23w5096: Systematic Effects and Nuisance Parameters in Particle Physics Data Analyses

Olaf Behnke (DESY),
Sara Algeri (University of Minnesota),
Lydia Brenner (Nikhef),
Richard Lockhart (Simon Fraser University),
Louis Lyons (Imperial College London and Oxford University)

April 23-28, 2023

1 Introduction to the topic of the workshop

This meeting dealt with systematic uncertainties in Particle Physics analyses. Such analyses are affected by statistical as well as systematic uncertainties. The former arise either from the limited precision of the apparatus and/or observer in making measurements, or from the random fluctuations (usually Poissonian) in counted events. They can be detected by the fact that, if the experiment is repeated several times, the measured physical quantity will vary.

Systematic effects, however, can cause the result to be shifted from its true value, but in a way that does not necessarily change from measurement to measurement. Measurements nearly always have some bias, and the question is by how much are they biased. Results are corrected for known biases, and the uncertainties in these corrections contribute to the total uncertainties. Systematic effects are not easy to detect, and in general much more effort is needed to identify them and to evaluate the corresponding uncertainties. The most worrying sources of systematics are the 'unknown unknowns'.

Typical sources of systematic effects include:
1: Estimated energies of jets of particles.
2: Identification efficiency of electrons.
3: Contamination by the background in selected event sample.
4: Estimates of various theoretical correction factors, and the role of theory in interpreting the data.
5: Total number of colliding beam particles for the selected events.
Such effects are usually parametrised by nuisance parameters.

After identifying possible systematic effects, the next tasks are to estimate their magnitude, and then to incorporate them into the analyses. This meeting focussed on the last aspect. It was considered that this was the topic that would most benefit from interactions between Statisticians and Physicists.

2 Participants and workshop elements

Some of the organizers of the present workshop are also involved in managing the PHYSTAT [1] event series on statistical issues and methods in particle physics data analyses. A key element of PHYSTAT workshops and seminars is the active involvement of statisticians and we are happy that at the present workshop 13 statisticians attended on-site together with 29 physicists. Regarding *equity, diversity and inclusion* we provide here some benchmark numbers: 13 on-site participants registered themselves as women and 29 as men (compared to the overall average of $\sim 20\%$ of particle physicists being female); 14 as early career researchers, ranging from PhD student to post-doc level, and 28 as seniors. In addition to the 42 on-site participants, 28 persons registered for joining via ZOOM from remote.

A workshop format was chosen for maximizing the amount of discussion. For this purpose, the workshop was structured into $\sim$ 25 topical sessions, listed in Table 1, each consisting of an introductory talk of 25 minutes followed by a 35 minutes discussion session. The participants quickly adapted to this scheme and, to the satisfaction of the audience, the allocated time for discussion was fully utilised. Indeed the discussions continued into the coffee and meal breaks, evenings and excursions. For a few sessions, the introductory talks were shared between physicists and statisticians, complementing each other. The workshop was transmitted in ZOOM, and Slack channels were used for intense continued discussions.

Session title	Speaker(s)	Chair
Introductory summaries of virtual Phystat-Systematics 2021 workshop [2]		
Physicist's view	N. Wardle	O. Behnke
Statistician's view	S. Algeri	O. Behnke
Macro Theme 1: Frequentist versus Bayesian		
Marginalizing versus profiling	R. Cousins, A. Davison	G. Cowan
Bayesian approaches in Astrophysics	F. Capel	W. Rolke
Pragmatic versus full Likelihood approaches	D. van Dyk	A. Brazzale
Likelihood-free frequentist inference	A. B. Lee	L. Heinrich
Simulation-based inference of atmospheric Cosmic-ray showers	A. Shen	L. Heinrich
Macro Theme 2: Modeling uncertainties		
Model selection	C. Schafer	N. Wardle
Background model building	L. Kania	N. Wardle
Background and signal shapes at the LHC	N. Morange	S. Williams
Template morphing	L. Brenner	H. Gray
Optimal Transport	P. Winischhofer, T. Manole	R. Lockhart
Systematics in Monte Carlo Simulation	G. Jones	T. Junk
Theory uncertainties	F. Tackmann	I. Volobouev
Macro Theme 3: Nuisance parameters		
Asymmetric uncertainties	R. Barlow	I. Volobouev
Error on error	E. Canonero	N. Wardle
Macro Theme 4: Machine Learning		
Machine Learning	M. Kagan	L. Heinrich
ML for reducing systematic uncertainties	T. Dorigo	H. Gray
Systematics in ML model independent searches	G. Grosso	I. Ochoa
Semi-supervised classifiers	P. Chakravarti	P. Windischhofer
Special experiment specific uncertainties, inverse problems and Banff Challenge		
Systematics in Neutrino physics analyses	E. T. Atkins	R. Cousins
Systematics in selected flavour physics topics	S. Stefkova	K. Tackmann
Systematics in unfolding	M. Stanley, R. Zhu	G. Cowan
Banff Challenge 3	T. Junk	P.-L. Tan
Thoughts on meeting		
Physicist's view	A. David	L. Lyons
Statistician's view	M. Kuusela	L. Lyons

Table 1: Topical sessions, speakers and session chairs.

Most of the topical sessions were distributed over four ”Macro themes”, listed in Table 1, comprising the topics of frequentist versus Bayesian inference, model uncertainties, nuisance parameters and machine learning. Another session covered topics such as experiment-specific uncertainties and inverse problems (aka unfolding) and one of the participants provided a practical data analysis challenge. The program started with two talks summarizing the outcome of the virtual PHYSTAT-Systematics 2021 workshop [2], which, among other things, served as a preparatory meeting for the present event and gave examples of problems that still need a solution and could be discussed during the meeting. A session with thoughts on the outcome of the current meeting concluded the workshop.

3 Review of the Meeting

This section contains a brief description of the individual talks; the two ‘Thoughts on the meeting’; and a list of cooperative topics as a result of the Banff meeting.

3.1 Presentations

Below are summaries of each of the talks. Each was followed by an extensive discussion period.

PHYSTAT-Systematics 2021: Physicist’s review Nicholas Wardle, Imperial

In this talk, I summarized the recent PHYSTAT-Systematics workshop held virtually in 2021 [2]. The presentation focused on reviewing the three major categories of sources of systematic uncertainties that arise in experimental particle physics; Those that arise from in-house calibrations (largely statistical in nature), those that arise due to model assumptions or using imperfect analysis methods and those that arise due to imprecise theoretical calculations which are extremely difficult to interpret as statistical uncertainties.

The talk discussed the two main methods for including these uncertainties in experimental analyses,

1. Error propagation: By changing some part of the model “one parameter at a time” (OPAT) and recording the change in a measured parameter, the uncertainty is propagated to the final result. Similarly, one can randomly sample from the systematic variations and interpret the “spread” of results as the uncertainty.

2. Nuisance parameters: Representing systematic uncertainties as parameters of the statistical model (potentially with constraints or priors) allows us to include their effect in the final result and often has the added bonus that profiling/marginalising can constrain these parameters further using the data.

Both methods have advantages and disadvantages which are more or less relevant depending on the type of source of systematic uncertainties. Several examples of experimental measurements that made use of methods 1 and 2 above and the various pitfalls and issues they faced were discussed. A list of discussion items or “open questions” to statisticians was provided:

- When reporting uncertainties on more than one parameter of interest via OPAT, is providing covariance enough?

- When modelling systematic uncertainties using nuisance parameters, should we sample many different parameter values to build suitable parameterisation and are there smarter ways (e.g. Gaussian Processes/Machine Learning?) to automate this?

