Timeseries of Determinisic Dynamic Systems

Sven Laur

November 1, 2004

1 Deterministic dynamic systems

Starting from Renaissance rationality has dominated in the science. W. G. Leibnitz and I. Newton based foundations to mathematical and physical doctrine of determinism—everything in the nature is defined by few deterministic laws, and thus can be explained (computed) automatically from initial conditions. However, the 20th century made quite serious amendments to the core idea of determinism. It turned out that an adequate mathematical model of the process and initial conditions are not sufficient for determining the state of the process in many cases. Small disturbances in initial conditions or changes in model will yield to totally different behaviour. The concept of such ill-posed problems was noted already in 1902 by J. Hadamard. The three body problem (solved by H. Poincaré and K. F. Sundman) gave a second warning: under some initial conditions orbits of three bodies under the gravitational forces can reach a state of chaos—all orbits are non-periodic and do not approach any limiting trajectory. The similar effects were noted for many physical phenomenas, but still there was a lack of mathematical insight why such mysterious fluctuations happen. In a certain sense, the the chaos theory got really popular after publications of Edward Lorenz, who discovered that a simple mathematical model for weather forecasting is unstable and found a simple explanation of the phenomenon.

Nowadays it is well established fact that even simple deterministic nonlinear systems can have truly chaotic behaviour. Hence, if we assume that some phenomenon is governed by (simple) deterministic rules there are three possible behavioural patterns. The system can head to a catastrophe—trajectory in the phase space grows without limit. Recall that phase space[1] is a vector space $\mathbb{R}^n$ that incorporates all parameters which determine the state of system. In case of three body problem, the phase space might consist of three dimensional coordinates of all bodies and their velocities. In case of catastrophe, one of the bodies departures from the others without ever returning.

Systems can also evolve in a stable way—the trajectory in the phase space is either periodic or quasi-periodic (oscillates around stable trajectory). The third alternative is chaos—a trajectory in the phase space jumps seemingly

[1] By some odd reason statisticians name it state space, therefore we use that term afterwards.

randomly around different sub-paths. Clearly, it is quite easy to make long-term predictions, when system is either periodic or quasi-periodic. On the other hand, it is inherently impossible to make meaningful long-term predictions of time series, when the series are generated by deterministic chaotic process.

The most obvious way to describe the evolution of the system is to build a sound well-behaving mathematical model, then extract enough information about initial state, and just compute the necessary output. However, in many cases such approach is impossible: either we cannot build an accurate model or the data is incomplete. Therefore, we should pose a question slightly differently: is it possible to construct an adequate deterministic model for a time serie that is based on the information encoded in the time serie itself. The celebrated Takens theorem provides a partial answer. In the following, we consider practical ramifications of Takens theorem and some possible applications.

2 Basic properties of dynamic systems

The following is written in mathematically loose terms: our intent is to give basic intuition behind the notions and avoid tiresome technical details. For more rigorous treatment, we suggest to consult PhD thesis [Bor98] or any other suitable mathematical monography.

In principle, time series can either be deterministic or probabilistic. The latter means that the observations are only partially determined by the internal state of the system—some information about the internal state is bound to be lost. Formally, the Takens theorem requires that measurements are completely determined by the system state. However, we can extend it to noisy time series, when the noise ratio is small. If the noise ratio is large, too much information about the internal state is lost and we fail to reconstruct the state with sufficient granularity.

Formal description of discrete dynamical system Let us denote the the set of possible states of the system by $\mathcal{X}$. In the following, we assume that $\mathcal{X}$ is a bounded set in a finite dimensional vector space $\mathbb{R}^d$. Informally, we could say that system is somewhat stable and has a compact mathematical description. Let $T : \mathcal{X} \to \mathcal{X}$ be a deterministic rule that uniquely determines the next state given the current state of the system. In other words, we assume that time is discrete or measurements are done with a fixed frequency. Given an initial point x_0, we get a positive orbit of the system

$$\boldsymbol{X} = (x_0, x_1, x_2, \ldots) = (x_0, Tx_0, T^2x_0, \ldots) = (x_0, T(x_0), T(T(x_0)), \ldots).$$

