Tutorial

1: Error Analysis

Joan J. Cerdà*, Nadezhda Gribova

April 21, 2010
ICP, Uni Stuttgart

Contents

1 Introduction **1**

2 Generating Correlated Series. **2**

3 Analysing Correlated Series. **2**

4 Estimating errors from correlated series. **4**

5 To learn more **5**

6 References **5**

1 Introduction

This tutorial is intended to enlarge your practice on Error Analysis techniques applied to series of correlated data. As you know from preceding lectures, for a given observable, Monte Carlo and Molecular Dynamics simulations can lead to a set of data with a high degree of correlation among them. Thus, being able to quantify the degree of correlation, as well as being able to estimate properly the statistical error inherent to our observable is of crucial importance.

The present tutorial is based on the very interesting article by Wolfhard Janke, (see references and download address below in this tutorial). Specially relevant for this tutorial are pages 430 to 438 of that article. It is advisable to read it before starting to do homewirk for this tutorial. Do

*jcerda@icp.uni-stuttgart.de

first a fast reading (do not worry about the fine details), and later once you have an idea of what is the contain, proceed to do a more detailed reading.

Comments for presenting the homework: For each task in the homework that asks to produce/compare some results you need to present a plot and some explanation of it. Deadline for handing in this homework is 14.05.09.

2 Generating Correlated Series.

Janke's article on page 435 (see equation 41) describes a bivariate times series model to obtain a series of correlated numbers following a Gaussian distribution population with zero mean, and variance unity. Such numbers have the property of being correlated with a certain exponential autocorrelation time τ_{exp}. Such number can be generated e.g. by

$$e_0 = e_0^{'}, \tag{1}$$

$$e_i = \rho e_{i-1} + \sqrt{1-\rho^2}e_i^{'} \tag{2}$$

where the $e_i^{'}$ are non-correlated (independent) Gaussian random numbers with zero mean, $\langle e_i^{'} e_j^{'} \rangle = \delta_{ij}$ and ρ is correlation coefficient $(0 \leq \rho \leq 1)$. In the folder *sources* you will find the program *"example_generator_series.c"* which provides a valuable example of how to implement such a Gaussian-correlated-generator.

Tasks

1. Create a series of 10^5 elements which follows a Gaussian Population distribution with mean $\mu = 0$, standard deviation $\sigma = 1$, and $\tau_{exp} = 10$. Hereby we will refer to this series as *Series-A*.

2. Create a similar *Series-B* but $\mu = 10$, standard deviation $\sigma = 20$, and $\tau_{exp} = 150$. **Hint:** if $G = N(\mu, \sigma^2)$ then $N(a\mu + b, (a\sigma)^2) = aG + b$

3. Check that *Series-A* and *Series-B* obey the expected Gaussian distribution by plotting the normalised histograms of the series against the theoretical curve

$$N(\mu, \sigma, x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

   (Here we recommend to use the *xmgrace* software, because histograms are straightforward to be made using such software).

3 Analysing Correlated Series.

In the previous section we have familiarised with a toy model which generates a series of data with a certain specific correlation. The advantage of working with such toy model is that we can do the analysis of the series both analytically and numerically. Therefore, we can compare the analytical predictions with the results we will obtain in this tutorial.

The normalised autocorrelation function $A(k)$ (see equation 25 on page 431) is defined as

$$A(k) = \frac{\langle O_i O_{i+k} \rangle - \langle O_i \rangle \langle O_i \rangle}{\langle O_i^2 \rangle - \langle O_i \rangle \langle O_i \rangle} \tag{3}$$

which in the limit $k \to \infty$ decays as

$$A(k) \to \exp\left(-k/\tau_{exp}\right) \tag{4}$$

The program *"get_autocorrelation_function.c"* gives an example of how to obtain $A(k)$ from a given series. If we use our toy model, the previously generated series, *A* and *B*, should exhibit an autocorrelation function purely exponential, i.e,

$$A(k) = \exp\left(-k/\tau_{exp}\right) \tag{5}$$

The proper integrated autocorrelation time of a series, τ_{int} is defined (see equation 24 on page 431) as

$$\tau_{int} = \frac{1}{2} + \sum_{k=1}^{N} A(k) \left(1 - \frac{k}{N}\right) \tag{6}$$

and the "running" autocorrelation time estimator,$\tau_{int}(k_{max})$ is defined (see equation 33 on page 433) as

$$\tau_{int}(k_{max}) = \frac{1}{2} + \sum_{k=1}^{k_{max}} A(k) \tag{7}$$

Theoretical predictions for τ_{int}, and $\tau_{int}(k_{max})$ in the case of bivariate Gaussian Series are (see equations 45 and 47 on page 437)

$$\tau_{int} = \frac{1}{2} \coth\left(\frac{1}{2\tau_{exp}}\right) \sim \tau_{exp}\left[1 + \frac{1}{12\tau_{exp}^2} + O\left(\frac{1}{\tau_{exp}^4}\right)\right] \tag{8}$$

$$\begin{aligned} \tau_{int}(k_{max}) &= \tau_{int}\left[1 - \frac{2\exp\left(-(k_{max}+1)/\tau_{exp}\right)}{1+\exp\left(-1/\tau_{exp}\right)}\right] \\ &\sim \tau_{int}\left[1 - \frac{2\tau_{exp}}{1+2\tau_{exp}} \exp\left(-k_{max}/\tau_{exp}\right)\right] \end{aligned} \tag{9}$$

Tasks

1. Obtain the autocorrelation function for a series of 10^5 elements following a Gaussian Population distribution in which no correlations exist among the numbers of the series. Is the Gaussian random number we have used until now a suitable one? Is it possible to get negative values for the autocorrelation time? Justify the answer.

