

# Physics G6080 – Problem Set 3

Fall 2022

Due by midnight Tuesday October 25, 2022

## 1 Statistical Analysis of Mock Dataset

In this problem, you will analyze a mock data set that is available on the course web site in the Assignments area for Problem Set 3.

There are 5 plain text files given, called v1, v2, v3, v4 and v5. (v1 is short for variable 1, etc.) Each text file contains 1,600,000 lines, which correspond to measurements of the given variable from a simulation. There are autocorrelations (correlations in time or measurement number) for each variable. There are also correlations between variables. In this problem, you need to analyze this data, using the methods discussed in class, to find average values, standard deviations of means, autocorrelations, etc.

Numpy has a statistics package that calculates many of these items. You can use this, or write out the calculations explicitly. For the jackknife analysis in problem 2, you can also use the numpy packages, but you will need to be careful about where the factors of  $N$  go. I sometimes will write out an explicit calculation for something like a covariance and compare it with a package/library call, just to make sure I am confident of the normalizations.

We will use  $M$  to represent the total number of values, *i.e.*  $M = 1,600,000$ .  $M$  is large enough so that it represents an effectively “infinite” sample size. Thus, the true data mean can be well approximated by averaging over all  $M$  values. Averages, standard deviations, etc. determined from the “infinite” sample will be denoted with a hat, *i.e.*  $\hat{v}_1$ , etc.

We will use  $N$  to represent the number of measurements in a sample of the data.  $N$  corresponds to the amount of data you might actually collect in a simulation.

1. Determine the true means,  $\hat{v}_a$ , for  $v_1, v_2, \dots, v_5$  from all  $M$  values.
2. Consider two values for our sample size:  $N = 1,000$  and  $N = 10,000$ . There are  $M/N$  samples of this size in our  $M$  values. Histogram the sample means for these values of  $N$  and determine the true standard deviation of the means  $\hat{\sigma}_{\bar{v}_a, N}$ . You can determine  $\hat{\sigma}_{\bar{v}_a, N}$  by assuming each of the  $M/N$  samples is independent. This is reasonable, provided the autocorrelations in the data are smaller than the values of  $N$  you use. Do your two values for  $\hat{\sigma}_{\bar{v}_a, N}$  show the correct behavior with  $N$ ?
3. You can now determine the true autocorrelation function for each variable,  $\hat{C}_{v_a, n}$ , which is given by

$$\hat{C}_{v_a, n} = \frac{1}{M-n} \sum_{i=1}^{M-n} (v_{a, i+n} - \hat{v}_a) (v_{a, i} - \hat{v}_a) \quad (1)$$

Here  $n$  goes from 0 to some maximum value  $n_{\text{cut}}$  with  $n_{\text{cut}} \ll M$ . Plot  $\hat{C}_{v_a,n}/\hat{C}_{v_a,0}$  versus  $n$  for  $a = 1 \dots 5$ .

4. Find the integrated autocorrelation times

$$\hat{\tau}_{\text{int},v_a} \equiv \frac{1}{2} \frac{1}{\hat{C}_{v_a,0}} \sum_{n=-n_{\text{cut}}}^{n_{\text{cut}}} \hat{C}_{v_a,n} \quad (2)$$

Estimate a value for  $n_{\text{cut}}$  from your plots.  $n_{\text{cut}}$  should be large enough that  $\hat{C}_{v_a,n}/\hat{C}_{v_a,0}$  has gotten close enough to zero that the value of  $\hat{\tau}_{\text{int},v_a}$  is not effected by modest changes in  $n_{\text{cut}}$ .

5. Calculate the true standard deviation of the data, *i.e.*

$$\hat{\sigma}_{v_a}^2 \equiv \frac{1}{M-1} \sum_{i=1}^M (v_{a,i} - \hat{v}_a)^2 \quad (3)$$

For a sample of size  $N$ , we should have

$$\hat{\sigma}_{\hat{v}_a,N} = \sqrt{\frac{2\hat{\tau}_{\text{int},v_a}}{N}} \hat{\sigma}_{v_a} \quad (4)$$

Does this relation hold for your analysis?