For a formal statistical analysis, we need a probability distribution over the state space $\mathcal{X}$ that is stationary with respect to the evolution rule T. More precisely, if the system state $x_i \in B$ then the next state $x_{i+1} \in T(B)$, hence we require invariance

$$\Pr[x_i \in B] = \Pr[x_{i+1} \in T(B)] = \Pr[x_i \in T(B)]$$

A natural candidate for such probability distribution is average presence time

$$\Pr[x \in B] = \lim_{n\to\infty} \frac{1}{n+1} \sum_{i=0}^{n} \Pr\left[T^i x_0 \in B\right]$$

that counts how long on average orbit stays in the set B.

System evolution and attractors Consider a an orbit $\boldsymbol{X} = (x_0, x_1, x_2 \ldots,)$. As the state space $\mathcal{X}$ is bounded there are many converging sub-sequences of $\boldsymbol{X}$. They form an attractor—a set A such that after sufficiently long evolution all consecutive states of the systems are less than ϵ units away from A. More formally, we get a requirement

$$\forall \epsilon > 0 \quad \exists i_0 : \qquad \forall i > i_0 \quad d(A, x_i) = \inf_{a\in A} \|x_i - a\|_\infty < \epsilon.$$

We should also require that A is stationary $A = T(A)$ and in some sense minimal. Intuitively, the attractor A describes long-term properties of the system.

Correlation integral It is easy to understand that even if the state space is high dimensional, the behaviour of the system might be simplistic. Therefore, we have evaluate the dimensionality of the attractor set (Recall that attractor describes the long-term behaviour of the system). The ordinal dimensionality scale is too coarse and we need a more refined notion. The fractional dimension of A is defined via correlation integral

$$C(r) = \Pr\left[\|X - Y\|_\infty \leq r\right],$$

where X and Y are independently drawn from the stationary distribution defined before. Clearly, the correlation integral depends on r. If $C(r) \sim \text{const} \cdot r^\alpha$ as $r \to 0$, we say that the correlation dimension of A is α. In oder words

$$\text{cdim}(A) = \lim_{r\to 0^+} \frac{\log C(r)}{\log r},$$

if the latter exists. It can be shown that $\text{cdim}(A) \in [0, d]$, where $d = \dim(\mathcal{X})$. The correlation dimension is sometimes referred as intrinsic dimensionality.

Time serie as an observation sequence Unfortunately, we cannot directly observe the evaluation of the system in time. Instead, we make some measurements that somehow characterise the system. Let $f : \mathcal{X} \to \mathbb{R}$ be the read-out function—a measurement scheme. Then for each orbit $\boldsymbol{X}$ there is a corresponding time serie

$$\boldsymbol{Y} = (y_0, y_1, y_2 \ldots) = (f(x_0), f(x_1), f(x_2), \ldots)$$

It might be impossible to restore the internal state x_i from measured quantities $\boldsymbol{Y}$, but the latter is not our primal goal in time series prediction. We need just a reasonable model $g : \mathbb{R}^k \to \mathbb{R}$ that predicts the behaviour of $\boldsymbol{Y}$. The Takens theorem states that under some weak assumptions the latter is doable (at least in theory).

3 Reconstruction. Takens theorem

It is easy to grasp that a single measurement y_i cannot describe the internal state of a complex system. The only reasonable alternative is to add memory, i.e. consider k-tuples $(y_i, y_{i+1}, \ldots, y_{i+k})$ of observations. We denote the corresponding mapping

$$\mathsf{Rec}_k(x) = (f(x), f(Tx), f(T^2x), \ldots, f(T^{k-1}x)).$$

Now for any k and each orbit $\boldsymbol{X}$, we get an extended observation orbit $\mathsf{Rec}_k(\boldsymbol{X})$. Ideally, we would like to distinguish between observation orbits of different systems that is $\mathsf{Rec}_k(\boldsymbol{X_1}) \neq \mathsf{Rec}_k(\boldsymbol{X_2})$ for any $\boldsymbol{X_1} \neq \boldsymbol{X_2}$. Secondly, we should be able to detect "critical" points, where external forces cause a change of orbit. The Takens theorem states that we can detect jumps (discontinuities in evolution) and angle-points (sudden changes of orbit direction without jumps).

