Data Analysis — Algorithms and Tools

This content has been downloaded from IOPscience. Please scroll down to see the full text.
2015 J. Phys.: Conf. Ser. 608 012059
(http://iopscience.iop.org/1742-6596/608/1/012059)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 131.169.4.70
This content was downloaded on 26/05/2015 at 20:48

Please note that terms and conditions apply.
Abstract.
Modeling of detector response, modeling of physics, and software tools for modeling and analyzing are three fields among others that were discussed during 16th International workshop on Advanced Computing and Analysis Techniques in physics research – ACAT 2014. This short report represents a summary of track two where the current status and progress in these fields were reported and discussed.

1. Introduction
There are three fields connected with the data analysis: modeling of the detector response, modeling of the physics and extracting physics from data, and software tools for modeling and analyzing. Each effort in any of these fields requires a knowledge of the status and progress in other fields. Many efforts also require a feedback to be given to make the work successful and useful. In this short report, we summarize the presented methods and tools along with some of the feedback given during the discussions. Each section below is dedicated to one of these three fields.

2. Modeling of detector response
The standard candle in the field of modeling of detectors is Geant4 [1]. Geant4 is a toolkit for the simulation of passage of particles through the matter. It is used in most of the high energy physics experiments including all experiments at the LHC for a reliable full simulation of the detector response. It has also found users in the field of medical science (for simulation of interactions of radiations used for treatment), in the field of space science (to study interactions between the natural space radiation environment and space hardware or astronauts), nuclear physics, and other fields where ionizing effects have to be taken into account. Recent updates of Geant4 consist of updates in the design and updates in the modeling of physics. Updates in the design are namely the use of multi-threading and a strong reproducibility of events (that is the independence of simulation of one event on the history of previous running). Updates in the physics modeling are namely fixing the energy response in QGSP_BERT physics list and improving lateral shower shape, improving the FTF physics list, and the use of a new unique solid library (USolid) [2].

Successful Geant4 project is being followed by GeantV [3] that will have several important new features. Main motivation for the new Geant is the need to speed up the process of particle
transport simulation. Indeed, most of the time is spent on few percents of the volume that is simulated. The main new features of GeantV are following: the vectorization and locality (not a single track but a group of tracks are transported), multi-threading, adding new entities to control the work flow, optimizing simulation geometry, and a possibility to include some fast simulation.

Fast simulation itself will certainly be needed during the LHC run II. Robustness of Geant4 simulation was a part of the success of LHC run I and the Higgs discovery \[1\] \[5\]. However, at the same time, many important physics analyses suffered from the lack of Monte-Carlo (MC) statistics. Thus, fast simulation will be indispensable during the run II. The core of the fast simulation is the use parameterizations of the response or pre-generated samples. Each LHC experiment has its own fast simulation of the detector (e.g. Atlfast of ATLAS or FastSim of CMS). A typical enhancement by about a factor of twenty in the computation speed was reported by ATLAS when using the fast simulation compared to the full simulation. An alternative to the fast simulation only is to combine the fast simulation with the full simulation (e.g. to use the fast tracking in tandem with the full simulation of the rest of the detector) or to replace costly physics objects by pre-clustered ones (e.g. frozen showers used by ATLAS).

Besides the fast simulation tools that are specific for a given experiment, there are tools on the market that can be used to simulate any detector or its part. One of such tools is Delphes \[6\] which is being developed since 2007. Delphes has several attractive features: it contains the simulation of propagation in the magnetic field and the calorimetry; it can simulate multiple simultaneous proton-proton collisions (pileup); it can be used to simulate realistic detector response to high-luminosity LHC (HL-LHC) environment; it has particle-flow algorithm (a strategy for jet reconstruction by CMS). Delphes is community based, modular tool that has many interfaces (e.g. to MadGraph, Hep-MC, FastJet) and a standard root output.

