Using covariance weighted euclidean distance to assess the dissimilarity between integral experiments

Fan Kai[1,*], *Li* Fu[2,3], *Wang* Jiangmeng[1], *Yin* Yanpeng[1], *Song* Linli[1], and *Wang* Sanbing[1]

[1]Institute of Nuclear Physics and Chemistry, China Academy of Engineering Physics, Mianyang 621900, Sichuan, China
[2]Institute of Nuclear and New Energy Technology, Tsinghua University, Beijing 100084, China
[3]Key Laboratory of Advanced Reactor Engineering and Safety, Ministry of Education, Beijing 100084, China

Abstract. Integral experiments especially criticality experiments help a lot in designing either new nuclear reactor or criticality assembly. The calculation uncertainty of the integral parameter which is introduced in by the nuclear data uncertainty is larger than the experimental uncertainty for most high-enriched uranium metal experiments, therefore the integral experiment is still very useful. There are lots of integral experiments have been done and documented. It should be considered carefully that which integral experiments should be used in applications. For instance, if the aim of the application is to validate the criticality design of a new reactor, integral experiments which are similar to the new reactor should be used. There are several similarity measures which have been used to assess the similarity between integral experiments, such as E similarity measure, G similarity measure and C similarity measure. But, there is no standard rule to choose which similarity measure should be used to assess the similarity between integral experiments in specific application. Another shortage of these similarity measures is that the thresholds of these similarity measures which should be set to judge whether the integral experiments are similar to each other or not have no clear physical meaning. In this paper, we will analyze the existing similarity measures which have been used to assess the similarity between integral experiments, and test some other similarity or dissimilarity measures which have been used in other research fields. After testing the Tanimato similarity measure and Euclidean distance, we find that the covariance weighted Euclidean distance is well suit to assess the dissimilarity between integral experiments, and the physical meaning of its threshold is clear. We recommend using covariance weighted Euclidean distance to assess the dissimilarity between integral experiments.

1 Introduction

Integral experiments are a kind of experiments which could be used to validate criticality designs (include the criticality safety), nuclear data and transport calculation programs. Integral experiments include criticality experiments, reaction rate ratios experiments and so on. Up to now, there are about 5,000 integral experiments on public in the world, and have been collected into ICSBEP (International Criticality Safety Benchmark Evaluation Project) handbook [1] and IRPhEP (International Reactor Physics Experiments Evaluation Project) Handbook [2]. These experiments perform important roles in fields of criticality design and criticality safety.

Takanobu Kamei and Tadashi Yoshida's work is a representative example which shows an important application of integral experiments [3]. They used integral experiments to predict the performance parameters of large liquid-metal fast breeder reactor core, and estimated the uncertainty of the predicted parameter. First, they selected integral experiments which are similar to the reactor whose parameters are to be predicted. Then they analyzed the biases between the calculation and experimental results of these integral experiments, and predicted the parameters of the new reactor based on the analysis. Finally, by analyzing the discrepancy between the integral experiments and the new reactor, they estimated the uncertainty of the predicted results. By this process, the performance parameter could be predicted so as the uncertainty of the estimated parameter before doing the experiment.

The most important requirement when selecting integral experiments is to select integral experiments which are similar to the new reactor. The existing similarity measures which have been used to assess the similarity between integral experiments include E similarity measure, G similarity measure and C similarity measure. Their definitions are shown in equation (1), equation (2) and equation (3). In these equations, S is the sensitivity vector of the integral parameter (to the nuclear data), and M is the covariance matrix of the nuclear data.

$$E_{ij} = \frac{S_i^{\,t} S_j}{|S_i| \left|S_j\right|} \tag{1}$$

$$G_{ij} = 1 - \frac{\left|S_i - S_j\right|^2}{|S_i|^2 + \left|S_j\right|^2} \tag{2}$$

[*]e-mail: fankai19891027@163.com

$$C_{ij} = \frac{S_i^t M S_j}{\sqrt{S_i^t M S_i}\sqrt{S_j^t M S_j}} \tag{3}$$

Each of these three similarity measures could assess the similarity between integral experiments, but there are two shortages in them. The first shortage is that there is no standard rule to choose which similarity measure to be used, and there might have a large difference between the results of these three similarity measures. The second shortage is that the thresholds of these three similarity measures which should be set to judge whether the two experiments are similar to each other or not have no clear physical meaning.

In this paper, we will analyze the existing similarity measures which have been used to assess the similarity between integral experiments, and test more similarity or dissimilarity measures which have been used in other research fields like pattern recognition. Finally, we recommend using covariance weighted Euclidean distance to assess the dissimilarity between integral experiments.

