Breaking barriers in analysis: modularizing analysis framework

Y. Kubota[*1,*2] and Y. Ono[*2,*3]

In the field of experimental nuclear physics, data analysis is frequently conducted using programs built on a single analysis framework, such as ROOT. However, these analysis frameworks make it difficult to introduce cutting-edge analysis methods, leading to problems such as difficulties in data visualization. These problems persist even when switching to different analysis frameworks, as those designed for the physics community often lack a comprehensive software ecosystem that allows continuous adoption of state-of-the-art techniques. Recent trends surpass confirmatory analysis, which is widely used in nuclear physics. Exploratory analysis is an emerging trend, driven by advances in mathematical and statistical methods such as deep learning. These techniques have rapidly progressed with implementations in statistical analysis platforms such as R,[1] making them easily accessible. However, while R is useful in analyzing tidy data, it is not effective for raw data preprocessing requiring conversion and calibration, which are inevitable and essential in experimental physics analyses.

We propose modularizing the analysis framework into a collection of loosely-coupled programs, which offers advantages in terms of continuity and future scalability. The analysis can be roughly divided into four stages: decoding binary data obtained from digitizer circuits into typed data, conversion to an intermediate representation of raw data reflecting the experimental conditions, conversion to objectively comparable physical quantities, and conversion to the final result. Each process is handled by one or more dedicated programs. This approach enables introducing parallel processing using graphics processing units (GPUs) or field-programmable gate arrays (FPGAs), making it possible to handle significantly large data. It is essential to select the appropriate data structure for the efficient data processing. For instance, while TTree implemented in ROOT offers the flexibility of storing arbitrary classes, it is not compatible with data processing on GPUs. Furthermore, the supported languages are limited to C++ and Python. Apache Arrow[2] can be considered as an alternative; it provides a language-independent, efficient columnar memory format for analytics on modern hardware and supports fast data access.

In this study, as a first step towards such an attempt, we analyzed data using the logic regression method[3] developed in the field of the genome analysis. Logic regression is an adaptive regression method that attempts to construct predictors as Boolean combinations of binary covariates. This allows analysis, generally described procedurally in the field of nuclear physics to be described in a more mathematically oriented manner. The cosmic-ray data from the GAGG:Ce calorimeter[4] was used in this analysis. First, the raw data obtained using RIBF DAQ were converted into TTree format by a decoder implemented in the framework of anaroot. Subsequently, a preprocessing program based on ROOT transformed the data into physical quantities. The TTree-formatted data was then converted into Apache Arrow format using a newly implemented C++ program. The data was analyzed using the logic regression implemented in the R language. We attempted to visualize correlations in the data structure without *a priori* knowledge of the setup or measurement conditions. Figure 1 illustrates a schematic estimating the hit pattern of GAGG:Ce crystals. Although only the hit patterns (Boolean values) were analyzed in this preliminary result, we obtained hints regarding the geometrical setup of the crystals only from the data. This method may allow systematic correction of data inconsistencies due to wiring errors *etc.*, which have been fixed in a random manner in the past.

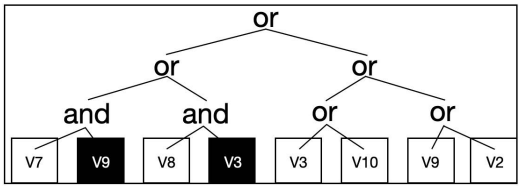

Fig. 1. Schematic representing the estimated equation in logic regression. Each box corresponds to each GAGG:Ce crystal. Black color represents logical negation (no hit of cosmic rays).

In summary, nuclear physics analyses rely on the ROOT; however, this dependency has been posing challenges for modernization, such as utilizing recent statistical methods. We propose establishing a modular, loosely coupled analysis scheme to overcome this problem. As a practical example, we demonstrated the application of logic regression to analyze cosmic-ray data from GAGG:Ce scintillators. Our scheme enables incorporatin pure statisticians, who previously could only join the discussion after the potentially-biased analysis, to directly analyze the data. This opens the door to incorporating insights from other disciplines and exploring possibilities for physics that may have been overlooked before.

*1 RIKEN Nishina Center
*2 RIKEN Cluster for Pioneering Research
*3 St. Luke's Graduate School of Public Health

References

1) R Core Team, `https://www.R-project.org/`.
2) Apache Arrow, `https://arrow.apache.org/`.
3) I. Ruczinski *et al.*, J. Comput. Graph. Stat. **12**, 475 (2012).
4) T. Sugiyama *et al.*, in this report.

DOI:10.34448/RIKEN.APR.57-073