Homework 1 (6 points)

1. Obtain the autocorrelation functions of *Series-A* for three different maximum correlation times $k_{max} = 1000$, $k_{max} = 100$ and $k_{max} = 50$. Compare them against the theoretical expression. Which of the k_{max} seems to give a better agreement against the theoretical predictions ? Why? **Hint:** Take a look at page 433 of Janke's article.

2. Redo the case $k_{max} = 100$ for three different seeds of the random number generator. What do you observe?

3. Which conclusions should be derived from the previous two experiences?

4 Estimating errors from correlated series.

Let's suppose that our *Series-A* and *Series-B* are the outcome of measuring two physical observables *A* and *B*. We want to estimate the expected value of this observables $\langle A \rangle$ and $\langle B \rangle$, as well as to estimate the statistical error we have for these mean values $\epsilon_{\bar{A}}$ and $\epsilon_{\bar{B}}$ respectively. In order to estimate the error, we will use the "one-sigma" definition, (see page 431 on Janke's paper, specially equation 23)

$$\epsilon^2_{\bar{O}} \equiv \sigma^2_{\bar{O}} = \frac{2\tau_{int}\sigma^2_{O_i}}{N} \tag{10}$$

Notice the difference between the variance of the individual measurements $\sigma^2_{O_i}$, and the variance of the expectation value $\sigma^2_{\bar{O}}$.

Tasks

1. Using the integrated autocorrelation time for *Series-A* and *Series-B*, obtain an estimate for $\epsilon_{\bar{A}}$ and $\epsilon_{\bar{B}}$. Express the results of $\langle A \rangle$ and $\langle B \rangle$ and their errors in the appropriate round-error format.

Homework 2 (2 points)

1. Compare the previous results with the errors one would obtain when using the typical formula for uncorrelated measures (see equation 19 on page 430), $\epsilon^2_{\bar{O}} \equiv \sigma^2_{\bar{O}} = \frac{\sigma^2_{O_i}}{N}$. What is the moral?

As surely you have already appreciated, the computation of autocorrelation functions, integrated autocorrelation times, etc. is quite tedious. An alternative way of obtaining an estimate of the errors $\epsilon_{\bar{A}}$ and $\epsilon_{\bar{B}}$ consist on using the so-called Binning Analysis (aka Blocking Analysis), or the so-called Jack knife Analysis. Both techniques are explained in detail on pages 434 and 435 of Janke's the paper. The program *"get_variance_by_binning.c"* provides a way of implementing the binning analysis. Nonetheless, remember always that a roughly knowledge of the degree of correlation we have in our data set is necessary to use properly the Binning and the Jack knife analysis.

Tasks

1. Use the Binning analysis technique to obtain an estimate of the errors $\epsilon_{\bar{A}}$ and $\epsilon_{\bar{B}}$. Express the results of $\langle A \rangle$ and $\langle B \rangle$ and their errors in the appropriate round-error format. On doing the binning analysis, select the length of the blocks in a proper way, and justify the election.

Homework 3 (2 points)

1. How do the integrated autocorrelation times obtained in the binning analysis compare with the ones obtained through the autocorrelation function.

5 To learn more

Binning and Jack knife techniques are just the tip of the iceberg in statistical error analysis of data series. Other related methods are cross-validation, randomisation tests, permutation tests, non-parametric bootstrapping (parametric, and non-parametric). As a further reading see for instance

- John Fox, *Bootstrapping Regression Models* , http://cran.r-project.org/doc/contrib/Fox-Companion/appendix-bootstrapping.pdf
- Moore, D. S., G. McCabe, W. Duckworth, and S. Sclove, *Introduction to the Practice of Statistics*, (2003): *Bootstrap Methods and Permutation Tests*, http://bcs.whfreeman.com/pbs/cat_140/chap18.pdf
- For a more recent version, see Hesterberg, T. C., D. S. Moore, S. Monaghan, A. Clipson, and R. Epstein, *Introduction to the Practice of Statistics*,(2005): *Bootstrap Methods and Permutation Tests*, http://bcs.whfreeman.com/ips5e/content/cat_080/pdf/moore14.pdf

6 References

Wolfhard Janke, *Statistical Analysis of Simulations: Data Correlations and Error Estimation*, published in: *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*, Lecture Notes, J. Grotendorst, D. Marx, A., Muramatsu (Eds.), John von Neumann Institute for Computing, Jülich, NIC Series, Vol. **10**, ISBN 3-00-009057-6,pp 423-445, 2002.

Download for classroom purposes: *http://www.fz-juelich.de/nic-series/volume10*