To simulate precisely the tracking system e.g. for a new detector to be used at ILC or other future collider facility one can use Genfit \[7\]. Genfit is a generic track-fitting toolkit that provides a modular track fitting framework and simulation of a tracker. Genfit can use fitting based on standard Kalman Filter (which does a linearization around predictions) \[8\] or it can use an improved fitting algorithm, which utilizes a Deterministic Annealing Filter \[9\]. This improved algorithm allows namely a better handling of out-layer hits.

Tracking and the vertexing may be improved by a use of multivariate analysis techniques (MVA). This was demonstrated by a project of robust tracking with neural network for JLab. JLab experiments investigate the hadron structure using polarized $e^-$ beams at the energy up to 12 GeV. In the new setup, experiments will have to deal with large pile-up and backgrounds. It was proposed that the new tracking with Gas Electron Multiplier (GEM) may use the neural network implemented within a Mean Field theory framework \[10\] to get the association of hits into tracks. Energy which is being minimized is defined in terms of a natural distance measure (reciprocal value of distances between hits times the cosine of an angle formed by line segments connecting the hits) and matrices quantifying connections between two points in subsequent GEM planes. An improved resolution by a factor of eight was reported, however it was also concluded during the discussion that a careful evaluation of computation time requirements is needed.

The use of neural network is planned also to improve the ability of vertex triggering at Belle II at SuperKEKB accelerator. Future SuperKEKB is expected to provide a luminosity of $8 \cdot 10^{35}$ cm$^{-2}$s$^{-1}$ which is forty times larger than the world record by KEKB accelerator. Topological and time information will be provided at Belle II by a Central Drift Chamber. This information is the input to a simple and fast neural network with one hidden layer, one neuron, and fully forward connected multi-layer perceptron. The training can be done with back-propagation algorithm and the output is scaled $z$-vertex. This trigger is being implemented in firmware on Virtex 7 FPGA board. Estimated vertex resolution is 1.3 – 2.3 cm.
Besides track and vertex reconstruction, the particle identification is also important for some of the physics analyses. A promising tool for particle identification without modeling the response, called Pridix [11], was introduced and discussed. Pridix uses minimization of generalized Kullback-Leibler divergence. Kullback-Leibler divergence (also known as information gain or relative entropy) is a non-symmetric measure of the difference between two probability distributions $P$ and $R$. Specifically, the Kullback-Leibler divergence of $R$ from $P$ is a measure of the information lost when $R$ is used to approximate $P$. The technique utilizing this tool from probability theory was shown to be useful and to provide a very good performance of PID at PHENIX detector at RHIC collider.

An interesting application of the knowledge of particle interaction with matter was reported by members of Muon Portal Project [12], an innovative non-destructive scanning based on muon tomography to be used to scan containers in docks to search for weapons and radioactive material. The deflection of muon trajectory in the presence of materials with high proton number was demonstrated to be sufficient for detecting the problematic materials. The basic approach for reconstructing the tracks uses the point of closest approach (POCA) between incoming and outgoing tracks. The scanning algorithm with POCA neglects multiple scattering and it has generally a poor resolution. A performance of an alternative approach which uses the density based clustering with friends-of-friends percolation algorithm (also used in cosmology) was reported. This clustering algorithm acts also as a filter and may be combined with other algorithms to further boost the performance. The estimated start-up of the muon-based scanning is March 2015.

3. Software tools for modeling and analyzing
The standard candle software tool used in high-energy physics is ROOT [13]. ROOT is an object-oriented program and library tailored for particle physics data analysis but it is used in many other fields, e.g. in astronomy, nuclear physics, and data mining. Current ROOT 5 is a versatile tool but sometimes it is mentioned that has also some drawbacks such as difficulty for beginners, heavy use of global variables, or limitations coming from CINT interpreter. Some of these issues are addressed in the new ROOT 6 which contains many great improvements. The main new feature is the use of new interpreter called CLING which use CLANG frontend for LLVM compiler which are standards being used for C++. This allows e.g. for interpreting from includes without dictionaries, improved compilation diagnostics, or C++14 support. Besides the new interpreter, there are many other news such as a new TTreeReader which simplifies reading from trees, a new TFormula which profits from the intime compiler, a new possibility to export graphics to LaTeX, or improved GUI with better axis zooming or guides for object placements. The ROOT 6.00 was released in May 2014. The ROOT 6.02 which is targeted to LHC run II was released in September 2014.

