Inference, Models and Simulation for Complex Systems CSCI 7000-001
Lecture 7 15 September 2011
Prof. Aaron Clauset

1 Random walks and data

Suppose you have some time-series data $x_1, x_2, x_3, \ldots, x_T$ and you want to model it using a random walk. As we saw in Lecture 5, however, there are a very large number of different kinds of random walks, e.g., bounded, semi-bounded or unbounded, biased or unbiased, those with size-dependent fluctuations, multiplicative or additive, combinations of these, and many more. How can you make sense of the underlying structure of your observed data in this framework?

1.1 Serial correlation

Recall that the simplest kind of random walk has the form

$$x_t = x_{t-1} + \lambda \ , \tag{1}$$

where λ is a random variable drawn iid from a distribution $\Pr(\lambda)$. Thus, a good first step in characterizing the structure of a random walk is to test whether the data exhibit *serial correlation*. That is, can you reject the hypothesis that sequential events are iid from some unknown but stationary distribution $\Pr(x)$?

You can get a quick idea of whether this hypothesis is reasonable by making a scatterplot of x_t against a "lagged" variable x_{t-k}, where typically $k = 1$. For instance, Figure 1 shows the results of doing this for data drawn iid from a normal distribution (upper) and for differences drawn iid from a normal distribution, i.e., for a process with serial correlation (lower). The middle column of figures shows the scatter plot of x_{t+1} vs. x_t. In the case of no-serial correlation, the scatter shows no structure. A formal hypothesis test of whether x_{t+1} correlates with x_t in a statistically significant way would lead us to reject the serial correlation hypothesis.

In contrast, data generated by a random walk (lower) shows a very strong diagonal structure on the scatterplot, and a formal hypothesis test would verify that this structure is unlikely to be accidental. This example is a particularly strong illustration of serial correlation. Processes that exhibit weaker serial correlation structures would exhibit a more diffuse diagonal cloud on such a scatterplot. That is, the stronger the correlation, the more diagonal the shape; the weaker the correlation, the more spherical (or square) the shape.

A statistic called the Hurst exponent H (after Harold E. Hurst, 1880-1978) is sometimes used to characterize the strength of serial correlation. $0 \leq H < 1/2$ indicates negative autocorrelation,

