., Cavagna A., Giardina I., Mora T., Silvestri E., Viale M. and Walczak A., Proc. Natl. Acad. Sci. U.S.A., 109 (2012) 4786.
[8] Stephens G. J., Mora T., Tkacik G. and Bialek W., Phys. Rev. Lett., 110 (2013) 018701.
[9] Fiedor P., Maximum Entropy Production Principle for Stock Returns, arXiv:1408.3728 (2014).
[10] Marre O., El Boustani S., Fregnac Y. and Destexhe A., Phys. Rev. Lett., 102 (2009) 138101.
[11] Cavagna A., Giardina I., Ginelli F., Mora T., Piovani D., Tavarone R. and Walczak A., Phys. Rev. E, 89 (2014) 042707.
[12] Van der Straeten E., Entropy, 11 (2009) 867.
[13] Voit J., The Statistical Mechanics of Financial Markets (Springer, Berlin) 2005.
[14] Brockwell P. J. and Davis R. A., Time Series: Theory and Methods (Springer, Berlin) 1991.
[15] Leone F. C., Nelson L. S. and Nottingham R. B., Technometrics, 3 (1961) 867.
[16] Chлимовічч G., Dupuis A. and Chopard B., Maximum Entropy Rate Reconstruction of Markov Dynamics, in preparation (2015).