Combinations

J. Baudot (IPHC Strasbourg), D. Greenwald (TUM Munich), G. Inguglia (DESY), K. Kinoshita (Cincinnati), T. Kuhr (LMU Munich), F. Le Diberder (LAL Orsay), E. Prencipe (Juelich Furschungszentrum), Diego Tonelli (INFN Trieste), and B. Yabsley (Sydney)

Belle II Physics Week - KEK Oct 29, 2019

Beware

This talk will not discuss anything you don’t know already.

This talk is shorter than the slot allocated for it (I only discovered yesterday night that my slot was 45 minutes…)

Blame me for any mistakes — my fellow SC members had no real chance to comment on a decent draft.

Why

To improve the results

“Improve” isn’t universally defined — really depends on your scientific goal.

Usually it’s intended with “reducing the variance”.

But there are cases where mild variance increase might be acceptable, if the combination achieves a more unbiased estimate, or a reduction of the relative impact of the systematic component on the total variance.

When

No-brainers

- independent inputs with similar uncertainties dominated by the statistical component, and no updates of the inputs foreseen in the near future ==> combine!

- inputs with similar statistical and systematic uncertainties dominated by a fully correlated systematic component ==> don’t waste time in combining!

All the other cases depend on the details of the inputs, priorities, timing etc..

Quick recap of basics / notation

- Measurement: combine observed data x into statistical model p(x|m) to infer the value of parameter m and its uncertainty.

- p(x|m) is called pdf when interpreted as a function of data, and likelihood when interpreted as a function of the parameter

- The value $\hat{m}$ of the parameter m that maximizes the likelihood has optimal (asymptotic) properties: unbiased, efficient (smallest variance), Gaussian distributed.

- In case of Gaussian likelihoods, the maximum likelihood estimator can be recasted in terms of the least-squares statistics ($-2 \ln L(m) \propto \chi^2$), which has attractive properties (namely, a ***known*** distribution) if (i) expected values m are known, values of the control variable x are known, the variances of the observations are known, observations are Gaussian-distributed observations around expected values.

Basics

In its most common incarnation, a measurement of one parameter m is typically a point-estimate (central values with uncertainty) $x_i \pm \sigma_i$

Interpret it as single sampling from a pdf $p(x_i|m) = L(m)$, which is the assumed model that relates the observed quantity with the parameter to be estimated.

Under some regularity assumptions, L(m) approaches a Gaussian

$$p(x_i \,|\, m) = L(m) \propto \exp\left[-\frac{(x_i - m)^2}{\sigma_i^2}\right]$$

The maximum likelihood estimator of m is x_i, with variance deviation σ_i^2. Our result will be $\hat{m} = x_i \pm \sigma_i$ Nothing new so far. Nothing has been combined.

Note also that $-2 \ln L(m) \propto \frac{(x - m)^2}{\sigma^2}$

(Maximizing the likelihood in this Gaussian case amounts to minimizing the least-squares)

How — many data, one parameter

Extend to N independent measurements of one parameter m. Since they are independent, multiply their probability densities to get the total density

$$p(\vec{x}\,|\,m) = \prod_i^N p(x_i\,|\,m) = L(m) \propto \prod_i^N \exp\left[-\frac{(x_i - m)^2}{\sigma_i^2}\right]$$

The maximum likelihood estimator of m is

$$\hat{m} = \sum_i w_i x_i, \quad \text{with} \quad w_i = (1/\sigma_i^2) \Big/ \left(\sum_k 1/\sigma_k^2\right) \quad \text{and} \quad \sum_i w_i = 1 \quad \text{with variance}$$

$$\hat{\sigma}^2_{\hat{m}} = 1 \Big/ \left(\sum_i 1/\sigma_i^2\right) \quad \text{that is, the weighted mean}$$

which also minimizes the least-squares $-2\ln L(m) \propto \sum_i^N \frac{(x_i - m)^2}{\sigma_i^2}$

whose distribution approximates that of χ^2 if the numerators are truly Gaussian and the observed variances σ_i^2 are sufficiently close to the true ones