A new tool for automatic differentiation (CLAD) [14] was discussed that may also be available in later releases of ROOT 6. CLAD extends the symbolic differentiation to the compiler as a module. This allows for automatic defining the function used for differentiation which is only declared by the user. Such versatile tool can be used also to boost the performance of minimization and fitting.

More complex tools connected with the ROOT environment are RooFit, HistFactory, RooStat, and RooWorkspace [15]. RooFit [16] is a toolkit for modeling of expected distribution of events. In RooFit, each mathematical concept (such as variable, probability density function (PDF) or its integral) is implemented as a RooFit class. This allows for an abstract manipulation with the mathematical objects within a simple and transparent code. HistFactory [17] allows structured statistical model building for binned likelihood template models. RooStats [18] provides a wide set of statistical tests that can be performed on RooFit models. RooWorkspace [19] is a concept that allows to persist complete likelihood models. These tools were used during
the Higgs discovery. The Higgs discovery and subsequent analysis of its properties are truly a result of collaborative statistical analysis of many signal and control samples which is allowed by the use of these tools.

A framework which is built on top of RooFit, HistFractory, and RooStat is HistFitter [20]. HistFitter is a programmable framework to build and test complex data models. It allows to construct and fit PDFs and provide a statistical interpretation. It has a built-in concepts of control, validation and signal regions with rigorous treatment of extrapolations. It allows simple book-keeping of multiple models and provides tools for graphical representation of the results of complex statistical analysis. HistFitter was used within ATLAS since 2012 and now the framework is publicly available [21].

Reconsideration of data model is being done by all the experiments at LHC to anticipate the upcoming LHC run II. In particular, ATLAS has reported an improvement in the data handling by replacing the Analysis Object Data (AOD). The AOD is a data structure that contains information about reconstructed object such as four-momenta of jets or charged particles. The AOD is legible only in a complicated software of ATLAS called ATHENA which made it practically unusable for the end user and it had to be translated to a legible structure called Derived Physics Data. Now, the AOD is replaced by so called xAOD which is a structure legible both in ATHENA and ROOT. This should generally save PB of memory and lot of computation time. The goal of the new data model is also to provide a simple access to events and kinematics of objects used in the data analysis by a light weighted code structure. CMS has reported the upgrade in the simulation which, among other issues, targets to solve the problem of demanding simulation of pileup.

4. Modeling the physics and extracting physics from data
MVA techniques boosted the ability of high energy physics (HEP) to analyze complex event topologies. The important question raised in the conference was if the field of HEP is still the state of the art in MVA, in particular whether some novel techniques used in other fields of science should be adapted to further enhance the performance of HEP analyses. A partial answer to that question was provided by a report discussing a comparison of different MVA techniques within inclusive top pair production cross-section measurement in D0 at Tevatron. In that study, several MVA techniques have been compared:

- Generalized Linear Models (GLM) that are used namely in econometrics
- Model based clustering (MBC) that is used often in acoustics
- Neural networks with cooperative switching units (NNSU) – an enhanced model of neural network designed originally for classification tasks
- Boosted Decision Tree (BDT) from standard TMVA package [22]
- Multilayer Perceptrons (MLP) from TMVA package

Last two above mentioned techniques are golden standards used for years in HEP. It was demonstrated that if the techniques are used in an appropriate way and with a good training, then the standard HEP methods exhibit similar performance as alternative techniques used in other fields of science.