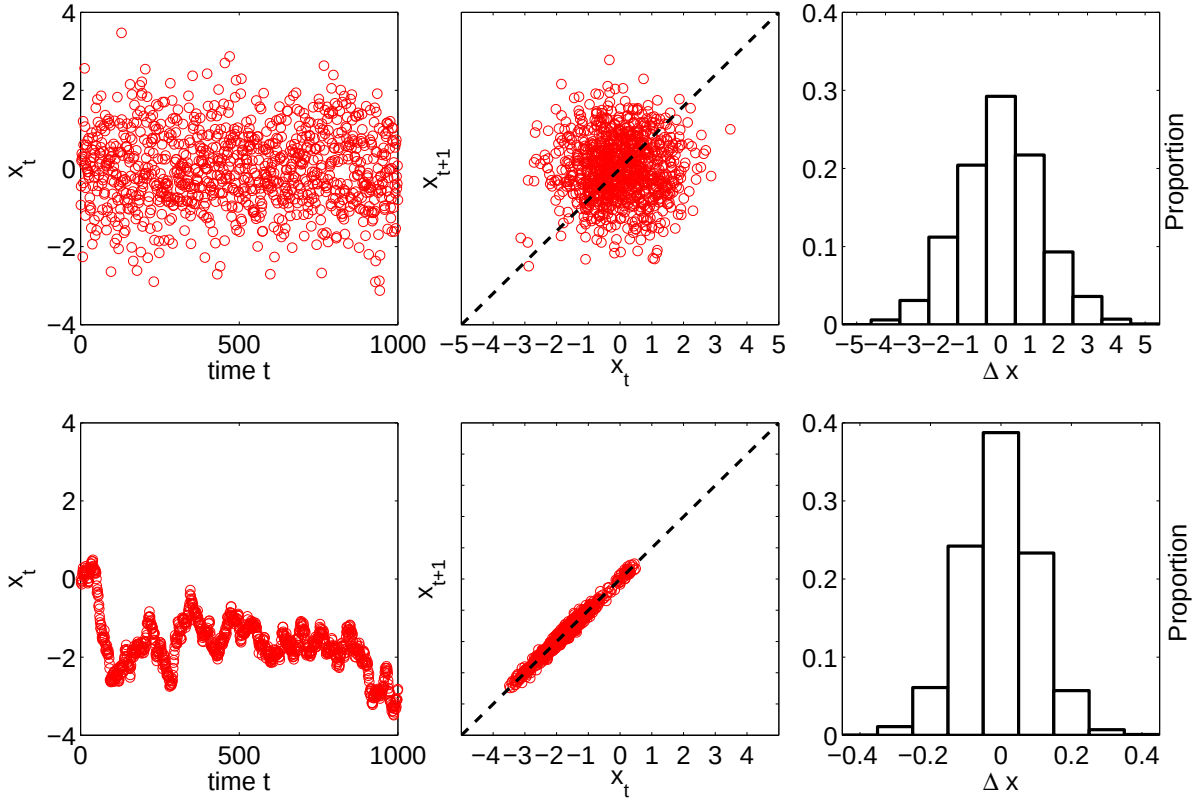

Figure 1: (upper) 1000 points drawn iid from a normal distribution with mean zero and unit variance; a scatter plot of x_{t+1} vs. x_t; and the distribution of differences $\Pr(\Delta x = x_{t+1} - x_t)$. (lower) The same for 1000 points drawn from a random walk with a normal diffusion kernel.

i.e., an increase is likely to be followed by a decrease; $1/2 < H \leq 1$ indicates the reverse (positive autocorrelation), i.e., increases tend to appear together (and similarly for decreases). The special case of $H = 1/2$ appears when increases are just as likely to follow decreases (and vice versa), which is precisely the behavior of an unbiased symmetric random walk. These facts are derived analytically from mathematical models, and so there are natural questions about accurately estimating H from empirical data. It turns out, it can be difficult to pin down the uncertainty in an empirical estimate of H, which thus makes it an ambiguous measure of the long-time behavior of the system.

1.2 Characterizing $\Pr(\lambda)$

A more powerful diagnostic tool is to measure and characterize the distribution of fluctuations itself $\Pr(\lambda)$. That is, tabulate the empirical sequential differences $\Delta x_i = x_i - x_{i-1}$ and examine *their* distribution. Mathematically, this is equivalent to measuring the vertical residuals from the $x_{t+1} = x_t$ diagonal line on the serial correlation plot and considering *their* distribution.

For the iid and random walk processes used in Section 1.1, the third column of Figure 1 shows the empirical fluctuation distributions (a.k.a. the diffusion kernel). Notably, they're both normally distributed. Why is this? The iid process exhibits it because the difference of two iid normally distributed random variables is itself a normal distribution but with greater variance, while the random walk does because that's its underlying structure. This means that, in general, $\Pr(\lambda)$ cannot help us with the question of serial correlation, since the distribution of differences for iid variables is often itself a simple distribution.

In the simplest cases, the empirical fluctuation distribution is where all the independence in the system lies. Thus, if we can show that the empirical distribution of differences can plausibly be explained by some simple distributional form, e.g., a normal distribution with mean and variance[1] estimated from our empirical differences, then we're done and we could simulate the entire underlying process by simply drawing new fluctuations from the fitted model. Even if we cannot pin down the general form of $\Pr(\lambda)$, which will be common, we can still estimate the overall bias in the system by computing $\langle\hat{\lambda}\rangle$. Even if we have missed a lot of interesting underlying structure, the Central Limit Theorem is general enough that our estimate of the bias will be fairly trustworthy.[2]

Real-world systems rarely exhibit such nice behavior and the empirical fluctuation distribution will often exhibit non-trivial structure. In this case, we must look deeper for independence.