- Are we OK that our fit updates our knowledge of certainty nuisance parameters?

- Is there a better way to include uncertainties due to model choice than taking the difference between (e.g.) simulations? Or approaches such as inflating uncertainty to cover potential bias / discrete profile method [3]?

These were mirrored in a similar summary talk given by statistician Sara Algeri and much of the discussion focused on these points.

PHYSTAT-Systematics 2021: Statistician’s review Sara Algeri, UMN

In my talk, I provided a statistical overview of the types of systematic uncertainties that we may encounter in physics data analyses, and how they differ from statistical uncertainties. I also highlighted some open questions and possible points of discussion, which were either raised during our Phystat-Systematics workshop in November 2021 [2] or triggered by some of the talks presented at that meeting. They can be summarized as follows:

- In the context of background mismodelling, can the difference between methods be used to acquire a measure of systematic bias?

- When dealing with nuisance parameters, is it at all possible to reach a consensus on what to do and when? (e.g., marginalizing or profiling)
- Can statisticians effectively access published likelihoods?
- How can we check the validity of regularity conditions required by classical statistics when dealing with complex models?
- What is needed to robustly bridge the gap between the statistics and physics communities?

Marginalising vs Profiling of Nuisance Parameters (Physicist's View) Robert D. Cousins, UCLA

I first spoke generally about the possibilities of eliminating nuisance parameters by using profile likelihood functions (the natural way in the context of a frequentist likelihood-based analyses) and Bayesian-inspired integrated likelihoods, advocated by Berger et al. [4], even when treating the parameter of interest by frequentist techniques.

One must distinguish between two stages in an analysis namely 1) in constructing a test statistic, typically a likelihood ratio, and 2) in using Monte Carlo simulation to generate synthetic data (pseudo-experiments) to obtain the distribution of the test statistic under the null or alternative hypotheses. I focused on the latter. In the fully frequentist parametric bootstrap, one generates the pseudo-experiments by using the profiled values of the nuisance parameters. This is the most commonly used method at the LHC.

I then discussed the simple example of the ratio of Poisson means, which appears in HEP and gamma-ray astronomy [5]. Integrating out the nuisance parameter gives the same numerical p-value as the standard "exact" frequentist solution that conditions on the ancillary statistic (total number of events), while the parametric bootstrap gives a smaller, more extreme p-value. But the discrete nature of the observables means that the exact method is conservative while the parametric bootstrap has excellent coverage. I ended by urging more studies.

Marginalise or profile: A statistician's view Anthony Davison, EPFL

Statistical inferences should be relevant, calibrated and secure. Relevance means that statements of uncertainty that are based on frequentist comparison with a reference set should restrict the latter to be appropriate to the data actually observed, by conditioning on the information content of the latter. Calibrated means that statements of confidence should be accurate, and secure means that they should not be too dependent on secondary details of the problem formulation. Nuisance parameters and discreteness can lead to poorly calibrated statements. In exponential families, the former can often be dealt with by suitable conditioning, and more general models can be approximated locally using the tangent exponential model (TEM) due to Fraser and Reid. Profiling out nuisance parameters typically leads to poor calibration for a scalar interest parameter, but the resulting error can be reduced using TEM approximation or parametric bootstrap simulation. Marginalisation is often associated with the use of a prior density on the nuisance parameters, but it can be regarded as resulting in a Bayesian inference in which a flat prior is placed on interest parameters, in which case the resulting inference may be close to the result of the TEM approximation, though a poorly-chosen prior may lead to a poorly-calibrated inference. Discreteness of responses leads to conservative inferences (confidence intervals are too long on average, leading to coverage exceeding its nominal level).

The subsequent discussion touched on a range of points, including the discreteness, the effect of the prior on nuisance parameters, and the desirability of improved approximations using the TEM. It was very helpful to talk to Bob Cousins.

Bayesian Approaches in Astrophysics Francesca Capel, MPI Munich and TUM

My talk gave an overview of Bayesian approaches to handling systematic uncertainty in astrophysics, which have become increasingly widespread over the past 30 years. Generally speaking, the Bayesian approach to systematics is conceptually straightforward, as systematic uncertainties can be parametrised and handled in the same way as statistical uncertainties. Marginalisation can be used to summarise results in terms of the parameters of interest. However, doing so requires introducing extra free parameters and a choice of priors, and results derived in this way do not guarantee coverage in the frequentist sense. I highlighted a few applications of Bayesian analysis to different types of systematics; those which can be constrained, such as in the case of the background parameter in “on/off” signal measurements of gamma-ray data [6], those which involve uncertain modelling assumptions, such as the discrete models for the Galactic magnetic field in ultra-high-energy cosmic ray studies [7], and those with uncertain theoretical frameworks, such as the shape of the binary black hole mass distribution inferred from gravitational wave data [8]. We discussed that the correct strategy would depend on the details of the problem. Still, workflows including validation against simulated data, model checking through, e.g. posterior predictive checks, or model averaging could be used to reveal and incorporate systematics into a Bayesian model.

Pragmatic vs full likelihood approaches David van Dyk, Imperial

Considering systematic errors that arise from uncertainty in quantities measured in a preliminary experiment that are treated as inputs or nuisance parameters in a primary statistical analysis, we formulate the coherent analysis of systematics in high-energy physics as a multi-stage statistical analysis. Such analyses combine multiple models and data sources into a single omnibus analysis, where outputs from one analysis are the inputs of subsequent analyses. In/Outputs may be high-dimensional with difficult-to-quantify correlations. Researchers involved in such a chain of analyses may work quite independently with different research groups having different areas of expertise and different levels of statistical sophistication. We consider different Bayesian approaches to multi-stage statistical analyses and how they can be applied to handle systematics in high-energy physics. A naïve approach ignores uncertainty in the nuisance parameters, whereas in a fully Bayesian analysis, all data (from both the preliminary and primary measurements) is used to estimate and compute errors for all unknown quantities (including nuisance parameters). In a pragmatic Bayesian analysis, on the other hand, only data from the preliminary analysis is used to estimate and compute errors for the nuisance parameters. We discuss the pros and cons of these approaches using several examples from high-energy astrophysics, including instrument calibration [9, 10] and the disentangling of overlapping sources [11, 12].

Subsequent discussion focused, for example, on the relationship of the pragmatic Bayesian approach to methods for “cutting feedback” or “modularization”, i.e. informally limiting the influence of potential misspecification of parts of a multicomponent Bayesian model (see, for example, [13] and references therein). We thank Tom Loredo for pointing out the relevant literature. This work has already led to a publication [14].

Likelihood-Free Frequentist Inference Ann B Lee, CMU

Many areas of science, such as high-energy physics, make extensive use of computer simulators that implicitly encode the likelihood functions of complex systems. Classical statistical methods are poorly suited for these so-called likelihood-free inference (LFI) settings, outside the asymptotic and low-dimensional regimes. Although new machine learning methods, such as normalizing flows, have revolutionized the sample efficiency and capacity of LFI methods, it remains an open question whether they produce confidence sets with correct conditional coverage. In my talk, I summarized recent developments in unifying classical statistics and machine learning methods for (i) constructing Neyman confidence sets with finite-sample guarantees of nominal coverage in the presence of systematic effects, and for (ii) computing diagnostics that estimate conditional coverage over the entire parameter space.

This workshop gave me an excellent overview of the relevance and main challenges of systematic effects in HEP. The workshop also gave me an invaluable opportunity to interact closely with particle physicists for a week. These interactions have, for example, led to a new collaboration with physicist Tommaso Dorigo on optimizing detector design for atmospheric cosmic-ray showers. I am also following up on a connection between calibrated forecasting and optimal transport that I was not aware of until I started discussions at Banff with physicists Michael Kagan, Lukas Heinrich and Philipp Windischhofer. Many congratulations to the Banff organizers for running such a successful and high-quality workshop with both statisticians and physicists as participants!

