Tutorial 6

Statistical analysis of time series

Peter Košovan[*], Marcello Sega

May 4, 2011

ICP, University of Stuttgart

We return back to analysis of simulation data, error estimation and identifying the equilibration period. We re-use some simulation programs from previous semester and focus mainly on data processing and obtaining reliable and unbiased estimates of the observables. We compare several methods of estimating correlation times in terms of efficiency and accuracy as well as computational cost.

1 Correlation functions and correlation times

In the last tutorial of the previous semester, we have been estimating correlation times of observables generated as a time series. Our estimates were done in a “quick and dirty” way which relies on assumptions about the properties of the correlation functions. In the current tutorial we revisit the same problem again. We will compute the correlation functions and correlation times explicitly, and compare different methods. This should provide more insight into reliability of the estimation methods.

Autocorrelation function of an observable $\mathcal{O}$ can be defined as

$$C(\tau) = \langle \, ((\mathcal{O}(t) - \langle \mathcal{O} \rangle \,) \cdot (\mathcal{O}(t+\tau) - \langle \mathcal{O} \rangle) \, \rangle \tag{1}$$

where $\langle \cdot \rangle$ refers to an ensemble average. Considering that $\langle \mathcal{O}(t) - \langle \mathcal{O} \rangle \rangle = 0$ it simplifies to

$$C(\tau) = \langle \mathcal{O}(t) \cdot \mathcal{O}(t+\tau) \rangle - \langle \mathcal{O} \rangle^2 \; . \tag{2}$$

It can also be written as the integral

$$C(\tau) = \lim_{t_{\max}\to\infty} \frac{1}{t_{\max}} \int\limits_0^{t_{\max}} \mathrm{d}t \, (\mathcal{O}(t) - \langle \mathcal{O} \rangle \,) \cdot (\mathcal{O}(t+\tau) - \langle \mathcal{O} \rangle) \, . \tag{3}$$

[*] kosovan@icp.uni-stuttgart.de

Note that $C(\tau)$ has the form of convolution of $\mathcal{O}(t)$ with itself. In a system in equilibrium (no time-drift), $C(\tau)$ is translationally invariant in time, *i. e.* it only depneds on the time difference (lag time), τ, but not on the absolute time t. It may be useful to normalize the autocorrelation function such as

$$\Gamma(\tau) = C(\tau)/C(0)\,.$$

The time-correlation function measures how quickly the system loses memory of its previous state. There are two important limits: $\Gamma(0) = 1$ and $\Gamma(\infty) = 0$. Asymptotically for large τ, correlations often decay exponentially

$$\Gamma(\tau) \sim \exp\left(-\frac{\tau}{\tau_{\mathrm{exp}}}\right)\,, \tag{4}$$

where τ_{exp} is the exponential correlation time. In a more general case, $\Gamma(\tau)$ can be composed of several exponential functions

$$\Gamma(\tau) = \sum_{i=0}^{i_{\max}} A_i \exp\left(-\frac{\tau}{\tau_i}\right)\,,$$

or it does not have to be exponential at all. The integrated autocorrelation time is defined as

$$\tau_{\mathrm{int}} = \frac{1}{2} + \int\limits_0^\infty \Gamma(\tau)\mathrm{d}\tau\,, \tag{5}$$

which in the case of a multi-exponential decay of $\Gamma(\tau)$ is a weighted average of the individual correlation times

$$\tau_{\mathrm{int}} = \frac{1}{2} + \sum_{i=0}^{i_{\max}} A_i \tau_i\,. \tag{6}$$

1.1 Computing correlation functions from a simulation trajectory

We will distinguish the ensemble average of an observable, $\langle \mathcal{O} \rangle$, which is a property of the system under study, and its estimator, $\overline{\mathcal{O}}$, which is computed by stochastically sampling the ensemble, *e. g.* from a given simulation trajectory using N samples which may be correlated. If the system is ergodic, then $\overline{\mathcal{O}} \overset{N\to\infty}{\longrightarrow} \langle \mathcal{O} \rangle$ but $\overline{\mathcal{O}} \neq \langle \mathcal{O} \rangle$ for a finite N. Similarly, symbol $\Gamma(\tau)$ will denote the autocorrelation function which is a property of the process which generates the time series, while $\overline{\Gamma}(\tau)$ is its estimator based on a finite sample of the time series.

We assume that the observable $\mathcal{O}$ is measured after fixed time intervals Δt. To simplify the notation, will furhter drop Δt and denote the value measured at time $i\Delta t$ as $\mathcal{O}(i)$. Similarly, value of $\Gamma(\tau = i\Delta t)$ will be denoted as $\Gamma(i)$.