1.3 Size-dependent fluctuations

Recall the example from Lecture 4 that exhibited size-dependent fluctuations, in which $\Pr(\lambda \,|\, x) = N[\mu(x), \sigma^2]$ and $\mu(x)$ was a symmetric cubic function. Dynamically, this system exhibit roughly bi-stable behavior, soft upper and lower boundaries and an equilibrium distribution. Given data time-series drawn from such a generative process, we can recover the underlying structure by empirically estimating the structure of the size-dependent diffusion kernel. We do this by tabulating the change in x as a function of x, that is, we create the table

[1] Or, some other parametric distribution, like the Laplace (exponential tails) or double-Pareto (power-law tails).

[2] The CLT holds whenever the variance in the diffusion kernel is not unbounded and the observations not too strongly correlated. The larger the variance, however, the slower the convergence of the averaging operation (think of the expression for "standard error" in the uncertainty of an estimated average value). And, the less independent the fluctuations, the more we underestimate the true variance in the generative process.

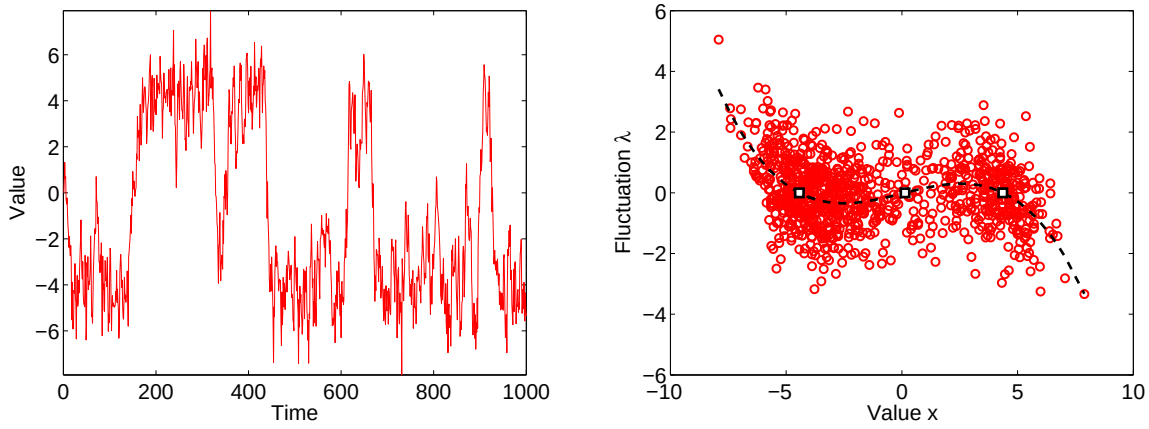

Figure 2: (a) A time series of length $T = 1000$ drawn from the size-dependent random walk described in Lecture 4, where $\Pr(\lambda \,|\, x) = N[\mu(x), \sigma^2]$ and $\mu(x)$ was a symmetric cubic function. (b) A scatter plot showing the empirical size-dependent diffusion kernel along with the maximum likelihood 3rd-order polynomial fit to the data and the estimated locations of the two attractors and the repeller.

x_t	$x_{t+1} - x_t$
x_1	$x_2 - x_1$
x_2	$x_3 - x_2$
$\vdots$	$\vdots$
x_{T-1}	$x_T - x_{T-1}$

These x,y pairs can then be visualized as a scatterplot and fitted with a generative model to determine their underlying structure. If these can be shown to be plausibly generated by the fitted model, then we're done and again we can simulate the process by drawing new fluctuations independently from the size-dependent diffusion kernel.

Figure 2a shows a simulated time series of length $T = 1000$ drawn from the size-dependent model used in Lecture 4; Fig. 2b shows the empirical size-dependent fluctuations as a scatter plot, along with the maximum likelihood 3rd-order polynomial fitted to the data, which nicely recovers the underlying structure.