Likelihood Free Frequentist Inference for Cosmic Ray Reconstruction Alex Shen, CMU

In my talk, I discussed an upcoming project that uses Likelihood Free Frequentist Inference (LF2I) to perform the reconstruction of cosmic rays from secondary footprint data with uncertainty quantification. The LF2I framework is outlined in [15].

Key Points

- Reconstruction is broken down into two steps: determining the identity of the primary cosmic ray (proton or photon), and estimation of key parameters (energy, orientation) for photonic cosmic rays only.
- The first step is complicated by the fact that the prior distributions of cosmic ray identity and energy are not independent. We take a nuisance parameterized LF2I hypothesis testing approach that ensures Type I error control no matter the energy of the primary cosmic ray.
- The second step uses hypothesis test inversion under the LF2I framework to construct confidence sets for parameters of interest for photonic cosmic rays.

Thoughts on model selection Chad Schafer, CMU

The general problem of background model selection in particle physics experiments presents some interesting statistical challenges, chiefly surrounding the issue of the tradeoff between using (1) a sufficiently realistic and flexible background and (2) a model that imposes sufficient structure that there is the power to detect the signal of interest. The *semiparametric* approach to inference seems to be a natural fit in this context. In the classic semiparametric setup, the parameter of interest (in this context, a quantification of the signal strength) is real-valued, while the nonparametric component (i.e, the background) is a nuisance parameter. It could be argued that most particle physics experiments implement a *de facto* semiparametric approach, in that even though parametric background models are utilized, their complexity is increased sufficiently to model the background to achieve an acceptable level of fidelity to the observed data. A well-implemented nonparametric background estimator would behave in a similar, but automated manner. [16] present theory regarding the performance of semiparametric estimators that could be useful in this case. In particular, the authors define versions of the *score function* and *Fisher information* adapted to the case where the nuisance parameter is estimated nonparametrically. The ideas are motivated by considering the *least favorable* background model, i.e., the model that makes the estimation of the parameter of interest the most challenging. If used in conjunction with the *method of sieves* [17] to constrain the space of background models, the result could be consistent estimation of, and more accurate and meaningful measures of uncertainty for, the signal strength.

Uncertainty quantification via influence functions Lucas Kania, CMU

A fundamental problem in High Energy Physics is separating collisions corresponding to known physical phenomena (called background) from those indicating new phenomena (called signal). This is challenging because the signal (if it exists) represents a small fraction of the available data. Current statistical approaches model the data as a mixture of the background and signal, and either (A) require access to an auxiliary sample from the background to estimate it non-parametrically and then attempt signal detection without a signal model [18] or (B) assume parametric background and signal models, and then try to find a signal via the profile likelihood ratio [19].

In this talk, we considered experiments where there is no auxiliary sample, but there is a known signal region. That is, if a signal exists, it must be contained in the signal region. Under this assumption, specific non-parametric background models can be identified, e.g. high-degree polynomials, and their influence functions [20] can be harnessed to perform valid signal detection. In summary, the method allows physicists to perform valid signal detection using complex background models without any auxiliary dataset.

In the discussion that ensued, Robert Cousins recommended analytically comparing our results to the standard profile likelihood machinery [21]. Additionally, Sara Algeri encouraged us to prove non-asymptotic guarantees for our methodology, which relies heavily on asymptotics.

This talk was based on joint work with Larry Wasserman, Mikael Kuusela and Olaf Behnke.

Background and signal shapes Nicolas Morange, Orsay, LAL

This presentation showed why getting accurate signal and background shapes is a major topic of analyses at the LHC, and discussed the main strategies used to increase the precision of the measurements by reducing the systematic uncertainties on the shapes.

With the vast amount of LHC data available, statistical uncertainties in the analyses are getting smaller and smaller. This in turn allows for precision calibration of the detection efficiencies and energy measurements.

In the quest for high-precision measurements of physics processes, the uncertainties in the background and signal shapes therefore become more and more relevant.

Getting accurate shapes and reducing the associated uncertainties relies on two main pillars. The first one is using better theoretical predictions and more accurate Monte Carlo samples, taking higher-order effects into account. Discussions highlighted the huge computing effort required to get enough statistics of accurate, modern Monte Carlo samples, and sketched ideas such as multi-dimensional reweighting to alleviate the issue. The second pillar consists in using the data as much as possible to calibrate the Monte Carlo in dedicated signal-depleted regions, or completely replace the Monte Carlo whenever possible, using fully data-driven techniques. Different ad hoc solutions to define systematic uncertainties are used depending on the analyses, and great care is taken to avoid biases and setting too small uncertainties.

Progress on this topic necessitates close collaboration between experimentalists, theorists and statisticians.

Template morphing Lydia Brenner, Nikhef

This talk describes and compares different interpolation/morphing methods that produce signal or signal+background models for fits to data. It is also described in which case different methods might break down. A relatively new method, effective Lagrangian morphing, is described in more detail [22]. The signal model is morphed in a continuous manner through the available multi-dimensional parameter space. An example is given of the calculation leading to the morphing function as well as an example use-case where the morphing function is applied to simulated data of a Higgs boson. The approach described in this talk is expected to substantially enhance the sensitivity to Beyond the Standard Model contributions. Among the considerations discussed are computational costs and uncertainty propagation of systematics.

Optimal transport in high-energy physics Tudor Manole, CMU and Philipp Windischhofer, Uni Chicago

Optimal transport is a branch of mathematics that has recently emerged as a methodological tool for statistics and machine learning. In our survey-style talk, we covered several aspects of data analysis in particle physics where optimal transport could extend, or replace, existing methods.

Optimal transport provides a principled way of defining mappings between probability distributions, which can be leveraged to construct unbinned calibrations of Monte Carlo simulations to data [23], perform data-driven background modelling [18], or build interpolated densities in the context of template morphing. Optimal transport also provides guidance in defining a distance metric on the space of distributions, known as the Wasserstein distance, lending a geometric interpretation to operations such as jet clustering [24]. In the realm of statistics, optimal transport can be used to generalize the notion of a quantile to multivariate distributions, in turn defining a swath of new goodness-of-fit and two-sample tests for multivariate distributions [25].

Recent advances in machine learning have made it possible to solve important classes of optimal-transport problems on large-scale data sets [26], thereby providing the technical framework to study the performance of these methods in real-world analysis settings, a process that has already begun within the experimental collaborations at the Large Hadron Collider.

Systematics in MC Galin Jones, UMN

My talk focused on systematic errors in Monte Carlo, especially in the context of systematics in high energy physics data analyses. The main points of discussion included (i) the effect of model error on Monte Carlo, (ii) the effect of nuisance parameters in Monte Carlo, (iii) Monte Carlo simulations producing erroneous output, (iv) determining Monte Carlo sample size, and (v) the effect of using the Monte Carlo output suboptimally. Some of the specific issues presented included assessing the Monte Carlo error to determine the simulation length, the construction of confidence intervals with convenient marginal interpretations and approaches to maximum likelihood-based inference for simulated likelihoods.

The session prompted substantial follow-up discussion, including after the conclusion of the workshop. Some of the discussion was prompted by concerns about the reproducibility of high-energy physics analyses. Other discussions concerned resolving Monte Carlo errors in neutrino analyses.