Takens Theorem (1981). *Let $\mathcal{X}$ be a bounded set. In the Cartesian product space of C^1-mappings on $\mathcal{X}$ and the space of C^1-functions from $\mathcal{X}$ to $\mathbb{R}$ there exists a open and dense subset U such that if $(T, f) \in U$, then the reconstruction map Rec_k is an embedding, whenever $k > 2 \cdot \dim(\mathcal{X})$. Moreover, the embedding is continuously differentiable and has also continuously differentiable inverse.*

Assumptions

- The read-out function $f : \mathcal{X} \to \mathbb{R}$ is continuously differentiable. In other words, it does not introduce new jumps or angle points.
- The deterministic transformation is also continuously differentiable—usually automatically fulfilled by physical systems.

Implications

- The embedding exists for "almost" all function pairs (T, f). More precisely, if there is no embedding from state space to extended observation space, then there exists a slightly different pair of a deterministic rule and read-out function $(\hat{T}, \hat{f})$ that has a corresponding embedding. We can even characterise the closeness

$$\begin{aligned} \forall \epsilon > 0 \quad \exists \hat{f} \in C^1(\mathcal{X} \to \mathbb{R}) : \quad \forall x \in \mathcal{X} \quad |f(x) - \hat{f}(x)| + |f'(x) - \hat{f}'(x)| < \epsilon, \\ \exists \hat{T} \in C^1(\mathcal{X} \to \mathcal{X}) : \quad \forall x \in \mathcal{X} \quad |T(x) - \hat{T}(x)| + |T'(x) - \hat{T}'(x)| < \epsilon. \end{aligned}$$

  In other words, we can pretend that there is a slight noise in the time serie and slight external disturbance of the observed system. Both assumptions are quite plausible in practice, and we do not have to care about $(f, T) \in U$ or not[2].

[2]However, there is a slight catch here. If system is chaotic then small errors in extended observation space may cause drastic changes in predictions—ideal regressor in not better than random guessing. This is true even $(T, f) \in U$.

- The embedding preserves structural properties of the system. For example, the the image $\mathsf{Rec}_k(A)$ of the system attractor A is also the attractor in the extended observation space. Moreover, the correlation dimensions coincide

$$\mathrm{cdim}(A) = \mathrm{cdim}(\mathsf{Rec}_k(A)).$$

- The observed orbit has roughly the same properties as the orbit of the system—no jumps and angle points are introduced or deleted.

Minimal Regressor Size

Theoretical implications of the Takens theorem are nice, but the most stunning ramification in the context of time series is captured in the following diagram

$$\begin{array}{ccc} \text{Current state} & \xrightarrow[T]{\text{Det. rule}} & \text{Next state} \\ {\scriptstyle \mathsf{Rec}_k}\downarrow\uparrow{\scriptstyle \mathsf{Rec}_k^{-1}} & & {\scriptstyle \mathsf{Rec}_k}\downarrow \\ (y_i, \ldots, y_{i+k}) & \xrightarrow[\mathsf{Rec}_k \circ T \circ \mathsf{Rec}_k^{-1}]{\text{Prediction}} & (y_{i+1}, \ldots, y_{i+k+1}) \end{array}$$

Explicitly stated, if $k > 2\dim(\mathcal{X})$ there exists a precise deterministic rule for predicting the next state of the time serie! Of course, the latter might be missing in our function class $\mathcal{F}$ that we try to fit on the data, but in principle we could find the ideal predictor.

However, the Takens theorem is non-constructive and in first sight does not provide any means to determine a proper size of k. In the following, we show how to estimate $\dim(\mathcal{X})$.