The most complex events at LHC come from heavy ion collisions. The question of the parton radiation and knowledge of properties of the quark-gluon plasma created in those collisions may shed light on fundamentals of physics [23, 24]. A new statistical method of Markov chain Monte Carlo was proposed for extracting the viscosity and equation of state of the quark-gluon plasma [25].

The knowledge of what fraction of momentum of hadron is carried by partons within protons, which is commonly characterized by parton distribution functions (PDF), is essential
for all physics studies at hadron colliders. Many physics observables strongly depend on PDFs. Relatively large uncertainties in the knowledge of PDFs thus translate into a large uncertainty in final state observables of interest. PDFs cannot be calculated within the theory of strong interaction, quantum chromodynamics, but they need to be extracted from the data. Parametrized PDFs at a certain momentum scale can then be evolved to a different scale which is relevant for a particular measurement of interest. HERAFitter [26] is a framework which allows the extraction of PDFs from data as well as comparing theory predictions with the data. PDFs can be represented in standard LHAPDF grids [27] and griding tools like APPLgrid [28] allowing a fast use of beyond-leading order pQCD observables. HERAFitter was used in more than fifteen LHC publications.

Typical problem in LHC analyses is a reconstruction of particles from decay chain. A typical example is the decay of top quark to $b$-quark and $W$ boson which further decays to lepton and neutrino or a quark anti-quark pair. Cut based analysis or an analysis utilizing MVA techniques are often used to reconstruct the kinematics of original particle (e.g. the top quark). CMS presented a performance of a different technique called Matrix Elements method (ME). ME method establishes a direct link between theory and event reconstruction. It builds a discriminant in terms of probability that the event with a given reconstructed kinematics matches a given hypothesis about the original particle. This probability can be written as a convolution of PDFs, perturbative QCD matrix element, and a transfer function which is obtained from simulation. Advantages of the ME method are: its versatility (it can be used in many different physics analyses); no need for complicated training (such as in the case of MVA methods); maximizing the amount of theoretical information in the discrimination. The disadvantage is the fact that it can easily use only the information from leading order pQCD calculation and that it may be relatively slow.

An example of the use of MVA techniques in a similar problem of the decay of $\Upsilon (4S)$ to two $B$ mesons was reported by members of Belle II Collaboration. Hierarchical $B$ meson reconstruction takes the advantage of events in which one $B$ meson is accurately reconstructed which allows to estimate the four-momentum of the other $B$ meson without a need to explicitly reconstruct the full decay chain. This is done by combining the particles into different decays, each with its own multivariate classifier. The order of 100 classifiers can handle order of 1000 exclusive $B$ decay modes. Hierarchical reconstruction is implemented in an automated framework which needs only minimal interventions by users. The MVA techniques are in use also for the Higgs boson reconstruction. An alternative technique for the Higgs boson reconstruction in $\gamma \gamma$ decay channel was presented which formulates the Quadratic Unconstrained Binary Optimization problem in speech of a Hamiltonian and use the Adiabatic Quantum Annealing to find a optimal classifier [29] for identifying the Higgs decay.

One of important problems present in different fields of physics is the deconvolution of a signal. Generally, a signal information can be deteriorated by a presence of noise or by a finite detector resolution. There is a lot of techniques present on the market, e.g. RooUnfold package [30] where several unfolding techniques are implemented. One new promising unfolding technique was presented here. It uses a Mixture Density Model for representation of unknown true distribution and a cross-validation approach to define optimal parameters of the unfolding. Despite to the fact that method seems to be promising it was generally concluded that any new method is useful only when being followed by a publicly available implementation which allows for an independent testing and comparison with other existing techniques. An overview comparing different deconvolution methods used in particular for $\gamma$-ray spectroscopy was provided. Besides other techniques, it discussed a comparison between Gold deconvolution algorithm, Richardson-Lucy deconvolution, Maximum A Posteriori Deconvolution Algorithm, and other algorithms implemented with TSpectra class in root.
Acknowledgment: This work was supported by Charles University in Prague, projects PRVOUK P45 and UNCE 204020/2012.