Theory uncertainties Frank Tackmann, DESY

Given a measurement of some quantity f, we obtain the parameter of interest p from it by comparing to the expected value $f(p)$ as a function of p. The formula we use for $f(p)$ is the *theory prediction*. The *theory uncertainty* is due to the fact that the formula itself is never really exact but derived in some approximation. To account for the theory uncertainty, we typically quote the prediction with an uncertainty estimate $f(p) \pm \Delta f$. In my talk, I discuss the current prevalent method in high-energy physics for how to estimate Δf, namely scale variations. In perturbative predictions, $f \equiv f(x)$ depends on a small quantity x (the coupling constants), in which we can expand $f(x) = f_0 + f_1 x + f_2 x^2 + ...$ Our approximate prediction comes from truncating this expansion, e.g. at the next-to-leading order (NLO) we would keep the first two terms and drop the quadratic and higher terms. The inexactness and thus theory uncertainty is then due to the dropped terms $f_2 x^2 + ...$ Scale variations essentially amount to performing a variable transformation $\tilde{x}$ such that $x = x(\tilde{x}) = \tilde{x} + b_0 \tilde{x}^2 + ...$, performing the expansion for f in terms of $\tilde{x}$, and using the difference between the expansions in x and $\tilde{x}$ as an estimate of Δf. While this method is convenient, it has many limitations, which I discuss. Another important issue is that of correlations between different predictions. Say we have two predictions $f(x) = f_0 + f_1 x \pm \Delta f$ and $g(x) = g_0 + g_1 x \pm \Delta g$ and we are interested in their ratio $f(x)/g(x)$. We expect or hope that some of the uncertainties would cancel in their ratio, but to quantify this we need to know the precise correlation between Δf and Δg. The perhaps most severe shortcoming of the scale-variation method is that it cannot tell us anything about correlations. The best we can currently do is after estimating Δf and Δg to make an educated guess on how to correlate them.

In the last part of the talk, I discuss the development of a new concept of theory nuisance parameters [27], which promises to overcome the limitations of scale variations. The basic idea is to consider the actual source of uncertainty, i.e. the missing coefficient f_2 in the above example, and parametrize it as $f_2 \equiv f_2(\theta)$, where the *theory nuisance parameters* θ are genuine nuisance parameters that have a true but unknown value. In the simplest case when f_2 is just a number, we could use it itself as the nuisance parameter, $f_2(\theta) = \theta$. This concept turns the theory uncertainty into a truly parametric uncertainty, and much of the following discussion was related to all the resulting advantages. For example, the theory of nuisance parameters could be constrained by auxiliary measurements. In the absence of such constraints, it is also possible to put a constraint on their size from theoretical considerations on the typically expected size of terms in perturbation theory.

Asymmetric Errors Roger Barlow, Huddersfield

In this talk, I presented the development of ideas about the handling of asymmetric errors in particle physics, which are quoted in many results and which are certainly not handled correctly. Preparation for the presentation forced me to develop my original ideas (which had been published only as preprints) and it became clear that the two classes of asymmetric error originally denoted 'systematic' and 'statistical' are, at a more fundamental level, asymmetries in the pdf and in the likelihood. Thus an apparently arbitrary taxonomy has been replaced by a logical division.

That this analysis was accepted by the audience was, for me, very encouraging, and a major paper on the subject is now in preparation. In the discussion session there were helpful suggestions about the use of the Edgeworth expansion and also Azzalini's skew Normal distribution, which I am incorporating in the treatment, in addition to the original two models I had considered.

I also received an offer of assistance with writing the software package(s) that will be needed to get this treatment adopted by practitioners in particle physics experiments. This working collaboration is very encouraging and helps move the project along. The paper and the package will produce a real impact on the treatment of this problem in the community.

Error on Error Enzo Canonero, RHUL

The Gamma Variance Model (GVM) is a statistical model that implements uncertainties in the assignment of systematic errors (informally called *errors-on-errors*), a pressing issue for many LHC analyses, currently or soon-to-be dominated by systematic uncertainties. The GVM treats discrepancies in input data as an indication of additional uncertainty, hence making a negligible impact on results when data are consistent. Conversely, when data are inconsistent, *errors-on-errors* significantly affect the results, thus influencing our final understanding of the parameters we aim to measure, both in terms of their central value and associated confidence intervals. To showcase this property of the model, an investigation of the tension between the latest

CDF W-mass measurement, the Standard Model prediction, and the ATLAS measurement was presented, emphasizing how uncertainties in systematic error assignment can substantially affect our understanding of the W-mass.

When *errors-on-errors* are small, one can use standard asymptotic methods to construct confidence intervals for the parameters of interest and compute p-values. This is analogous to the familiar scenario with a large data sample, where all adjustable parameter estimators follow Gaussian distributions. However, if *errors-on-errors* are not small, first-order asymptotic distributions are not an accurate approximation. During the talk, we presented advanced test statistics based on higher-order asymptotics (r^* approximation and Bartlett correction), which are corrections to the first-order statistics so that their asymptotic distributions are a better approximation of the truth even in scenarios with small sample sizes, or equivalently large *errors-on-errors*.

The discussion after the talk focused on how *errors-on-errors* could be assigned in real applications to uncertain systematics, paying particular attention to theory systematics and two-point systematics. The impact of *errors-on-errors* on the W-mass application was also discussed, given the significant and non-trivial effects observed. A comparison was made between the GVM method of combining incompatible measurements and the approach of the Particle Data Group (PDG), along with potential Bayesian approaches to the problem.

Systematics in Neutrino physics data analyses Edward Thomas Atkin, Imperial

Systematic uncertainties in long-baseline neutrino experiments come from the description of the neutrino beam, the near and far detector responses and the modelling of neutrino-nucleus interactions. In particular, systematic uncertainties related to neutrino-nucleus interactions are very important as they change the reconstructed energy spectra of neutrino events which directly impacts the measurement of neutrino oscillation parameters. Unfortunately, neutrino interactions can still have large, often ill-formed, and ad-hoc uncertainties. The near detector is essential in reducing systematic uncertainties and measures the unoscillated neutrino beam. The far detector then measures neutrinos after they have oscillated and gives sensitivity to neutrino oscillation parameters. However, due to limited statistics, the far detector has very little power to constrain systematic uncertainties. Ensuring that the near detector correctly constrains systematic uncertainties is one of the main challenges of long-baseline experiments. If the constraint from the near detector is biased then measurements on oscillation parameters at the far detector will also be biased. This is often a challenge due to near and far detectors being non-identical detectors that span different neutrino fluxes which can also have quite different acceptances.

There were discussions on how neutrino interactions impact the analysis and the checks on the bias that a poor near detector constraint can have on the final contours. In addition, useful advice on how to treat systematics which migrate events between analysis bins was given.

I hope to continue working with Galin Jones on studying some of the statistical problems encountered by neutrino experiments, especially those related to the Markov Chain Monte Carlo analysis which T2K and other experiments use.

Machine Learning Michael Kagan, SLAC

As machine learning (ML) is used more and more in high energy physics (HEP) data analysis, a question that frequently arises is “when do we need uncertainties on the ML model?”. Before answering this question, one must first ask what uncertainties do we mean?

