Note on statistics of descriptors based on product of MD simulations frames

INTRODUCTION

Purpose of this note is to construct statistical test which would examine if the two sets of blocks show significant differences. Assume that we have two sets of frames from different MD simulations and a related to them descriptor f (e.g. RMSD) operating on their subsets. Let each set consists of n blocks of equal size, from which corresponding descriptor values could be calculated and then put into $n \times n$ matrices. For pair of simulations' frames sets three different matrices could be created, consisting of descriptor values of: 1) product between n blocks of the same set, 2) product between different sets of blocks. Let denote $\boldsymbol{A}$, $\boldsymbol{C}$, self-product matrices from first and second simulation subsequently and $\boldsymbol{B}$ descriptor matrix of product between two different simulations. In that notation $a_{i,j}$ denotes the descriptor calculated from i and j block of first and second set used to build $\boldsymbol{A}$ respectively. Let also $\{a_i\}$ represent set of frames from block i of simulation no. 1.

PROBLEM

First obvious approach to this problem would be to gather all descriptor values and examine statistical differences of means of matrices. Because $\boldsymbol{A}$ and $\boldsymbol{C}$ matrices are products of the same sets of blocks in general each of them consists of $n^2/2$ different values (in other words they are symmetrical), while $\boldsymbol{B}$ have n^2 unique values, which summs up to $2n^2$ sized sample. This could lead to carry out two simple tests between differences of matrices: $(\boldsymbol{A};\boldsymbol{C})$ and $(L_{\boldsymbol{A}} + U_{\boldsymbol{C}} - \mathrm{diag}((L_{\boldsymbol{A}} + U_{\boldsymbol{C}})/2; \boldsymbol{B})$, where $L_{\boldsymbol{A}}$ and $U_{\boldsymbol{C}}$ are lower and upper triangular parts of matrices $\boldsymbol{A}$ and $\boldsymbol{C}$ respectively. Such approach would be justified only if the values sampled from simulations were independant.

Recalling the procedure of generating $\boldsymbol{A}$ matrix, one can find its element from following formula

$$a_{i,j} = f(\{a_i\}, \{a_j\}),$$

which clearly shows that the result is obtained by some transformation of sets of frames. Expanding this observation for other elements one can easily notice that $a_{i,j}$ is somehow correlated to element $a_{i,k}$ for every $j \neq k$, because $a_{i,k}$ is also transformation of set $\{a_i\}$ and some another set. In another words following relation is valid

$$(\forall j \in \{1, ..., n\})(r_{\{a_{i,j}\},\{b_{i,j}\}} \not\approx 0), \qquad (1)$$

where $r_{\{a_{i,j}\},\{b_{i,j}\}}$ is Pearson's correlation coefficient of sets of elements $a_{i,j}$ and $b_{i,j}$ with fixed index i. In the language of probability it could be reformulated as $\mathrm{P}(a_{i,j}, b_{i,j}) \neq \mathrm{P}(a_{i,j})\mathrm{P}(b_{i,j})$. One can also view this phenomenon as using one dataset (e.g. frames from one simulations) of independantly generated samples, where each sample is subsequently used to generate few different results, because every block of sim. no. 1 produces n results. Summing it up, some of matrices elements should not be treated as independantly drew samples.

POSSIBLE SOLUTION

According to the statistical inferring some of the elements should be treated as pairs of correlated results in such order that different pairs fulfil the condition of independance. Recalling the equation 1 one finds that $a_{i,j}$ should be correlated to all elements which were calculated using block i - in our case $\forall k \{a_{i,k}, b_{i,k}\}$, because $b_{i,k}$ were also obtained by some transformation of $\{a_i\}$.

This reasoning leads us to the problem of construction of independant random variables. Let's consider sample of $2n$ results X_i generated as in the following formula

$$\mathrm{X}_i = \left(\sum_{m=1}^{n} a_{i,m}, \sum_{m=1}^{n} b_{i,m}\right) \text{ for } i <= n$$
$$\mathrm{X}_i = \left(\sum_{m=1}^{n} c_{m,i}, \sum_{m=1}^{n} b_{m,i}\right) \text{ for } i > n,$$

where all elements of each X_i are transformation of one block i (for first n results comes from frames $\{a_i\}$, for last from $\{c_i\}$). In such manner each both elements of pair X_i are correlated, but results X_i and X_j are independant with the agreement to at most single components inside sums which repeats in both of them (e.g. for $n = 8$, $a_{1,2}$ will be a part of X_1 and X_2, because $\boldsymbol{A}$ is symmetrical and for $b_{1,2}$ in X_1 and X_{10}), thus the correlation between the different X_i's should be minimized. Moreover it should be possible to estimate it by calculating proper coefficient between the sets of values, which were used to X_i construction similar to the formula below

$$\begin{aligned} r(\mathrm{X}_i, \mathrm{X}_j) = r(&\{a_{i,m}\, b_{i,m} : m \in \{1, ..., n\}\}, \\ &\{a_{j,m}\, b_{j,m} : m \in \{1, ..., n\}\}) \end{aligned}$$

In this example $i, j <= n$. Having that kind of results one can now perform the Student's t-test for paired samples if they come from some normal distribution and Wilcoxon's non-paramteric signed-rank test otherwise. Latter one assumes that distribution of such results should be symmetrical around 0. One can wonder why elements of $\boldsymbol{C}$ were used for creation of samples X_i? Answer is simple - wider problem is to examine if the simulations no.1 and no.2 are significantly different, which means that null hypothesis tested assumes that distributions of $\boldsymbol{A}$ and $\boldsymbol{C}$ are identical thus their elements can be treated in the same way.

Example values and ending remarks. In our case n is 8, so we have $2n = 16$ independant results for pair of simulations, because each simulations consists of 8 blocks. This construction assures that all different elements of $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are used at least once. Although testing normality of the distribution could not produce relevant results, it should not be a problem for non-parametric test. Forementioned three 8×8 matrices are easily reconstructable from 10 matrices of size 4×4. Overally 28 tests could be carried on in this manner (multiplying by number of descriptors $28 * 3 = 84$). Not a word was mentioned about the independance between elements $a_{i,j}$ and $c_{i,j}$ because they are generated independantly (recall that $\boldsymbol{A}$ is self-product of sim no. 1 and $\boldsymbol{C}$ of sim no. 2).