1.4 Boundaries

Boundaries can be easy to identify if you have enough data, as they show up as clear limits on the range of the variable. Identifying what *kind* of boundary can be more difficult. This is because

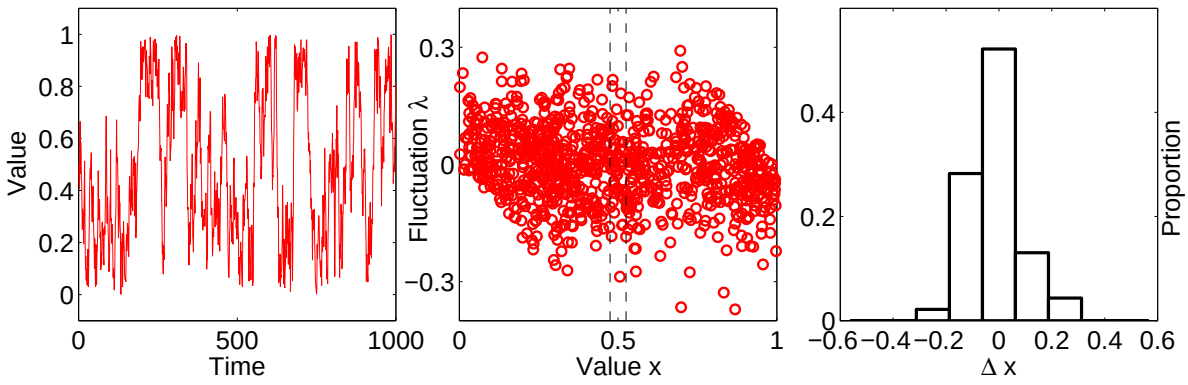

Figure 3: (a) A time series of length $T = 1000$ drawn from the normal random walk on the interval $[0, 1]$. (b) Scatterplot showing the empirical size-dependent fluctuations; dashed lines are at 0.475 and 0.525, a region "far" from both boundaries. (c) The empirical fluctuation distribution for walkers $0.475 < x_t < 0.525$ [that is, the distribution of the scatter between the vertical lines in (b)].

there are several kinds of behaviors that can appear in a random walk as it approaches a boundary. For a single trajectory, the typical behavior is that any steps that would cause the walker to cross the boundary are simply disallowed. This effectively truncates the fluctuation distribution when the walker is near a boundary, e.g., $\Pr(\lambda < x_{\min} - x_t)$ for a lower boundary and $\Pr(\lambda > x_{\max} - x_t)$ for an upper boundary, thus increasing the likelihood of taking a step away from the boundary.

But note that this simply induces size-dependencies in the fluctuation distribution near the boundaries, i.e., $x \approx x_{\min}$ or $x \approx x_{\max}$. This means that boundaries of this kind can be detected by estimating the size-dependent fluctuation function, as in Section 1.3. Visually, they appear as a "forbidden" or empty region in the lower-left or upper-right corners of the figure, because these are the places where a fluctuation would cause the walker to cross the boundary. A crude but simple way to eliminate the impact of the boundaries and recover the underlying diffusion kernel $\Pr(\lambda)$ is to only consider steps taken by walkers "far" from either boundary, where the biasing effects will be small.

Figure 3 illustrates this idea using time series data drawn from a normal random walker on the unit interval. In this case, there is nowhere that is "far" from both upper and lower boundaries and so the region we focus on is the small set of fluctuations around $x_t = 0.5$. Fig. 3c shows the corresponding distribution of step sizes for those steps taken in the target region, which has $\hat{\mu} = -0.010 \pm 0.015$ and $\hat{\sigma} = 0.098$. The true values in the underlying process were $\mu = 0.000$ and $\sigma = 0.100$. Not bad.

1.5 Multiplicative random walks

Often, it's immediately clear by inspection that some empirical time series data are more likely to be described by a multiplicative random walk than an additive one. For instance, if the distribution of values is very heavy-tailed, or the standard deviation is larger than the median value, or the range of values covers several powers of 10, a multiplicative random walk is likely to be a better model.

In this case, the steps or changes are proportional and are recovered from the empirical data by taking the ratio of sequential observations rather than their difference. All of the tools described above can easily be applied to multiplicative random walks, with the main difference being the way the results are interpreted and a few mathematical details.

2 Related topics

There are many other approaches—indeed even entire textbooks devoted—to analyzing time series data. We won't cover them in detail, but here are brief descriptions of some of the more common alternative techniques.