Within the topic of uncertainty quantification (UQ) in ML, uncertainties are typically broken into two categories: aleatoric uncertainties and epistemic uncertainties. Aleatoric uncertainties, often called statistical uncertainties, arise from the inherent randomness in a system, e.g. in stochastic systems, or noisy detectors. For example, a form of aleatoric uncertainty is detector resolution; the same incoming particle can give rise to a distribution of observed measurements. This uncertainty is often described as “irreducible” because no number of measurements will reduce these inherent stochastic effects. As aleatoric uncertainty arises from statistical sources, it is often described using probability distributions. Epistemic uncertainty, often called “model uncertainty”, arises from a lack of knowledge about the system, e.g. from missing measurements or limited amounts of data used for fitting a model to data. These uncertainties are often considered reducible, as more data can often help constrain the set of models one would fit to a given data set. While the above-mentioned uncertainty classes assume a training set that is drawn from the same distribution as the data in which the model will be applied, when such distributions differ, uncertainty can arise due to such a domain/distribution shift. These distribution shift uncertainties are akin to Systematic uncertainties in HEP.

Methods to model both Aleatoric and Epistemic uncertainty can be found in the study of UQ in ML. Aleatoric uncertainty, being of a statistical nature, is often described using neural parameterized probability distribution, such as mixture density networks, or generative models such as normalizing flows. Epistemic uncertainty, not being of a statistical nature, often attempts to describe what set of parameters, such as weights in a neural network, could have been fit from a given training set. Attempts to model this include ensembling techniques and Bayesian techniques that attempt to estimate posteriors on weight distributions. As such posteriors can't be calculated analytically, a host of approximation methods have been developed.

When would such epistemic / model uncertainties be needed? In most cases in HEP, such as reconstruction or signal/background classification, the aim of the ML model is to define a good summary statistic for downstream parameter inference. This choice of summary statistic will determine the power of the statistical test, i.e. it will determine the optimality of the statistical test but will not make the model of the data incorrect. In these cases, epistemic uncertainty is likely not needed. However, if an ML model affects the prediction of the rates of process or shapes of distributions given a test statistic, as in background estimation of ML-based simulations, poor predictions will affect the compatibility of the model with real data. As such, epistemic uncertainty estimation is likely needed, or the development of procedures to estimate systematic uncertainties using control data. Ultimately, this question of optimality versus correctness should be addressed for each application of ML in HEP in order to determine if an associated model uncertainty should be estimated.

Using ML to reduce Systematics Tommaso Dorigo, INFN and Uni Padova

A variety of machine learning-powered techniques have been developed over the course of the past decade with the purpose of incorporating the effect of systematic uncertainties in the extraction of statistical inference on parameters of interest. This is particularly useful in multi-dimensional problems, where direct parametrizations are not feasible. Dorigo's presentation briefly touched on the various ideas that have been used to develop models that include a correct treatment of nuisances in typical HEP problems. For a description of these techniques see a recent publication [28].

The realization that the incorporation of nuisance parameters in inference extraction is a holistic optimization problem, where consideration of nuisance-related effects corresponds to a realignment of the statistical procedures with the final goal of the measurement, and the parallel observation that machine learning today offers powerful solutions to such end-to-end optimization tasks, has recently led Dorigo to assemble a group of physicists and computer scientists in order to consider the possibility of modeling together in a single problem not only all aspects of statistical inference, but also the data collection and reduction procedures, and ultimately the data generation procedures. This allows to encode in a global pipeline the full design of an experiment, from the choice of materials and geometry of the detector layout to the pattern recognition and analysis procedures. Such an ambitious goal can today be achieved for systems of moderate complexity through the use of backpropagation of derivatives of a global utility function in a full model of the experiment. The utility must specify in a the faithful way the relative importance of different goals of the experimental endeavour, not dissimilarly to what is done when defining the bandwidth allocated to different triggers in a collider experiment. The MODE collaboration [29], which includes 40 computer scientists and physicists from 24 institutes in Europe, Asia and North America, has thus started a program of investigation to attack simpler end-to-end optimization tasks, to build expertise and a library of solutions and methods to tackle progressively harder optimization problems. A recent publication [30] describes the status of these activities.

Systematics in ML model-independent searches Gaia Grosso, CERN and Uni Padova

In my talk, I presented New Physics Learning Machine (NPLM), a novel machine-learning-based strategy to detect and quantify data departures from a Reference model with no prior bias on the source of discrepancy [31, 32]. The main idea behind the method is to approximate the log-likelihood-ratio hypothesis test parametrising the data distribution with a universal approximating function, and solving its maximum-likelihood fit as a machine-learning problem with a customised loss function. The method returns a p-value, which measures the compatibility of the data with the Reference model. NPLM has been recently extended in order to deal with the uncertainties affecting the Reference model predictions [33]. The new formulation directly builds on the specific maximum-likelihood-ratio treatment of uncertainties as nuisance parameters, that is routinely employed in high-energy physics for hypothesis testing. The main goal of this presentation was to convey the key steps needed to correctly set up the strategy and validate its robustness in the presence of systematic uncertainties. Some theoretical aspects to be further investigated were pointed out in the conclusions, and a rich discussion emerged from them. In particular, Alessandra Brazzale pointed out that our construction of the family of alternative hypotheses by means of universal approximators (like neural networks) with exponential parametrization could be seen as a generalization of the Exponential family. She also noticed possible connections between our work and the research on nonparametric extensions of the likelihood-ratio-test (see for instance [34]). I am studying the literature and I plan to further discuss this topic with Alessandra. After the presentation, we also discussed the possibility of estimating an error for the outcome of the test by running the algorithm multiple times with different random initializations of the model trainable parameters. We commented on the need to explore smarter regularization schemes to improve the test sensitivity. We discussed the globality of the p-value returned by the algorithm; in particular, we agreed that we should think about how to fairly compare our method with alternative approaches that quote local p-values. Finally, we started discussing similarities and differences between the NPLM approach and the approach presented in the following talk by Purvasha Chakravarti. The two groups are now interacting to work on a comparison of the methods.

Model-Independent Search using Interpretable Semi-Supervised Classifier Tests Purvasha Chakravarti, UCL

In my talk, I introduced some of the recent developments in model-independent searches using classifiers presented in [35]. The aim is to search for new signals that appear as deviations from known Standard Model physics in high-dimensional particle physics data. The key contributions and main discussion points during the talk were:

- Introduced three test statistics using the classifier: an estimated likelihood ratio test (LRT) statistic, an area under the ROC curve (AUC) based test statistic, and a misclassification error (MCE) based test statistic.

- Introduced a method for estimating the signal strength parameter and active subspace methods to interpret the classifier in order to understand the properties of the detected signal.

- Discussed the performance of the methods on a simulated data set related to the search for the Higgs boson at the Large Hadron Collider at CERN and showed that the semi-supervised tests have power competitive with the classical supervised methods for a well-specified signal, but much higher power for an unexpected signal which might be entirely missed by the supervised tests.

- Discussed that the introduced model-independent approaches are sensitive to the background reference data set and the systematic uncertainties in them. However, we can still use the methods to identify and characterize regions of high-dimensional space where the background is mismodelled and/or perform pilot analysis to guide future model-independent searches.

- Discussions included detailed comparisons with the approaches in [33], the first contribution towards dealing with systematic uncertainties in model-independent searches.

