Note on error estimate and superjackknife

Masaaki Tomii

May 14, 2024

0.1 Warm up

Let's consider observables x_a sampled from a set configurations, X_a. The number of samples

$$N_a = \#X_a, \tag{1}$$

$$N_{a\cap b} = \#(X_a \cap X_b). \tag{2}$$

The full ensemble $X \supset X_a$ for any subensemble a.

Mean value of the covariance matrix:

$$\bar{\sigma}^2_{ab} = \frac{1}{N_{a\cap b}-1} \sum_{i\in X_a\cap X_b} (\bar{x}_a - x^i_a)(\bar{x}_b - x^i_b) \xrightarrow{N_{a\cap b}\to\infty} \sigma^2_{ab}. \tag{3}$$

Consider $f(\vec{x})$, a function of $\vec{x}$, and a naïve approach,

$$\bar{f} = f(\vec{\bar{x}}), \tag{4}$$

with its squared error

$$(\delta f)^2 = \sum_{a,b} \frac{N_{a\cap b}}{N_a N_b} f_a f_b \sigma^2_{ab}, \tag{5}$$

where

$$f_a = \partial_a f|_{\vec{x}=\langle\vec{x}\rangle}. \tag{6}$$

Since we do not know the true values of σ_{ab} or f_a, we replace them as their mean values $\bar{\sigma}_{ab}$ and $\bar{f}_a = \partial_a f|_{\vec{x}=\vec{\bar{x}}}$ and give an error estimate

$$(\overline{\delta f})^2 = \sum_{a,b} \frac{N_{a\cap b}}{N_a N_b} \bar{f}_a \bar{f}_b \bar{\sigma}^2_{ab}. \tag{7}$$

0.2 Jackknife

For $i \in X$, define

$$\bar{x}_a^i = \begin{cases} \frac{N_a \bar{x}_a - x_a^i}{N_a - 1} & \text{if } i \in X_a \\ \bar{x}_a & \text{else} \end{cases} . \tag{8}$$

Evaluate a kind of deviation a follows:

$$\begin{aligned} \sum_{i \in X} (\bar{f} - f(\vec{\bar{x}^i}))^2 &= \sum_{i \in X} \left(\sum_a \bar{f}_a (\bar{x}_a - \bar{x}_a^i) \right)^2 \\ &= \sum_{a,b} \bar{f}_a \bar{f}_b \sum_{i \in X} (\bar{x}_a - \bar{x}_a^i)(\bar{x}_b - \bar{x}_b^i) \\ &= \sum_{a,b} \frac{1}{(N_a - 1)(N_b - 1)} \bar{f}_a \bar{f}_b \sum_{i \in X_a \cap X_b} (\bar{x}_a - x_a^i)(\bar{x}_b - x_b^i) \\ &= \sum_{a,b} \frac{N_{a \cap b} - 1}{(N_a - 1)(N_b - 1)} \bar{f}_a \bar{f}_b \bar{\sigma}_{ab}^2. \end{aligned} \tag{9}$$

This error estimate is slightly different from the naïve one given in Eq. (7). If we substitute the jackknife samples as

$$\bar{x}_a^i \to \bar{x}_a + \sqrt{\frac{N_a - 1}{N_a}} (\bar{x}_a^i - \bar{x}_a), \tag{10}$$

the diagonal ($a = b$) contributions to the error becomes the same as those in Eq. (7). But that is not necessarily the case for the off-diagonal ($a \neq b$) part. The difference from Eq. (7) in the off-diagonal part is only up to a factor which disappears in the limit $N_{a \cap b} \to \infty$ (and hence $N_a, N_b \to \infty$). Thus, this kind of discussion is not important and it may be more valuable to spend our time for other works. If anyone is nevertheless eager to know how to achieve the same error estimation as in Eq. (7) with the jackknife method, the answer is given as follows.

Instead of the definition of jackknife samples given in Eq. (8), let's view the definition as a linear map centering the mean value,

$$\bar{x}_a^i = \bar{x}_a + \sum_{a'} L_{aa'} (\bar{x}_{a'} - x_{a'}^i), \tag{11}$$

where we define $x_a^i = \bar{x}_a$ for $i \notin X_a$. For the standard jackknife, $L_{aa'} = \delta_{aa'}/(N_a - 1)$. With this definition, the jackknife average stays the mean value $\bar{x}_a$, and Eq. (9) becomes

$$\sum_{i \in X} (\bar{f} - f(\vec{\bar{x}^i}))^2 = \sum_{a,b,a',b'} \bar{f}_a \bar{f}_b L_{aa'} L_{bb'} (N_{a' \cap b'} - 1) \bar{\sigma}_{a'b'}^2. \tag{12}$$

In order for this to be the same as Eq. (7), it is enough to require

$$LKL^T = H, \tag{13}$$

where

$$K_{ab} = (N_{a\cap b} - 1)\bar{\sigma}^2_{ab}, \tag{14}$$

$$H_{ab} = \frac{N_{a\cap b}}{N_a N_b}\bar{\sigma}^2_{ab}. \tag{15}$$

If K and H are real symmetric positive matrix, there are many solutions to Eq. (13). However, I do not see any solution independent of $\bar{\sigma}^2_{ab}$. Since Eq. (13) is real symmetric, there are $\frac{D(D+1)}{2}$ constraints on L, where D stands for the size of matrices under consideration. Therefore, the solution has $\frac{D(D-1)}{2}$ degrees of freedom, which may not be enough to eliminate $\frac{D(D+1)}{2}$ independent values of $\bar{\sigma}^2_{ab}$. This means L is expected to be dependent on at least D values from $\bar{\sigma}^2_{ab}$. Unfortunately, $\bar{\sigma}^2_{ab}$-dependence on L is not convenient as it means we need to perform this rescaling each time we introduce new observables sampled from a new subensemble.

My preference is the simple rescaling in Eq. (10) allowing tiny difference in the off-diagonal ($a \neq b$) contribution to the error. If we really need to take into account that difference, I would prefer Matsumoto's method than what we discussed here.

0.3 varying bin size