2 Analysis of the Existing Similarity Measures

2.1 Existing similarity measures

There are several existing similarity measures which have been used to assess the similarity between integral experiments. The definitions have been shown in equation (1), equation (2) and equation (3).

All these three similarity measures assess the similarity between integral experiments through the sensitivity vectors.

E similarity measure is the cosine of the angle between the two sensitivity vectors. If the angle between the two sensitivity vectors is 0°, the *E* similarity measure is equal to 1. As the angle between the two sensitivity vectors getting larger, the *E* similarity measure gets smaller.

E similarity measure assesses the similarity between two integral experiments only through the angle between the two sensitivity vectors. *G* similarity measure takes both the angle and the ratio of the two sensitivity vectors' norms into consideration. Only if the two sensitivity vectors are equal to each other, the *G* similarity measure is equal to 1. As the increasing of the angle between the sensitivity vectors or as the ratio of the two sensitivity vectors' norms getting away from 1, the *G* similarity measure decreases.

$$G_{ij} = 1 - \frac{\left|S_i - S_j\right|^2}{|S_i|^2 + \left|S_j\right|^2} = \frac{2c\cos\alpha}{1 + c^2} \tag{4}$$

It could be seen obvious in equation (4) that *G* similarity measure is a function of the angle and the ratio of the two sensitivity vectors' norms. In equation (4), c is the ratio of the two sensitivity vectors' norms, and α is the angle between the two sensitivity vectors. The function of *G* similarity measure has been presented in Figure (1).

E and *G* similarity measures treat all cross sections equal-weighted. But in some applications, especially in nuclear data validations, sensitivities of the integral parameter to different cross sections should have different weight. This is because the uncertainties of different cross sections are different, we care more about the cross sections with large uncertainties than the ones with small uncertainties. Therefore, *C* similarity measure uses both the covariance matrix of the nuclear cross sections and the sensitivity vectors to assess the similarity between integral experiments. In *C* similarity measure, different cross sections have different weights.

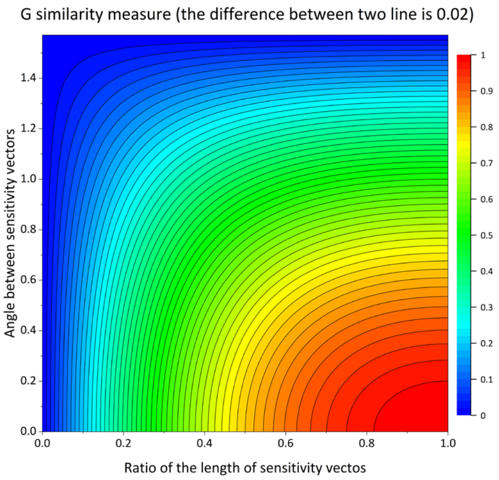

Figure 1. *G* similarity measure

2.2 Shortages of these three similarity measures

There are at least two shortages in these three similarity measures. The first shortage is that there is no standard rule to choose which similarity measure should be used in a specific application. All of these three similarity measures have been used to assess the similarity between integral experiments, but none of them has achieved recognition by most experts. *C* similarity measure has been used widely to assess the similarity between integral experiments in criticality safty validations [4], all of these three similarity measures have been used in nuclear data validations.

The second shortage of these three similarity measures is that the thresholds of these similarity measures which should be set to judge whether the integral experiments are similar to each other or not have no clear physical meanings. For instance, *C* similarity measure could be used in applications of criticality safety validations, and the threshold of it is sometimes set to 0.8. But the physical meaning of the value 0.8 is not clear. The same problem exists in *E* and *G* similarity measures. The setting of the thresholds of these similarity measures mostly depends on the experience and experts opinions, but doesn't depend on the physical meaning of these similarity measures.

3 Dissimilarity and Similarity Measures in other Research Fields

Because the two shortages of these three similarity measures, we analyzed some other similarity measures and dissimilarity measures which have been used in other research fields and tested whether they have the shortages mentioned in subsection 2.2 or not.

We tested Tanimato measure, Euclidean distance and covariance weighted Euclidean distance in this section. We recommend using covariance weighted Euclidean distance to assess the dissimilarity between integral experiments.