6. Calculate the true covariance matrix for the data, defined by

$$\hat{c}_{v_a,v_b} = \frac{1}{M} \sum_{i=1}^M (v_{a,i} - \hat{v}_a) (v_{b,i} - \hat{v}_b) \quad (5)$$

It is customary to define a normalized version of  $\hat{c}_{v_a,v_b}$  by

$$\hat{\rho}_{v_a,v_b} \equiv \frac{\hat{c}_{v_a,v_b}}{\hat{\sigma}_{v_a} \hat{\sigma}_{v_b}} \quad (6)$$

Since  $\hat{c}_{v_a,v_a} = \hat{\sigma}_{v_a}^2$ ,  $\hat{\rho}_{v_a,v_b}$  has ones on the diagonals and the off diagonals give a ready measure of the covariance between variables.

7. Now, pick two groups of data from the full universe of data. One should have  $N = 1,000$  and the other should have  $N = 10,000$ . These two groups represent results one might get from simulations. We want to see how well these groups reproduced the true statistical results for these data. Estimate the autocorrelation function  $C_{v_a,n}$  from these two groups and the integrated autocorrelation time. Use these to determine the standard deviation of the mean  $\sigma_{\hat{v}_a,N}$ . Compare this with the results from the universe of data. Also compare the normalized covariance matrix  $\rho_{v_a,v_b}$  from these small samples with the universe of data.

## 2 Jackknife Analysis

In this problem, we will see how the jackknife method can be used to find errors on functions on the mean values of data, *i.e.* on  $f(\bar{v}_a)$ . We can use the universe of data to calculate the errors and then calculate the same error from a sample of size  $N = 5,000$ .

We will consider two functions in what follows

$$f_1(\bar{v}_a) = \bar{v}_1/\bar{v}_2 \quad (7)$$

$$f_2(\bar{v}_a) = \exp(\bar{v}_3 - \bar{v}_4) \quad (8)$$

1. Break the  $M$  measurements up in to groups of size  $N$ , calculate  $\bar{v}_a$  for each group and then calculate  $f_i(\bar{v}_a)$  for each group. Calculate these functions of the data means for all  $M/N$  groups and find the standard deviation for  $f_i(\bar{v}_a)$ ,  $\hat{\sigma}_{f_i,N}$ .
2. Calculate  $\hat{\sigma}_{f_i,N}$  from naive propagation of errors, *i.e.* using  $\hat{\sigma}_{\bar{v}_a,N}$  and neglecting correlations between the  $v_i$ . Compare with your results from part 1.
3. We now want to estimate  $\sigma_{\bar{v}_a,N}$  and  $\sigma_{f_i,N}$  using the jackknife method from a single sample of size  $N$ . First we must deal with the autocorrelations in the data, and you have an idea of the integrated autocorrelation time from the first problem. We proceed as follows here. Average your  $N$  data values into bins of size  $b$ . This will produce  $N/b$  data values. Then use the jackknife method to estimate  $\bar{v}_a$  and  $\sigma_{\bar{v}_a,N}$  from these  $N/b$  data values. The jackknife method resums these  $N/b$  values as done in class, *i.e.*

$$v'_{a,k} = \frac{1}{N/b - 1} \sum_{i=1, i \neq k}^{N/b} v_{a,i} \quad (9)$$

From these jackknife values, you can determine  $\bar{v}_a$  and  $\sigma_{\bar{v}_a,N}$ . Do this for a few different values of  $b$  comparable to the integrated autocorrelation time to check that your results do not depend strongly on  $b$ .

4. Now calculate  $f_i(v'_{a,k})$  for each of the  $N/b$  jackknife blocks. You can then determine  $\sigma_{f_i,N}$  from

$$\sigma_{f_i,N}^2 = \frac{N/b - 1}{N/b} \sum_{k=1}^{N/b} (f_i(v'_{a,k}) - f_i(\bar{v}_a))^2 \quad (10)$$

Again, do this for a few values of  $b$  that are comparable to the integrated autocorrelation time. How does  $\sigma_{f_i,N}$  compare with  $\hat{\sigma}_{f_i,N}$  from part 1?

### 3 Argon Molecular Dynamics

We can now apply these statistical ideas to the results of your argon MD simulation. Run as long a simulation as is practical and make measurements of the temperature, potential energy and the time average of the virial, which is given by

$$\sum_i \sum_{j>i} r_{ij} \frac{\partial V_{ij}}{\partial r_{ij}} \quad (11)$$

every MD time step. You should be able to run a few thousand steps, after thermalization. Measure the autocorrelation times for the temperature, potential energy and virial. Also measure the covariance matrix for these 3 quantities. Use your estimate of the autocorrelation times, along with binning and the jackknife method to give an error on the pressure from your simulation.