Common tools — the statistical-combination challenge.

A.k.a:

Sara Bolognesi (CEA Saclay),

Mathieu Lamoureux (INFN Padova),

D. Tonelli (INFN Trieste)

JENNIFER2 kick-off meeting,
Austrian Academy of Sciences

Wien, Sep 12, 2019

What we promised

- The main deliverable will be **a document** detailing recipes on how to properly combine results from different experiments, in presence of multi-parameter analysis:

  emphasis on **combination of likelihoods** as a function of the parameter of interests (~5) and the nuisance parameters (~hundreds)

  (complete likelihood at their highest possible level of dimensionality to preserve coherence of information for further manipulation: profiling/marginalization...)

- Second optional deliverable (if personpower): **software tool** for storing and combination of user-provided likelihoods

A conceptual, technical (and sociological) challenge

In detail

c) Statistical methods for combinations of experimental results

The reach of many crucial measurements of the T2K and Belle2 programs is severely limited by the small size of the event samples used. In this scenario, completely common for neutrino and quark flavour experiments, the combination of the statistical information from multiple measurements has significant potential to enhance the physics reach over the bare combination of the final results. Past results combination attempts have typically been conducted on an ad-hoc basis and after the individual measurements and their methodological choices and approximations had been consolidated. This results in suboptimal combinations limiting the statistical power of the outcomes.

Each individual measurement typically involves a large number of estimated parameters: the physics parameters of interest and many nuisance parameters correlated with them. While the former can be reasonably cast in an universal experiment-independent format and treated consistently in combinations, the latter are partly universal and partly experiment-dependent. This leads to a variety of possible options for the approximations and approaches needed to include their effect in the combination.

We propose a systematic and consistent plan for obviating the above pitfalls that consists in:

- A survey of the Belle2 and T2K physics topics and specific measurements where inter-experiment combinations (with NOvA, LHCb, etc.) have the potential to lead to significant reach enhancements.
- A survey of past and present combination efforts aimed at forming a global picture of the variance of the approaches adopted, the approximations made, and the possible pitfalls/inconsistencies encountered.
- A unified proposal for: (i) restricting the definition of the relevant physics and nuisance parameters for each measurement to one or few variants; (ii) restricting the approximations associated with the modelling of the interplay between nuisance and physics parameters to a few consistent variants. The proposal will be documented in a report that will serve as a reference for experimental groups willing to combine their results, which will be invited to conform to the selected prescriptions.

A possible development of such work could be the set up a software framework (e.g., a data base) explicitly suited and optimized for (i) accepting as inputs the values of multivariate likelihoods from each individual measurement and (ii) operating consistently the combination (likelihood multiplication) taking properly into account the commonalities between physics and global nuisance parameters and treating coherently experiment-dependent nuisance parameters. If successful, this work will enhance the physics reach of the single experiments both in neutrino and quark flavour physics.

Paris recap

In Paris we laid down the concept

Common tools:
statistical methods for combination of experimental results

JENNIFER2 meeting
(October 2018, Paris)

S.Bolognesi (CEA Saclay) and D.Tonelli (INFN Trieste)

https://agenda.infn.it/event/16350/timetable/#20181030.detailed

- Guidelines/tools to assist inter-experiment combinations of results — a T2K-Belle II commonality with enabling potential to boost the science.
- Outlined a few benchmark physics cases and previous approaches to combinations
- Identified the optimal solution in a “likelihood multiplier”, based on a carefully chosen set of common inputs and assumptions

Today

An “organizational” roadmap to get us there

Intelligence — setting the stage

- Which are the standard statistical procedures in T2K? Which in Belle II? How are these implemented (tools? assumptions? choice of variables/observables)

- What are the typical limitations they suffer or issues they run into?

- Which constraints do such procedures impose on the combination options (if any)?

- Which technical constraints, if any?

- What are the topics/measurements most likely to benefit from combinations? Which of these are easier/harder to consider for a centralized-combination effort?

This is ongoing. No major logistical bottlenecks — remote communications suffice.

Synthesis — sorting the options

- What realistic options for combinations exist that are compatible with the picture emerged from the previous step?
- Which are conceivably implementable in software within the existing constraints?
- What's the extent of the Belle II-T2K commonalities?
- Can/should our guidelines be experiment independent? And our tool?
- Is the climate within experiments toward this effort collaborative/neutral/hostile?
- Are there options more likely to get traction beyond Belle II/T2K?