Systematics in rare B-decays Slavomira Stefkova, KIT

Rare B-decays are excellent probes of new physics (NP). These are either flavor-changing-neutral-current transitions and/or helicity-suppressed decays and therefore the Standard Model (SM) contribution is small. NP contribution in comparison, if present, could be significant. Two flavour physics experiments, LHCb and Belle II, search for NP in rare B-decays directly (dedicated NP searches) and indirectly (SM precision measurements). In this talk, systematic uncertainties and their treatment in rare B-decays were presented. Given that LHCb and Belle II make use of different collider technologies, not only measurement style but also considered systematic uncertainties are different. Systematic uncertainties are included as nuisance parameters in binned or unbinned likelihoods when building a statistical model. There are three main challenges when building statistical models for rare B-decays. Many analyses find themselves in the low-statistics regime after all the selection and asymptotic approximation may no longer be valid. In addition, the trade-off between smoothness and fit stability, which heavily depends on the included systematic uncertainties, is often an important part of the analysis. The second challenge is especially applicable to LHCb, where all the branching fraction measurements are done relative to another control channel so that the systematic uncertainties due to accelerator and acceptance effects cancel in the ratio. For rare B-decays it is sometimes challenging to find a good channel to normalise to. The final challenge is that backgrounds for rare B-decays could be themselves rare and they may have never been measured or even theoretically predicted. The difficulty with such cases is how to correctly include these rare B-decay backgrounds in the statistical model and assign the appropriate systematic uncertainties to them. Apart from these challenges, the main discussion item evolved around the possibility to make a combination between LHCb and Belle II measurements. At the moment, HFLAV collaboration makes naive combinations but in the future, this can be further refined by publishing the full likelihoods and by correlating the systematic uncertainties, wherever possible, from the two experiments.

Accounting for systematic uncertainties in unfolding uncertainty quantification Michael Stanley, CMU

This talk discussed the computation of bin-count confidence intervals for particle unfolding uncertainty quantification. We first provided the mathematical framework for unfolding and then outlined four system uncertainty sources; regularization bias, wide-bin bias, missing auxiliary variables, and response kernel uncertainty. The first two sources were discussed in some detail in which we identified the primary systematic uncertainty culprit as the Monte Carlo ansatz used in place for the true particle intensity function. As such, our solution aims to reduce the error resulting in this ansatz misspecification by first unfolding to fine bins followed by an adjacent bin aggregation to a desired wide-bin resolution. Further, we show the implementation of two optimization-based statistical methods to directly compute confidence intervals for each desired aggregated wide bin.

This work is thoroughly explored and developed in [36], which is based on the following works [37, 38, 39]. To make these methods accessible to the particle physics community, we next plan to work with Lydia Brenner toward their implementation in RooUnfold.

Systematic uncertainties in unfolding of differential measurements Richard Zhu, CMU

In my talk, I addressed the systematic uncertainties in the forward model for unfolding differential measurements.

The goal of unfolding is to recover the true particle spectrum based on the smeared detector measurement. The forward model (response kernel) models the detector response, which is the conditional probability of the smeared observations given the truth. In practice, the forward model needs to be estimated using detector simulations. The imperfect knowledge of both the detector alignment and calibration as well as the theoretical predictions can affect the forward model in different ways. This leads to systematic uncertainties in the forward model and hence raises nontrivial effects on the unfolding solutions.

To formalize the problem, recall that we have the particle-level spectrum f and detector-level spectrum g related by the Fredholm integral operator $g = \int_T k(y,x)f(x)dx$. The response kernel k inside the integral represents the conditional density of the smeared measurement given the truth, i.e. $k(y,x) = p(y|x)$. The systematic uncertainty in k means that a set of alternative kernels $k_0, k_1, ..., k_m$ might be plausible during detector simulation. The question is how should we account for this systematic uncertainty in the unfolding solution. For simplicity, suppose there are two base kernels k_0, k_1. Then we propose to use optimal transport to morph between k_0 and k_1, which defines a geodesic of kernels $\{k_t : 0 \leq t \leq 1\}$ between the two kernels. Since optimal transport preserves the shape and geometry of the kernel, the geodesic of kernels can be served as the plausible candidate kernels. Then we propagate the geodesic of kernels and unfold with the 'One-at-a-time Strict Bounds' (OSB) intervals proposed by Stanley et al. [36], which results in a collection of confidence sets for λ. The collection of confidence sets has the correct frequentist coverage if the unknown correct kernel lies on the geodesic.

Some open problems and the following discussion include: (1) How can we summarize the collection of confidence sets (dependent on $\mathbf{t}$) into a single confidence set? Can we do profile likelihood? (2) We can view the weight $\mathbf{t}$ as a nuisance parameter. Can we learn $\mathbf{t}$ from the data? (3) How do we apply the method on real HEP and how well does it work?

Banff Challenge 3 Tom Junk, FNAL

Banff Challenge 3 is an exercise designed to test the ability of participants to estimate systematic uncertainties and apply these estimates in the calculation of confidence intervals in a contained environment, without requiring detailed knowledge of elementary particle physics or the accelerators and detectors experimental physicists use. Simulated random data and randomly-sampled model predictions are provided, drawn from two-dimensional probability distributions. The underlying functional forms of the probability distributions, and thus the parameters they depend on, are hidden from challenge participants. The simulated data samples contain mixtures of signal and background data, and the simulated model samples are separated into labeled sets of signal and background data. Participants are to provide confidence intervals on the rate of signal counts in each of 100 simulated data samples. The exercise is patterned after a CDF measurement of the W boson cross section [40], although the probability distributions and parameters are completely artificial. The full problem statement and data files are available at [41].

3.2 Thoughts on the meeting session

The concluding talks were given by Physicist Andre David and Statistician Mikael Kuusela. Rather than being required to provide a comprehensive review of the whole meeting, they were asked to give us their thoughts on what they considered were some of the interesting talks and discussions.

Physicist's view Andre David, CERN

Of signals and backgrounds When discussing systematic uncertainties in HEP analyses, one is easily taken to the notion of nuisance parameters and how to deal (usually meant in the "get rid of" sense) with said nuisances. A common trap is to think of the different physics processes involved in ways that are hierarchical, like signal and background, with backgrounds being considered nuisances and signals being of interest.

This is a shortsighted view since in most cases improvements come about by not only a better (more accurate) modelling of the "signal" process but also (if not more) by a better modelling of the "background" processes. There are also cases where the sensitivity to an underlying theory parameter is "spread" among different processes that, in some cases, are not distinguishable at a fundamental, quantum-mechanical, level. A more holistic view when dealing with uncertainties on different processes is expected to become even more important as the field moves towards performing broader combined analyses and interpretation is performed in terms of effective field theories, like the SMEFT, whose parameters have wide-reaching and pervasive effects across whole classes of processes.

One should carefully distinguish and treat the following: processes sensitive to the inference one wants to make, processes that are not sensitive but in many cases may limit the power of said inference, and detector limitations, like noise and finite resolution. The more broadly a result is expected to be used, the more important these matters are, since what is my signal may be your background and vice-versa.

Alternatives and morphing One important argument made during the workshop is that one should be careful to distinguish physical deformations for which the intermediate values have a physical meaning (e.g., the effect of an energy scale uncertainty) and alternatives between differing options, the interpolation of which is devoid of a concrete physical interpretation and meaning. The important distinction (made by Cousins) is that for the latter if at any point the results of the inference end up depending on them, then there will be enough power in the data and the model to likely rule one out as unphysical.

This reasoning removes some of the pressure on how to deal with "pure alternatives" and may point the way forward in establishing procedures for working through them.

Progress in unfolding Unfolding detector-level quantities is an important step to make experimental results more readily used by the broader community. E.g., unfolding allows theorists to compare their results without having to go through the computationally intensive and error-prone task of forward-folding detector effects that are specific to each detector.

In this area, a lot of progress has been made in the recent past thanks to new statistical methods and that effort continues unabated.

My measurement or your calibration? An interesting discussion that arose at the workshop relates to the difference between calibration and measurement and how there are even separate communities in astrophysics.

The interesting point here is how to ensure that the measurement part of the chain does not overstep the stated calibration accuracy and precision and, in case it can legitimately improve the calibration, how to go about that. In the LHC collaborations the "calibrators" and "measurers" are all part of the same community and it was reported that both types of interactions have taken place. I.e., there are cases where a measurement is "blunted" to stay within credible calibration boundaries and cases where a measurement led to improvements in the calibration procedures.