4 Estimators of correlation integral

In order to determine the optimal regressor size, we must somehow estimate the dimensionality of state space $\mathcal{X}$. The second theoretical implication of the Takens theorem

$$k > 2 \cdot \dim(\mathcal{X}) \qquad \Longrightarrow \qquad \mathrm{cdim}(A) = \mathrm{cdim}(\mathsf{Rec}_k(A))$$

provides necessary insight. Intuitively, if we increase the dimension of the extended observation space the correlation dimension $\mathrm{cdim}(\mathsf{Rec}_k(A))$ grows, as the orbit becomes more complex. If the attractor of the system is not redundant $\mathrm{cdim}(A) > dim(\mathcal{X}) - 1$, then we have a trivial inequality

$$k < \dim(\mathcal{X}) \qquad \Longrightarrow \qquad \mathrm{cdim}(A) > \mathrm{cdim}(\mathsf{Rec}_k(A)).$$

Combining both inequalities with *ad hoc* steady growth assumption

$$\mathrm{cdim}(\mathsf{Rec}_1(A)) < \mathrm{cdim}(\mathsf{Rec}_2(A)) < \ldots < \mathrm{cdim}(\mathsf{Rec}_d(A))$$

we get a digestible interpretation of the Takens theorem.

Interpretation of Takens Theorem. *If the correlation dimension of* $\mathsf{Rec}_k(\boldsymbol{Y})$ *and* $\mathsf{Rec}_{k+1}(\boldsymbol{Y})$ *are equal or close enough, then the dimensionality of the state space satisfies* $\dim(\mathcal{X}) \leq k \leq 2\dim(\mathcal{X}) + 1$. *The latter means that the optimal regressor size is between* k *and* $2k + 1$.

Grassberger-Proccacia estimator We want to estimate the correlation dimension of the extended observation orbit $\mathsf{Rec}_k(\boldsymbol{Y}) = (\boldsymbol{z_0}, \ldots, \boldsymbol{z_n}, \ldots)$, where $\boldsymbol{z_i} = (y_i, \ldots, y_{i+k})$. The correlation dimension itself is defined by the correlation integral $C(r) = \Pr\left[\|X - Y\|_\infty \leq r\right]$, where X and Y are random variables distributed by the average presence time. If there are enough orbit points then Grassberger-Proccacia estimator

$$C_n(r) = \frac{2}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \Pr\left[\|\boldsymbol{z_i} - \boldsymbol{z_j}\|_\infty \leq r\right]$$

is a Monte-Carlo approximation of correlation integral $C(r)$. If n is large, vectors $\boldsymbol{z_i}$ are drawn from a distribution that approximates the average presence time distribution. Strictly speaking, values $\boldsymbol{z_i}$ are not independent, but the summation decreases the effect of dependence. Under some weak but weird assumptions, one can even estimate convergence speed of $C_n(r) \underset{n}{\rightarrow} C(r)$. The Grassberger-Proccacia estimator of attractor dimension is a double limit

$$\alpha^{\mathrm{GP}} = \lim_{r \to 0^+} \lim_{n \to \infty} \frac{\log C_n(r)}{\log r}.$$

Of course we cannot compute it, instead we fix n and compute $\alpha_i = C_n(r_i)$. Then we fit a line through $(\log \alpha_1, \log r_1), \ldots, (\log \alpha_m, \log r_m)$ and the slope of the line is the estimate $\widehat{\alpha}^{\mathrm{GP}}$. There three kinds of errors:

- Systematic error. Points $r_1, \ldots, r_m$ are not infinitesimal and thus the fraction $C_n(r_i)/\log r_i$ is only approximately α^{GP}. If we decrease r_i the precision increases.
- Statistical error. We use a finite number of samples in the probability estimate of $\Pr\left[\|X - Y\|_\infty \leq r\right]$. If r is small the neighbourhood of $\boldsymbol{z_i}$ contains too few points and statistical fluctuations dominate. If we decrease r_i the precision decreases.
- Effect of noise. Instead of values $\boldsymbol{z_i}$ we really use $\widehat{\boldsymbol{z}}_{\boldsymbol{i}} = \boldsymbol{z_i} + \boldsymbol{\epsilon_i}$, thus for some pairs $\|\boldsymbol{z_i} - \boldsymbol{z_j}\|_\infty \leq r_i$ though $\|\widehat{\boldsymbol{z}}_{\boldsymbol{i}} - \widehat{\boldsymbol{z}}_{\boldsymbol{j}}\|_\infty > r$ and vice versa. The latter introduces a bias to the estimate: $C_n(r_i)$ will be slightly smaller than without noise. It is more probable to increase the distance between points than decrease. If we decrease r_i, the error increases.