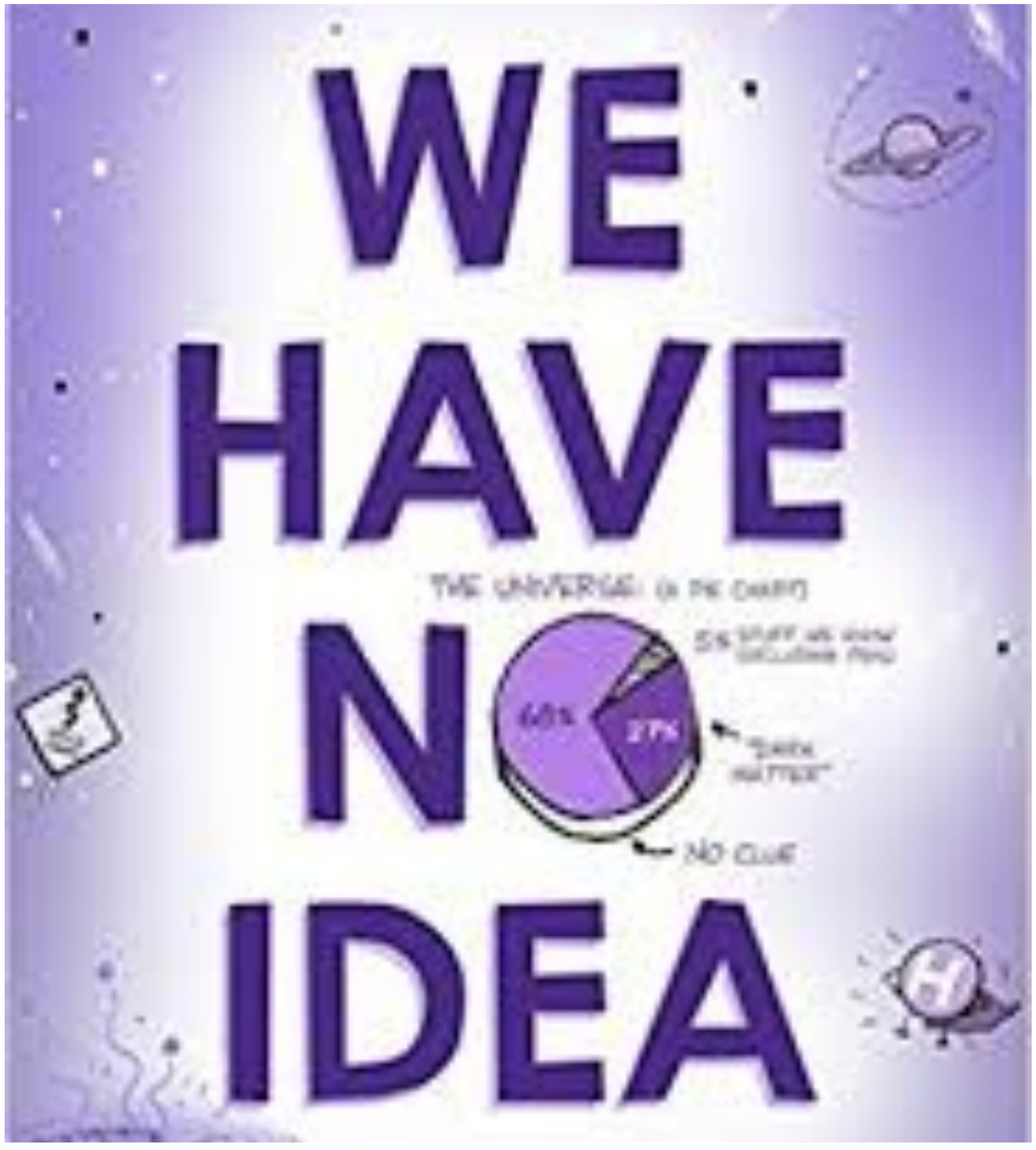
Will probably require more in-person interactions/discussions.

Creativity — the document

- Content? (just conceptual stuff? Demonstrations based on simulated/real data too?)
- Format/length?
- Arxiv or journal? Which target journal?
- Who writes what?
- Authors? Editors? Reviewers?

Additional volunteers welcome. A few in-person interactions. A lot of offline work.

Impact — the software



Well, not exactly…

The core combination engine could be as simple as setting up a wrapper for likelihoods that takes 3 experiment-provided inputs:

- a C++ function for each experiment (that is, the likelihood in whatever format experiments fancy, provided it has a defined C++ interface)
- a matrix of correlations between nuisance parameters for each experiment
- a mechanism to define correlations between nuisance parameters of different experiments.

There’s much more than that to it…

A significant overhead of scientific and nonscientific work to allow for the combination engine to operate within the proper framework and achive its goal

- Common inputs: ensure that experiments are willing to produce their likelihoods as functions of a common set of physics parameters (same variables, same ranges, etc..) and a common set of parameters/assumptions for the shared nuisance parameters.
  ===> (This is hard: when one think of combining results, analysis choices are already frozen and established, analyzers are very territorial about their choices,).

- Collaborative benchmarking: availability of simulated or real data (and serious guidance on how to use them) to allow benchmarking the combiner before public release.

- Extend the scope? Could such a tool become an attempt at introducing inter-experiment data-sharing, something that’s quite taboo in collider physics but it’s the industry standard in other branches?.