3.1 Tanimato measure (similarity measure)

Tanimato measure has been used in pattern recognitions to measure the similarity between vectors. The definition of Tanimato measure is shown in equation (5).

$$T_{ij} \equiv \frac{1}{1 + \frac{(S_i - S_j)^t (S_i - S_j)}{S_i^t S_j}} \qquad (5)$$

Tanimato measure (T similarity measure) is a similarity measure. Figure (2) shows the contour lines of T similarity measure, the difference between the values of two neighbor contour lines is 0.02. It is obvious that the distribution of the contour lines of T similarity measure is more symmetrical than that of G similarity measure, which means that the linearity of T similarity measure is better than that of G similarity measure.

Tanimato measure performs better than E and G similarity measures, because the linearity of T similarity measure is better than that of E and G similarity measures. People prefer thinking in linear mode. Supposing there are three vectors, $(1.000, 0.000)$, $(0.866, 0.500)$ and $(0.500, 0.866)$. The angle between the first vector and the second one is 30°, and the angle between the first vector and the third one is 60°. The second vector is in the center between the first one and the third one. But if we calculate the similarities between them by E similarity measure, the similarity between the first vector and the second one is 0.866, the similarity between the first vector and the third one is 0.500. For whom they don't know the definition of E similarity measure, they might think that the second vector is closer to the first one than to the third one. Therefore, we think T similarity measure is better.

But Tanimato measure also has the shortages mentioned in subsection 2.2.

3.2 Euclidean distance (dissimilarity measure)

In application, either similarity measure or dissimilarity measure could assess the similarity between integral experiments. Thus, we try to use dissimilarity measure instead of similarity measure to assess the similarity between integral experiments.

Euclidean distance is a dissimilarity measure used widely, and the definition is shown in equation (6).

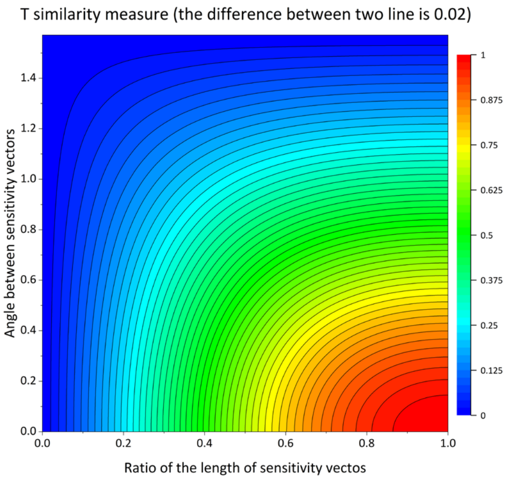

Figure 2. T similarity measure

$$O_{ij} = \sqrt{(S_i - S_j)^t (S_i - S_j)} \qquad (6)$$

To explain the physical meaning of Euclidean distance when it is used to assess the similarity between integral experiments, we suppose that there are two integral experiments, marked with i and j respectively. By sensitivity analysis, we could obtain the sensitivity vectors of these two experiments, S_i and S_j. We could assume that there is a hypothesized system whose sensitivity vector is $S_i - S_j$. These three experiments are shown in Figure (3).

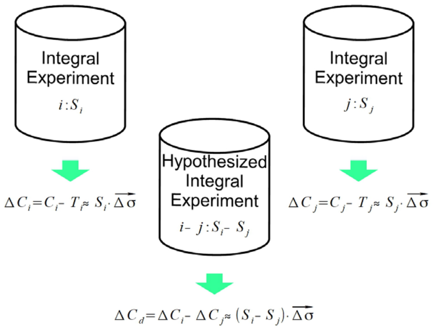
Integral Experiment
$i:S_i$
Integral Experiment
$j:S_j$
Hypothesized Integral Experiment
$i-j:S_i-S_j$
$\Delta C_i = C_i - T_i \approx S_i \cdot \overrightarrow{\Delta\sigma}$
$\Delta C_j = C_j - T_j \approx S_j \cdot \overrightarrow{\Delta\sigma}$
$\Delta C_d = \Delta C_i - \Delta C_j \approx (S_i - S_j) \cdot \overrightarrow{\Delta\sigma}$

Figure 3. Explanation of O dissimilarity measure

When we use a nuclear data library to calculate the integral parameters, the biases of the cross sections ($\overrightarrow{\triangle\sigma}$) are const. And the biases of the calculated results of integral experiments (i and j) could be expressed by equation (7) and equation (8) respectively. In these equations, $\triangle C$ is the bias between the calculated integral parameter (C) and the true integral parameter (T), and S is the sensitivity vector of integral parameter C to the nuclear data.