Therefore, we must do a fair tradeoff between three error components. There are also two principal chooses for fitting the line: least square or least median regression. The latter is more stable against outliers and should provide more conservative estimate.

Generally, Grassberger-Proccacia estimator is quite stable, but it has a serious downside—quadratic computational complexity. Hence, the computations quickly become intractable for large values of n. One possible caveat is to use sub-sampling.

Takens estimator Consider the case when $C(r) = c_0 \cdot r^\alpha$, whenever $r \leq r_0$. Then the conditional probability

$$\Pr\left[\|X - Y\|_\infty \leq r | r \leq r_0\right] = \frac{\Pr\left[\|X - Y\|_\infty \leq r\right]}{\Pr\left[\|X - Y\|_\infty \leq r_0\right]} = \left(\frac{r}{r_0}\right)^\alpha$$

Simple substitution shows that $S = \log \frac{\|X-Y\|_\infty}{r_0}$ has an exponential distribution $\Pr\left[S \leq s\right] = 1 - \exp(-\alpha s)$. Thus given an upper threshold r_0 we can estimate the dimension α by Maximum Likelihood estimate

$$\widehat{\alpha}^{\mathrm{ML}} = \left(\frac{1}{m}\sum_{i=1}^{m} S_i\right)^{-1}, \qquad S_i = \frac{\|\boldsymbol{z_i} - \boldsymbol{z_j}\|}{r_0} \leq 1$$

The formula is correct when $\boldsymbol{z_i}$ and $\boldsymbol{z_j}$ are independent. Since $\boldsymbol{z_i}$ and $\boldsymbol{z_j}$ are dependent, Takens suggested modified formula

$$\widehat{\alpha}^{\mathrm{T}} = -\left(\frac{2}{n(n-1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{n} \frac{\|\boldsymbol{z_i} - \boldsymbol{z_j}\|_\infty}{r_0}\right)^{-1}$$

Generally, the condition $C(r) = c_0 \cdot r^\alpha$ is not satisfied. However for sufficiently small r_0, $C(r) \approx c_0 \cdot r^\alpha$ and the estimate has a small enough bias.

The Takens estimator is computationally more efficient, but still has quadratic complexity. On the other hand, the estimate is computationally unstable. If $\boldsymbol{z_i} \approx \boldsymbol{z_j}$ the small variations of $\boldsymbol{z_j}$ have large impact on $\widehat{\alpha}^{\mathrm{T}}$.

Chord estimator If we take only two values r_0 and r_1 in the Grassberger-Proccacia estimator, the line-fitting becomes trivial

$$\widehat{\alpha}^{\mathrm{C}} = \frac{\log C_n(r_0) - \log C_n(r_1)}{\log r_0 - \log r_1}$$

But it seems fishy—depending on the choice of r_0 and r_1 the estimate $\widehat{\alpha}^{\mathrm{C}}$ might vary a lot. The trick here is to choose optimal values of r_0 and r_1 that minimise the expected variance of $\widehat{\alpha}^{\mathrm{T}}$. It has been proven that with properly chosen r_0 and r_1 the expected relative error is only 1.25 times worse than for the Takens estimator. On the other hand, it is computationally more stable.

Ellner estimator In the Takens estimator, we ignored pairs that are to far apart. The Ellner estimator ignores also pairs that are too close and adds a corresponding correction term into the formula of $\widehat{\alpha}^{\mathrm{T}}$. Intuitively, the number of very close orbit points is small and statistical fluctuations do not cancel out. Thus pruning improves stability of the estimate. From theoretical point, the Ellner estimator is much more appealing than the Takens estimate.

References

[Bor98] Svetlana Alfredovna Borovkova. *Estimation and prediction for nonlinear time series.* PhD thesis, University of Groningen, 1998. Freely available `http://www.ub.rug.nl/eldoc/dis/science/s.a.borovkova/`