This interplay between measurement and calibration ends up being slightly artificial and two sides of the same coin of what information can be exploited and whether it has been considered in the calibration process or not. In that sense, it is important in the context of systematic uncertainties to keep a two-way road between "calibrators" and "measurers".

The discrete profiling miracle Discrete profiling made a splash in the scene of HEP and after a decade it is not yet clear why it achieves the frequentist properties it seems to have. Perhaps there is hope to understand discrete profiling in the model selection context as discussed at the workshop. The attending question with model selection came as to how would one feel about model averaging being the (weighted) average of estimates across different models and physicists seemed to not be too comfortable with this idea.

There was also a very concrete discussion on the methods that CMS (discrete profiling) and ATLAS (spurious signal) have employed to tackle the lack of knowledge on the background shape and the main point is that both of them use the statistical uncertainty under the signal as their gauge. In that sense, they trace their use to the same underlying quantity and even though they then are used in different ways, can still be expected to have similar impacts.

Unleash the tails The workshop saw a very interesting discussion on how certain one is of a given systematic uncertainty by allowing for longer and heavier tails, therefore allowing for the corresponding nuisance parameters to be pulled farther than one would otherwise consider.

The technique seems to be very well suited for application to theory uncertainties, not the least because it could be interesting what experimentalists from experiment i would say about how certain they are of the systematic uncertainties of experiment k, especially when $i = k$.

"Last mile" corrections The workshop saw a large number of machine learning topics discussed and among the most interesting are those related to optimal transport in the context of fine corrections. In a typical experiment, there will be the main, general-purpose, corrections to observables that are called calibrations, and then there is a second (or third) level of corrections that are small and dedicated to particular corners of the phase-space of reconstructed properties, e.g., the electromagnetic shower shapes of low-energy photons.

"Last mile" corrections can be cast as finding (learning) awhat multidimensional function that takes MC-simulated events and matches actual data in a control region. This function is then used to "correct" MC events in the region of interest when performing inference.

Depending on the level of discrepancy, it may be that optimal transport methods presented at the workshop can become a new tool in performing this kind of "last mile" correction that will become ever more important in precision analyses.

ML for actual intelligence The rise of machine learning is reshaping the way we think about analyses and the use of information. Many applications are possible and almost all are being pursued: detector operation, construct (optimal) observables, design (optimal) detectors, model-independent methods (vs MC simulation), sample sizes, skirt systematically-affected phase spaces, etc. Now that creativity in the applications is partly exhausted, we need to become proficient at writing loss functions that allow these algorithms to explore outside the box. Only then will these algorithms become effective support systems for AI (actual intelligence, as defined by Cousins) i.e., us.

That all said, it is good to see that there is now a general consensus that a bad ML algorithm in constructing an observable will not lead to a wrong result, just a sub-optimal one. Of course, this cannot be said of all ML algorithms, especially generative ones, where inaccuracies can have dire consequences.

Statisticians view Mikael Kuusela, CMU

In my talk, I highlighted selected thoughts I had about the meeting topics and discussions from the perspective of a statistician who actively collaborates with high-energy physicists. The talk had six key messages:

1. **Beyond profiling/marginalization:** We heard a lot about profiling and marginalization for handling nuisance parameters during the meeting. But are there other ways of handling nuisance parameters? Specifically, in the setting of Gaussian linear models, we know how to get exact confidence intervals for individual model parameters despite the presence of nuisance parameters. How do we extend this to more general settings?

2. **Which confidence set to report?** High-energy physicists usually report ellipsoidal confidence sets but, in some settings, it could also make sense to report hyperrectangle confidence sets, as was illustrated in Galin Jones's talk. Similarly, many talks and discussions during the meeting raised the question of whether high-energy physicists should consider reporting simultaneous confidence intervals instead of one-at-a-time confidence intervals.

3. **Systematics in other fields:** Handling systematic uncertainties is a widespread challenge across the physical sciences. I showed an illustrative example of this from a recent intercomparison of ocean heat content estimates. The error-on-error model presented by Enzo Canonero could prove useful in these other fields.

4. **How to best learn likelihood ratios?** We heard from Gaia Grosso and Purvasha Chakravarti about model-independent searches of new physics using likelihood ratio tests where the test statistic is learned from the data using machine learning. In the case of Grosso, the alternative hypothesis is modeled using a neural network while Chakravarti trains a classifier to learn the likelihood ratio. This leads to two different ways of learning the likelihood ratio so I posed the important open question about the pros and cons of these two ways of performing the test.

5. **Model discrepancy:** Instead of using nuisance parameters, one can handle systematic uncertainties by writing down a stochastic model to capture the model discrepancy. This is a common strategy in computer model calibration. I briefly explained the popular Kennedy–O'Hagan method [42] for modeling model discrepancy as an additive Gaussian process.

6. **Interdisciplinary collaboration:** How do we sustain and expand the interdisciplinary interaction between physicists and statisticians beyond workshops like this? How do we achieve active in-depth collaborations between the two fields? I suggested co-supervision of Ph.D. students, where a physicist is a co-advisor of a statistics student and a statistician is a co-advisor of a physics student, as a potential avenue for this. This is a model that has worked successfully for us at Carnegie Mellon in several interdisciplinary projects.

3.3 Connections

The meeting benefitted from the unique atmosphere at Banff and was characterised by vigorous and productive discussions, especially between Physicists and Statisticians. From a subjective physicist's standpoint these exchanges can be roughly categorized in:

- Statisticians *agreeing* with what physicists do, e.g. the asymmetric uncertainty procedures presented by Roger Barlow.

- Statisticians *questioning* what physicists do, e.g. Sara Algeri about checking the validity of regularity conditions required by classical statistics when dealing with complex models.

- Statisticians *suggesting* physicists what to do, e.g. Anthony Davison on making use of the tangent exponential model for improved statistical inference.

- Statisticians *competing* with the physicists, e.g. on Machine learning algorithms for anomaly detection, as presented by Gaia Grosso (P) and Purvasha Chakravarti (S); the symbol S denotes a Statistician and P is for Physicist.

The organizers encouraged participants to continue and expand the established contacts, which could be very useful. Some of these are listed below:

- Galin Jones (S) and Ed Atkin (P): Implementing MCMC methods in neutrino analyses.
- Lydia Brenner (P) and Michael Stanley (S): Incorporating unfolding procedures in Particle Physics libraries.
- Roger Barlow (P) and Igor Volobouev (P): Methods for dealing with asymmetric uncertainties.
- Gaia Grosso (P) and Purvasha Chakravarti (S): Procedures for model-independent searches for New Physics.
- Tudor Manole (S) and Philipp Windishhofer (P): Optimal Transport.
- Tommaso Dorigo (P) and Ann Lee (S): End-to-end experimental design.
- Ann Lee (S) and Lukas Heinrich (P), Michael Kagan (P) and Philipp Windischhofer (P): Calibrated forecasting and Optimal Transport.
- Alessandra Brazzale (S), Purvasha Chakravarti (S) and Gaia Grosso (P): Properties of ML methods.
- Sara Algeri (S) and Lydia Brenner (P): Format for future interaction between Physicists and Statisticians.

4 Conclusions

We are very appreciative of BIRS having provided us with the opportunity of having this meeting on the important role of systematics in Particle Physics analyses. The ambiance of the Banff Centre was ideal for fostering dialogue, and we made use of this by having more time for discussions than for talks during the formal sessions. In addition, many of the interesting interactions took place during the coffee breaks, meals, pub visits, outdoor excursions, and on our dedicated Slack channel.