2.1 Chaotic time series and delay coordinate embedding techniques

These are techniques typically applied deterministic systems (in contrast to the stochastic systems we considered above). A *chaotic* system is one in which nearby trajectories (e.g., two systems with initial conditions x_0 and $x_0 + \epsilon$ for $\epsilon \to 0$) through the system's *state space* diverge exponentially, that is, they have a positive Lyapunov exponent. Examples of such systems include special kinds of coupled differential equations, such as the Lorenz equations, and the logistic map. A number of physical systems are known to exhibit chaotic dynamics.

Chaotic systems exhibit what are called "strange attractors," which are a special kind of pattern in the trajectories through the state space. Non-chaotic systems can exhibit a range of less strange behavior, such as limit cycles (loops in the state space) and basins of attractions (regions of the state space that are separated dynamically from each other).

Given a time series $x_1, x_2, x_3, \ldots, x_T$, a state space equivalent to the generating process can be recovered using a technique called delay coordinate embedding. The basic idea is to first estimate the delay coordinate τ that correctly samples the dynamics of the system and then sample the data at intervals of τ to estimate the embedding dimension d_E. There are several ways of doing each of these. The average mutual information and the false-nearest neighbors approaches are popular ways to the two steps, respectively. Figure 4 shows an example of the results of such an analysis,

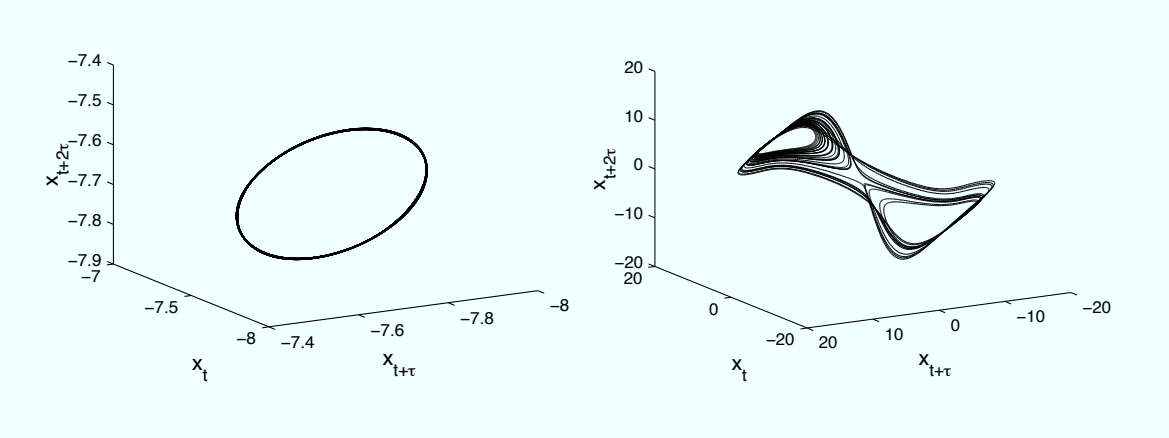

Figure 4: Reconstructed attractors, using the delay coordinate embedding technique, for the (a) stable (limit cycle) and (b) chaotic regimes of the Lorenz equations.

run on data generated by the Lorenz equations.

For more information, see S. Strogatz, *Nonlinear Dynamics and Chaos.* Addison-Wesley (1994) and H.D.I. Abarbanel, *Analysis of Observed Chaotic Data.* Springer Verlag (1996).

2.2 Auto- and cross-correlation functions

Recall from Section 1.1 that sequential-correlation is a hallmark of random walks. A more general idea is that of auto-correlation, in which the value at x_t is most correlated with the value at some $x_{t-\tau}$ for $\tau > 0$. If $\tau > 1$, a simple plot of the sequential correlation may not capture the true self-correlative behavior. The autocorrelation function (ACF) can pick out correlations at any length. Most compactly, the ACF can be computed using results from spectral theory, using a Fourier transform. Less elegant solutions also work for small to moderate sized data sets. In general, the ACF illustrates the correlation between the observed time series $f(t)$ with itself, but lagged by a variable amount, $f(t - \tau)$ for all τ.