- …

The software

Current people on the project can do the “preparation steps” within the experiments to set up a common combination grounds

But the chances of transforming anstract recommendations into a useful tool depends critically on the possibility that a dedicated, technically-competent person joins us nearly full time for ~12 months

- Survey existing tools to find the “optimal” framework
- Code it up…
- Test it, optimize it,
- Benchmark it

Where do we stand? A snapshot

Setting the stage: July 2019, first surveys...

Likelihoods how many data events are typically fit? Whats the dimensionality of the space of observables? What's the dimensionality of the space of parameters? How many of the latter are "physics/interesting" parameters? How many are "nuisance" parameters? How many of the latter are known from external measurements? How many are associated with theoretical inputs or other non-obviously gaussian inputs? Any specific issue with likelihood minima approaching the boundaries of the variable's domains?

** Technicalities** which toolkits are used to implement likelihoods? (Roofit?Roostat?Pyroot? custom?), how much computing power a minimization takes approximately? How much time toy generation (frequentist) or marginalization (bayesian)?

Inference methods: frequentist confidence region construction? Bayesian posterior probability? Others? For the frequentists methods, which ordering is used? How's the treatment of nuisance parameters implemented both in fitting and in generation? For the Bayesian methods, what choices are made for the priors and how sensitivity of results to priors is addressed?

** Combination ** What's have been previous experiences of combining results with other experiments (eg, NOvA)? What were the chief scientific difficulties? What were the chief non-scientific (technical, political..) difficulties?

The jungle…

- Events fed to fits: anywhere from few 10's to billions

- Dimensionality of the space of observables (yes, that's the number of fit parameters): anywhere from few 10's to nearly 1000 —dominated by nuisance parameters.

- Inference choice: anything goes — frequentist, pseudo-frequentist, Bayesian.

- Assumptions for the nuisance parameters: Gaussian very popular but known to be critical for inputs that are not externally measured. E.g, theory inputs (unknown distribution, if any, by definition) or mixed theory-experimental inputs.

- Likelihood nonlinearities: sure —multiple minima, parameters hitting physical bounds, you name it

- Tools: the whole spectrum from Roofit (Belle II) to custom tools (T2K). Fit timing from few minutes to few months…

Have fun with that..

$\vec{x}$ is the combination of oscillation parameters ($\sin^2\theta 12$, $\sin^2\theta_{23}$, $\sin^2\theta_{13}$, Δm^2_{32}, Δm^2_{21}, δ_{CP}) and the **751** nuisance parameters: flux parameters $\vec{b}$, interaction parameters $\vec{xs}$ and detector uncertainties parameters for ND280 $\vec{d}_n$ and for SK $\vec{d}_s$.

The likelihood

Oscillation priors: only priors for $\sin^2\theta_{13}$, $\sin^2\theta_{12}$ and Δm^2_{21}, and no correlations, so that the diagonal terms representing $\sin^2\theta_{23}$, $|\Delta m^2_{32}|$ and δ_{CP} are zero (uniform prior) and all the off-diagonal terms are zero. The prior for θ_{13} can also be chosen to be uniform (not using reactor neutrino experiments constraints).

19

$$\mathcal{L}(N_e^{obs}, N_\mu^{obs}, \mathbf{x}_e, \mathbf{x}_\mu, \mathbf{o}, f) = \mathcal{L}_e(N_e^{obs}, \mathbf{x}_e, \mathbf{o}, f) \times \mathcal{L}_\mu(N_\mu^{obs}, \mathbf{x}_\mu, \mathbf{o}, f) \times \mathcal{L}_{\text{syst.}}(f)$$

- $\mathcal{L}_e$ and $\mathcal{L}_\mu$ are binned Poisson likelihood with bins in 1D E_{rec} for *mu*-like events and 2D $p-\theta$ for e-like events.
- $\mathcal{L}_{\text{syst}}$ is a multivariate gaussian in dimension $d = 119$ systematic parameters: 50 parameters for flux ; 20 parameters for neutrino interactions (after near-detector constraints).

To make things worse

Technical challenges can be overcome — and have been overcome, as recent history of combinations shows (e.g., LHCb+CMS, CDF+D0, Belle+Babar).

Such history also tells us that the most severe limitations are perhaps conceptual/political: have big experiments agreeing on underlying models, analysis choices, and concepts will be hard.

Roadmap

- Setting the stage (mid-2019—mid-2020)
- Sorting the options (early 2020— late 2020)
- The document (late 2020—mid-2021 using the intermediate evaluation deadline as a target)
- The document II (final version published by mid 2022)
- The software (first incarnation by mid-2022)
- The end (software benchmarked on one or two real use-cases and available online, mid-2022-late 2023)

Conclusions

There are no conclusions (we are just at the beginning).

But at least we started.

Conclusions

There are no conclusions (we are just at the beginning).

But at least we started.