The presence of a good number of Statisticians encouraged the idea of incorporating them in Physics analyses, for the mutual benefit of Physicists and Statisticians. We are pleased that, as a result of this Workshop, there appear to be several examples in which collaborative work is already happening. We hope this will significantly improve the treatment of systematic uncertainties in our data analyses.

References

[1] PHYSTAT events series, `https://espace.cern.ch/phystat`.

[2] PHYSTAT Systematics 2021, `https://indico.cern.ch/event/1051224/`.

[3] P. Dauncey et al., Handling uncertainties in background shapes: the discrete profiling method, *J. Inst.* **10** (2015) P04015.

[4] J. Berger, B. Liseo and R. Wolpert, Integrated likelihood methods for eliminating nuisance parameters, *Stat. Sci.* **14** (1999), 1–28.

[5] R. Cousins, R., J. Linnemann, and J. Tucker, Evaluation of three methods for calculating statistical significance when incorporating a systematic uncertainty into a test of the background-only hypothesis for a Poisson process, *Nucl. Instrum. Meth. A* **595** (2008) 480–501.

[6] P. Gregory, Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with Mathematica Support, Cambridge University Press, 2005.

[7] S. Mollerach and E. Roulet, Progress in high-energy cosmic ray physics, *Progress In Particle And Nuclear Physics* **98** (2018), 85–118.

[8] A. Ray, I. Hernandez, S. Mohite, J. Creighton and S. Kapadia, Non-parametric inference of the population of compact binaries from gravitational wave observations using binned Gaussian processes, `https://arxiv.org/abs/2304.08046`.

[9] H. Lee et al., Accounting for calibration uncertainties in x-ray analysis: effective areas in spectral fitting, *The Astrophysical Journal* **731** (2011) 126.

[10] J. Xu et al., A fully Bayesian method for jointly fitting instrumental calibration and x-ray spectral models, *The Astrophysical Journal* **794** (2014), 97.

[11] D. Jones, D., Kashyap and D. van Dyk, Distentangling overlapping astronomical sources using spatial and spectral information, *The Astrophysical Journal* **808** (2015) 137–160.

[12] A. Meyer et al., eBASCS: Disentangling overlapping astronomical sources II, using spatial, spectral, and temporal information, *Monthly Notices Of The Royal Astronomical Society* **506** (2021) no.04, 6160-6180.

[13] P. Jacob, M. Murray, C. Holmes and C. Robert, Better together? Statistical learning in models made of modules, *ArXiv Preprint ArXiv:1708.08719 [stat.ME].*

[14] D. van Dyk and L. Lyons, How to Incorporate Systematic Effects into Parameter Determination, *ArXiv Preprint ArXiv:2306.05271 [hep-ex].*

[15] N. Dalmasso, L. Masserano, D. Zhao, R. Izbicki and A. Lee, Likelihood-Free Frequentist Inference: Confidence Sets with Correct Conditional Coverage, `https://arxiv.org/abs/2107.03920`.

[16] S. Murphy, and A. W. Van Der Vaart, On Profile Likelihood, *Journal Of The American Statistical Association* **95** (2000) no.450, 449-465.

[17] U. Grenander, Abstract Inference, Wiley, 1981, `https://books.google.com/books?id=ng2oAAAAIAAJ`.

[18] T. Manole, P. Bryant, J. Alison, M. Kuusela, M. and L. Wasserman, Background Modeling for Double Higgs Boson Production: Density Ratios and Optimal Transport, `https://arxiv.org/abs/2208.02807`.

[19] D. Van Dyk, The Role of Statistics in the Discovery of a Higgs Boson, *Annual Review Of Statistics And Its Application* **1** (2014), 41-59.

[20] O. Hines, O. Dukes, K. Diaz-Ordaz and S. Vansteelandt, Demystifying Statistical Learning Based on Efficient Influence Functions, *The American Statistician* **76** (2022) no.03, 292–304.

[21] G. Cowan, K. Cranmer, E. Gross, E. and O. Vitells, Asymptotic Formulae for Likelihood-Based Tests of New Physics, *Eur. Phys. J. C* **71** (2011), 1554.

[22] ATLAS Collaboration, A morphing technique for signal modelling in a multidimensional space of coupling parameters, ATL-PHYS-PUB-2015-047.

[23] C. Pollard and P. Windischhofer, Transport away your problems: Calibrating stochastic simulations with optimal transport, *Nucl. Instrum. Meth. A* **1027** (2022), 166119.

[24] P. Komiske, E. Metodiev and J. Thaler, The Hidden Geometry of Particle Collisions, *JHEP* **07** (2020), 006.

[25] M. Hallin, Measure transportation and statistical decision theory, *Annual Review Of Statistics And Its Application* **9** (2022), 401–424.

[26] B. Amos, L. Xu and J. Kolter, Input Convex Neural Networks, *PMLR* **70** (2017), 146–155.

[27] F. Tackmann, Beyond Scale Variations: Perturbative Theory Uncertainties from Nuisance Parameters, *DESY-19-021*.

[28] P. Calafiura, D. Rousseau and K. Terao, Artificial Intelligence for High-Energy Physics, World Scientific 2022, `https://doi.org/10.1142/12200`.

[29] `https://mode-collaboration.github.io`.

[30] T. Dorigo et al., Toward the End-to-end Optimization of Particle Physics Instruments with Differentiable Programming, *Reviews in Physics* **10** (2023), 100085.

[31] R. D'Agnolo and A. Wulzer, Learning New Physics from a Machine, *Phys. Rev. D* **99** (2019), 015014.

[32] R. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer and M. Zanetti, Learning multivariate new physics, *Eur. Phys. J. C* **81** (2021), 89.

[33] R. D'Agnolo, G. Grosso, M. Pierini, A. Wulzer and M. Zanetti, Learning new physics from an imperfect machine, *Eur. Phys. J. C* **82** (2022), 275.

[34] J. Fan and J. Jiang, Nonparametric inference with generalized likelihood ratio tests, *TEST-MADRID* **16** (2007), 409.

[35] P. Chakravarti, M. Kuusela, J. Lei and L. Wasserman, Model-independent detection of new physics signals using interpretable semi-supervised classifier tests, *ArXiv Preprint ArXiv:2102.07679 [stat.AP]* .

[36] M. Stanley, P. Patil and M. Kuusela, Uncertainty quantification for wide-bin unfolding: one-at-a-time strict bounds and prior-optimized confidence intervals, *Journal Of Instrumentation* **17** (2022), P10013.

[37] D. O'Leary and B. Rust, Confidence intervals for inequality-constrained least squares problems, with applications to ill-posed problems, *SIAM Journal On Scientific And Statistical Computing* **7** (1986), 473–489.

[38] P. Patil, M. Kuusela and J. Hobbs, Objective frequentist uncertainty quantification for atmospheric CO2 retrievals, *SIAM/ASA Journal On Uncertainty Quantification* **10** (2022), 827–859.

[39] P. Stark, Inference in Infinite-Dimensional Inverse Problems: Discretization and Duality, *Journal Of Geophysical Research* **97** (1992), 14055–14082.

[40] D. Acosta et al, First measurements of inclusive W and Z cross sections from Run II of the Tevatron collider, *Phys. Rev. Lett.* **94** (2005), 091803.

[41] `https://drive.google.com/drive/folders/1i2yDyiQo7wQOw0hGv2guwSPwAgIuCfdo`.

[42] M. Kennedy and A. O'Hagan, Bayesian calibration of computer models. *Journal Of The Royal Statistical Society: Series B (Statistical Methodology)* **63** (2001), 425–464.