The cross-correlation function (CCF) works the same way as the ACF, but instead of computing the correlation between a time series and itself at various lags, the CCF computes the correlation between two time series $f(t)$ and $g(t)$ at various lags τ. In both of these techniques, there remains the tricky problem of identifying whether the observed ACF or CCF could have been generated by chance. One way of estimating the underlying variability in the system, and thereby estimate whether the observed ACF or CCF is meaningful, is to use the bootstrap.

2.3 Autoregressive models

Autoregressive (AR) models are a very general family of generative processes, of which the simple random walks we discussed here and in Lecture 5 are special cases. A kth-order AR model is given by

$$x_t = \lambda + \sum_{i=1}^{k} \phi_i x_{t-i} \ , \tag{2}$$

where λ is an iid random variable drawn from our diffusion kernel and ϕ_i is a weight that specifies how much influence the x_{t-i} variable has on the x_t value. When $k = 1$ and $\phi_1 = 1$, we recover a simple additive random walk. Varying ϕ_i between 0 and 1 allows us to move between a fully iid process with no serial correlation $\phi_i = 0$ and full serial correlation $\phi_i = 1$.

AR models can be further generalized to include non-stationary behavior (see Section 3.1 below) such as moving averages. In general, the more complicated the model and the more free parameters it contains, the more difficult it is to fit the model to data and the less statistical power it has over alternative models. Many of the techniques we discussed in Lecture 4, however, can be used to test the plausibility of candidate such models or to choose the model complexity k.

AR models can also be generalized to vector autoregressive (VAR) models, which allow a vector of time series variables $\vec{x}_t$ rather than focusing only on a single time series variable x_t. These models are typically used in econometric contexts, where we can measure a large number of attributes of a system over time, e.g., the value of a set of n goods or stocks. The most general version of this model allows for arbitrary covariance structure in the fluctuation distribution; simpler versions assume conditionally independent fluctuations or other structure on the covariance matrix.

3 Complications

3.1 Non-stationarity

For a process to be *stationary*, the generative dynamics must be independent of the time variable t. For instance, in the example of size-dependent fluctuations in Section 1.3, the generative process exhibits dependencies on the current state, but the diffusion kernel itself $\Pr(\lambda \,|\, x)$ is the same at every time t. If we introduced some slight dependence, for example, by letting the constant a_0 in the cubic function increase slowly with time t, then the process would be non-stationary.

In general non-stationarity has to be modeled and controlled for in order to estimate the underlying structure of the time-series. The bad news is that the number of ways a process can be non-stationary is large, and it can be difficult to pin down. The most conventional approaches to dealing

with non-stationarity are to assume some kind of moving average, estimate it, and then derive the residual fluctuations relative to that moving average. These residuals can then be analyzed as a stationary process.

3.2 Statistical Forecasting

All empirical time series are finite since it's difficult to know the future behavior of a system without observing it first. In many applications, however, we might like to make forecasts about the future behavior of the system. Statistical forecasting is, at its most basic, the extrapolation of past dynamical patterns into the future, possibly with the aid of mathematical models. For instance, think about weather forecasting.

The fundamental difficulties, of course, are in accurately characterizing the generative process of the time series data and in accurately estimating any unknown parameters of that process. For instance, consider the difficulty of making accurate statistical forecasts in a system exhibiting deterministic chaos. Here, in addition to the problem of specifying and estimating a good model, e.g., using delay-coordinate embedding from Section 2.1, there is the problem of measuring the initial conditions accurately. Slight errors there imply exponentially worse forecasts the further out in time we try to forecast. On the other hand, if a system exhibits only short-term chaotic dynamics but long-term non-stationarity, as is believed to be the case with the global climate system, it can be possible to make statistical forecasts about the global behavior of the system even if it is not possible to make accurate forecasts about the short-term local behavior of the system. In other contexts, such as for social processes like the occurrence of a terrorist attack or the outbreak of war or biological processes like the evolution of the influenza virus, it's an open question as to how forecastable events truly are.