References
[1] S. Agostinelli et al. GEANT4: A Simulation toolkit. *Nucl.Instrum.Meth.*, A506:250–303, 2003.
[2] http://aidasoft.web.cern.ch/USolids.
[3] http://geant.cern.ch.
[4] ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys.Lett.*, B716:1–29, 2012.
[5] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Phys.Lett.*, B716:30–61, 2012.
[6] J. de Favereau et al. DELPHES 3, A modular framework for fast simulation of a generic collider experiment. *JHEP*, 1402:057, 2014.
[7] C. Hoppner, S. Neubert, B. Ketzer, and S. Paul. A Novel Generic Framework for Track Fitting in Complex Detector Systems. *Nucl.Instrum.Meth.*, A620:518–525, 2010.
[8] R. E. Kalman. A new approach to linear filtering and prediction problems. *J. Fluids Eng.*, 82:35, 1960.
[9] R. Frühwirth and A. Stradill. Track fitting with ambiguities and noise: A study of elastic tracking and nonlinear filters. *Comput.Phys.Commun.*, 120(2-3):197–214, 1999.
[10] N. Kurita and K. Funahashi. *Neural Networks*, 9:1531–1540, 1999.
[11] http://github.com/sangaline/pidrix.
[12] S. Riggi, V. Antonuccio-Delogu, M. Bandieramonte, U. Becciani, A. Costa, et al. Muon tomography imaging algorithms for nuclear threat detection inside large volume containers with the Muon Portal detector. *Nucl.Instrum.Meth.*, A728:59–68, 2013.
[13] http://root.cern.ch.
[14] http://github.com/vgvassilev/clad.
[15] http://root.cern.ch/root/html/ROOFIT_Index.html.
[16] http://root.cern.ch/drupal/content/roofit.
[17] https://cdsweb.cern.ch/record/1456844.
[18] https://twiki.cern.ch/twiki/bin/view/RooStats/WebHome.
[19] http://root.cern.ch/root/html/RooWorkspace.html.
[20] M. Baak, G.J. Besjes, D. Cote, A. Koutsman, J. Lorenz, et al. HistFitter software framework for statistical data analysis. arXiv:1410.1280.
[21] http://cern.ch/histfitter.
[22] http://tmva.sourceforge.net.
[23] Edward V. Shuryak. Quark-Gluon Plasma and Hadronic Production of Leptons, Photons and Psions. *Phys.Lett.*, B78:150, 1978.
[24] Martin Spousta. Jet Quenching at LHC. *Mod.Phys.Lett.*, A28:1330017, 2013.
[25] Scott Pratt, Claudia Ratti, and William Patrick McCormack. Chemical properties of super-hadronic matter created in relativistic heavy ion collisions. arXiv:1409.2164.
[26] P. Belov et al. Parton distribution functions at LO, NLO and NNLO with correlated uncertainties between orders. arXiv:1410.4294.
[27] J. Butterworth, G. DiSertori, S. Dittmaier, D. de Florian, N. Glover, et al. Les Houches 2013: Physics at TeV Colliders: Standard Model Working Group Report. arXiv:1405.1067.
[28] Tancredi Carli, Dan Clements, Amanda Cooper-Sarkar, Claire Gwenlan, Gavin P. Salam, et al. A posteriori inclusion of parton density functions in NLO QCD final-state calculations at hadron colliders: The APPLGRID Project. *Eur.Phys.J.*, C66:503–524, 2010.
[29] K Pudenz and D Lidar. Quantum adiabatic machine learning. *Quant. Inf. Proc.*, 12:2027–2070, 2013.
[30] Andreas Hocker and Vakhtang Kartvelishvili. SVD approach to data unfolding. *Nucl.Instrum.Meth.*, A372:469–481, 1996.