The simplest algorithm to compute the estimator $\overline{\Gamma}(\tau)$ is to correlate all possible pairs of values. For a given set of correlated consecutive values $\mathcal{O}_i$, we compute the estimator

$\overline{\Gamma}(j)$ as follows

$$\overline{\Gamma}(j) = \frac{1}{\overline{\mathcal{O}^2} - \overline{\mathcal{O}}^2} \frac{1}{(N-j)} \sum_{i=0}^{N-j-1} \left(\mathcal{O}(i) \cdot \mathcal{O}(i+j) - \overline{\mathcal{O}}^2 \right), \tag{7}$$

where

$$\overline{\mathcal{O}} = \frac{1}{N} \sum_{i=0}^{N} \mathcal{O}(i) \,. \tag{8}$$

Then we can perform numerical integration and obtain an estimate $\overline{\tau_{\text{int}}}$, for example using the parallelogram law

$$\overline{\tau_{\text{int}}}(k) = \frac{\Delta t}{2} \Big(1 + \sum_{j=0}^{k} \big(\Gamma(j) + \Gamma(j+1) \big) \Big) \tag{9}$$

The value of $\overline{\tau_{\text{int}}}$ can then be determined self-consistently from the value of $\overline{\tau_{\text{int}}}(k)$ at $k \geq 6\,\overline{\tau_{\text{int}}}(t)/\Delta t$ (see also [1]). However, this is only true if $\tau_{\text{int}} \ll N\Delta t$. Note that the last condition requires that the true correlation time, not the estimator be much smaller than the total simulation length.

In practice the formla of equation 9 $\overline{\Gamma}(j)$ is feasible up to $j_{\max} \approx 10^4$. The required number of operations is proportional to $N(N-1)$ and thus the effort of computing correlations grows with N^2. For longer lag times (higher $j_{\max}$), more complicted algorithms have to be used. When the time series can still be stored in computer memory, it is possible to use fast Fourier transform (FFT).

Another practical problem is that the error of estimating $\overline{\Gamma}(j)$ also grows very quickly with j and soon it becomes comparable to the measured value. Beyond this point, the estimators $\overline{\Gamma}(k)$ and $\overline{\tau_{\text{int}}}(k)$ attain random values which have nothing in common with the true values $\Gamma(k)$ and $\tau_{\text{int}}(k)$.

2 Simulation programs

To save time and keep our attention focused on physics and statistics, we avoid programming in this tutorial. Instead, we will re-use the programs from the last semester and provide one additional to compute autocorrelation functions. A brief help on the use of all the programs can be obtained upon calling them with no arguments.

Apart from earlier programs, a new program for autocorrelations is provided. It computes the estimator for $\overline{\Gamma}(j)$ and $\overline{\tau_{\text{int}}}(k)$ according to formulas 7 and 9 respectively. It also computes the self-consistent estimate of $\overline{\tau_{\text{int}}}$ but it does not check if the conditions named above are fulfiled. This has to be checked by the user, for example by plotting $\overline{\Gamma}(j)$ and $\overline{\tau}(k)$ in the concerned range.

As usual, the code is compiled by typing

```
make
```

and it takes the following arguments

```
autocorr <filename> <n_lines> <column> <tmin> <tau_max> <dt>
```

For convenience, you can get a short help when invoking the program without any arguments. Of course, you are encouraged to read the code and understand how it works.

3 The tasks

In the first task, we take the simple ising model and relatively well-behaved quantities. Well behaved in a sense that their correlations decay exponentially and relatively fast.

When analyzing the correlations, remember that useful information is contained at $\tau \lesssim 6\,\tau_{\text{int}}$. The tail is just noise and due to accumulation of errors the estimated $\overline{\Gamma}(\tau)$ can take arbitrary shape at high τ.

3.1 Simple correlation functions: follow the recipe

Task: (4 points)

1. *Perform the simulation of the ising model at temperature $T = 2.5$ with the following parameters:* `n_sweeps=100000`*,* `lattice_size=50`*,* `dump_frequency=10000`*. The output file* `magnetization.dat` *contains three columns: MC step, magnetization and energy, respectively.*

2. *Use the* `analyze` *routine from the ising tutorial to measure the averages and Estimate the correlation time of magnetization (column 2) using the binning method. Take* `tmin=10000` *and different numbers of blocks as $N_{\text{blocks}} = 4 + 2^n$, $n \in \mathbb{N}$. Use all technically possible values of N_{blocks}, even those which make* `analyze` *produce a hangman.*

3. *Plot the estimated correlation time as a function of the number of blocks and as a function of block size. Log-scale for N would be convenient!*