$$\triangle C_i = C_i - T_i \approx S_i \cdot \overrightarrow{\triangle\sigma} \qquad (7)$$

$$\triangle C_j = C_j - T_j \approx S_j \cdot \overrightarrow{\triangle\sigma} \qquad (8)$$

$$\triangle C_d = \triangle C_i - \triangle C_j \approx (S_i - S_j) \cdot \overrightarrow{\triangle\sigma} \qquad (9)$$

We could define a parameter $\triangle C_d$ as equation (9). We think $\triangle C_d$ could represent the dissimilarity between the two integral experiments and its physical meaning is obvious. If the two experiments are similar to each other, the calculation biases ($\triangle C_i$, $\triangle C_j$) of the two integral parameters should be close to each other, which causes $\triangle C_d$ approaching to 0. Therefore, the dissimilarity between the integral experiments could be measured through the difference between the calculation biases of the two integral experiments. The biases of the nuclear data are unknown, therefore it is convenience to use a norm of $(S_i - S_j)$ as the dissimilarity measure between integral experiments, and we recommend using the Euclidean distance (2-norm). With a fixed $\overrightarrow{\triangle\sigma}$, the dissimilarity (assessed by Euclidean distance) between two integral experiments gets larger as the difference between the calculation biases of the two integral experiments getting larger.

3.3 *F* dissimilarity measure

Euclidean distance treats all nuclear data equal-weighted, but in nuclear data validations, we focus on the cross sections with large uncertainties. Because of that, we use the covariance weighted Euclidean distance instead of Euclidean distance to assess the dissimilarity between integral experiments in nuclear data validations, and we name the dissimilarity measure with F, as equation (**??**), where M is the covariance matrix of the nuclear data.

$$F = \sqrt{(S_i - S_j)^t M (S_i - S_j)} \qquad (10)$$

There is a parameter which have been defined by Takanobu Kamei and Tadashi Yoshida [3] is similar to F dissimilarity measure, as equation (10). But this parameter wasn't used as a dissimilarity measure.

$$V = (S_i - S_j)^t M (S_i - S_j) \qquad (11)$$

The physical meaning of F dissimilarity measure is clear. Consider the hypothesized integral experiment in subsection 3.2 again whose sensitivity vector is $S_i - S_j$. The uncertainty of the calculation result of this system is $\sqrt{(S_i - S_j)^t M (S_i - S_j)}$ which is consistent with F dissimilarity measure. And because $\triangle C_d = \triangle C_i - \triangle C_j$, F dissimilarity measure represents the uncertainty of the difference between the calculation biases of the integral parameter caused by the biases of the nuclear data. If the two integral experiments are similar to each other, the calculation biases of these two integral experiments caused by the biases of the nuclear data should be approximately equal to each other, and the F dissimilarity measure approaches to 0.

The physical meaning of the threshold of F dissimilarity measure is clear too. For instance, if the threshold is set to 0.001, the two integral experiments are similar to each other means that the uncertainty of the calculation biases of the integral experiments caused by the biases of the nuclear data is less than 0 001.

4 Conclusion

In this paper, we analyzed the existing similarity measures which have been used to assess the similarity between integral experiments. We found there are shortages in these similarity measures. Then we tested other similarity or dissimilarity measures in other research fields, and finally we recommend using F dissimilarity measure (covariance weighted Euclidean distance) to measure the dissimilarity between integral experiments. The advantage of F dissimilarity measure is that its physical meaning is clear.

There is an important thing should be noticed is that the ranges of these similarity measures and dissimilarity measures are different. Mathematically, the ranges of E, G and C similarity measures are $[-1, 1]$, the range of T similarity measure is $[-\frac{1}{3}, 1]$. The ranges of O and F dissimilarity measures are $[0, +\infty)$.

Acknowledgements

Project 2017AA01 supported by Key Laboratory of Neutron Physics, CAEP

References

[1] International Handbook of Evaluated Criticality Safety Benchmark Experiments. July 2018. http://icsbep.inl.gov/; https://www.oecd-nea.org/science/wpncs/icsbep/.

[2] International Handbook of Evaluated Reactor Physics Benchmark Experiments. Paris: OECD Nuclear Energy Agency, 2017. (NEA; 7329).

[3] Takanobu Kamei and Tadashi Yoshida, Error Due to Nuclear Data Uncertainties in the Prediction of Large Liquid-Metal Fast Breeder Reactor Core Performance Parameters, Nuclear Science and Engineering, **84:2**, 83-97 (1983).

[4] B.L.Broadhead, B.T.Rearden, and et al., Sensitivity and Uncertainty Based Criticality Safety Validation Techniques, Nuclear Science and Engineering, **146**, 340-366 (2004).