4. *Based on the plot, try to estimate the correct value of τ_{int} and discuss the possible confidence interval. Keep in mind that the estimate of τ_{int} is reliable only when two contradicting requirements are satisfied simultaneously: $N \gg 1$ and $k \gg \tau_{\text{int}}$ where k is the block size.*

5. *Using the* `autocorr` *routine, compute the autocorrelation function and $\overline{\tau_{\text{int}}}(t)$ for both the magnetization and energy. Take* `tau_max=100`*. Plot the autocorrelation function and $\tau_{\text{int}}(t)$ and discuss their shapes.*

6. *Estimate the correlation time by identifying where $\overline{\tau_{\text{int}}}(k)$ reaches a quasi-plateau.*

7. *Estimate the correlation time by fitting the initial part of $\Gamma(\tau)$ using equation 4. Plot your data together with the fit function. Use x-range $[0 : 6\overline{\tau_{\text{int}}}]$ and logarithmic y-scale.*

8. *Put estimates from different methods in a table. Where available, include the error or try to make an educated guess. Discuss the (dis)agreement between different methods. Which methods would you trust and why? Are error estimates provided by the fitting program trustworthy?*

9. *Repeat the same procedure for energy density (column 3 of magnetization.dat). The energy density is not as well-behaved quantity as magnetization, so you need to be more careful analyzing the data.*

3.2 Explore the realm of ising correlations

Task: (2 points)

1. *Try to measure how the correlation time depends on temperature. Do the analysis only for magnetization.*
2. *Perform simulations also for T = 2.35, 2.4, 2.7, 3.0. Be aware that* each *temperature requires different length of the simulation and equilibration!*
3. *Plot* $\Gamma(\tau)$ *for different temperatures in one plot.*
4. *Plot* $\tau_{\text{int}}(t)$ *for different temperatures in one plot.*
5. *Estimate* τ_{int} *from your results.*
6. *Plot* $\log(\overline{\tau_{\text{int}}})$ *against* $\log(|T - T_{\text{c}}|)$ *where* $T_{\text{c}} = 2/\ln(1+\sqrt{2})$ *is the critical temperature of the 2D ising model.*

3.3 Correlations in a LJ liquid

Task: (2 points)

1. *Perform a simulation of a LJ liquid in the Langevin thermostat at* $T = 1.2$, $\rho = 0.5$, $N = 256$, $\Delta t = 0.01$ *and* $\gamma = 1.0$. *Again, you are responsible for determining the necessary simulation time and discarding the equilibration.*
   Hint: try starting with `ns=10000` *timesteps to get the first result quickly and then make your simulation longer until you find the plot of* $\overline{\Gamma}(\tau)$ *satisfactory.*
2. *Analyze average system temperature in a similar fashion as you analyzed magnetization in Homework 1.*
3. *Estimate correlation times of* T *at more values of* γ*: 0.5 0.7, 1.4 and 2.0.*
4. *Plot the estimated correlation times (including errors) as a function of* γ^{-1} *and discuss the plot.*

Optional task: (2 bonus points)

Error of the error

As it was mentioned in the introduction, errors in $\overline{\Gamma}(\tau)$ grow very quickly with increasing τ. Therefore, one would like to go one step further and obtain some error estimates also for the autocorrelation function.

1. *Modify the* `autocorr` *code so that it uses binning method to compute error estimates for $\overline{\Gamma}(j)$ at each j.*
2. *Use the error propagation formula [2] to use the errors of $\overline{\Gamma}(\tau)$ and obtain error estimates for $\overline{\tau_{\text{int}}}(t)$. For simplicity, neglect the covariance term.*
3. *Apply the program to analyze the results of Homework 1.*
4. *Plot how the absolute and relative errors of $\overline{\Gamma}(\tau)$ and $\overline{\tau_{\text{int}}}(t)$ grow with time.*
5. *Compare your error estimates to the self-consistent estimates from* `autocorr` *and those from fitting the exponential to the initial part of $\overline{\Gamma}(\tau)$. If desired, try running a longer simulation for better accuracy.*

Additional suggestions for playing around (no points)

1. Try what happens to $\overline{\Gamma}(\tau)$ if you forget to exclude the equilibration or a part of it.
2. Try to see how the correlation time increases with the lattice size in the Ising model.
3. Follow how the correlation times change with system size in the LJ liquid. Note qualitative differences between different quantities (T, p, E_{kin}).

References

[1] `http://www2.fz-juelich.de/nic-series/volume10/janke2.pdf`

[2] `http://mathworld.wolfram.com/ErrorPropagation.html`