How — many data, many parameters

Generalize to the case of N point estimates, each of a set of n parameters. The optimal combined value for $\vec{m}$ is achieved by minimizing

$$\mathrm{LS}(\vec{m}) = \sum_{i}^{N} (\vec{x}_i - \vec{m})^T V_i^{-1} (\vec{x}_i - \vec{m})$$

uncertainty is encoded in the covariance matrix $V^{-1} = \sum_i V_i^{-1}$

In addition, the value of LS ($\vec{m}$), offers an approximate measure of the consistency of the inputs to each other, knowing that the inputs are supposed to be distributed as a Gaussian and therefore LS($\vec{m}$) approximates a χ^2 distribution

$$\mathrm{LS}(\vec{m}) = \sum_{i}^{N} (\vec{x}_i - \vec{m})^T V_i^{-1} (\vec{x}_i - \vec{m}) \approx \chi^2(\vec{m})$$

Example #1 — Heavy Flavor Averaging Group

It is important for any averaging procedure that the quantities measured by experiments be statistically well-behaved, which in this context means having a (one- or multi-dimensional) Gaussian likelihood function that is described by the central value(s) $\boldsymbol{x_i}$ and covariance matrix $\boldsymbol{V_i}$. In what follows we assume that $\boldsymbol{x}$ does not contain redundant information, *i.e.*, if it contains n elements then n is the number of parameters being determined. A χ^2 statistic is constructed as

$$\chi^2(\boldsymbol{x}) = \sum_{\boldsymbol{i}}^{\boldsymbol{N}} (\boldsymbol{x_i} - \boldsymbol{x})^{\mathrm{T}} \boldsymbol{V_i^{-1}} (\boldsymbol{x_i} - \boldsymbol{x}) \ , \tag{1}$$

where the sum is over the N independent determinations of the quantities $\boldsymbol{x}$, typically coming from different experiments; possible correlations of the systematic uncertainties are discussed below. The results of the average are the central values $\hat{\boldsymbol{x}}$, which are the values of $\boldsymbol{x}$ at the minimum of $\chi^2(\boldsymbol{x})$, and their covariance matrix

$$\hat{\boldsymbol{V}}^{-1} = \sum_{\boldsymbol{i}}^{\boldsymbol{N}} \boldsymbol{V_i^{-1}} \, . \tag{2}$$

We report the covariance matrices or the correlation matrices derived from the averages whenever possible. In some cases where the matrices are large, it is inconvenient to report them in this document, and they can instead be found on the HFLAV web pages.

The value of $\chi^2(\hat{\boldsymbol{x}})$ provides a measure of the consistency of the independent measurements of $\boldsymbol{x}$ after accounting for the number of degrees of freedom (dof), which is the difference between the number of measurements and the number of fitted parameters: $N \cdot n - n$. The values of $\chi^2(\hat{\boldsymbol{x}})$ and dof are typically converted to a confidence level (C.L.) and reported together with the averages.

Example #2: Particle Data Group

Same as HFLAV, but with an important difference: the choice of the inputs, and the uncertainty of the combined value might be modified ad-hoc depending on how the LS of the combination

$$\mathrm{LS}(\vec{m}) = \sum_{i}^{N} (x_i - \vec{m})^T V_i^{-1} (x_i - \vec{m})$$

compares with the N-1 value expected for a chi2 distribution in case Gaussian distributions of uncertainties.

If $\mathrm{LS}(\vec{m}) \leqq$ N-1, then no modification

If $\mathrm{LS}(\vec{m}) >$ N-1, then scale the uncertainty of the combined value by $S = \sqrt{[\chi^2/(N-1)]}$

If $\mathrm{LS}(\vec{m}) \gg$ N-1, no average at all.

Complications

- The previous discussion fails if uncertainties are inconsistent (i.e., rely on different assumptions) or correlated (i.e., do not fluctuate independently) across inputs

- The previous discussion relies on the all-important assumption of Gaussian distribution of the measurements around their true values. If that assumption is no longer valid, the previous treatment fails

- The previous discussion is hard to apply to inputs where no point estimates are provided, but just confidence intervals.

Complication #1: inconsistent/correlated uncert.

Things get hairy when the uncertainties (typically the systematic ones) are inconsistent or correlated.

Usually the model p(x|m) is an approximation of p(x|m, ν) where ν is one or multiple auxiliary external parameters (so-called nuisance parameters).

While we are not interested to determine ν from our data, its value impacts the measurement as it modifies the shape p(x|m) and therefore our inference of m. Because of the presence of ν, the uncertainty on the inputs may get

- inconsistent, e.g., different values for such external parameter (a particle mass or lifetime, a decay constant, etc.) have been assumed to derive each input so that combining them will yield apple-with-oranges situation.

- correlated, e.g., the uncertainty on the external parameter ν introduces a common-mode component into the uncertainties of inputs.

BLUE

In case of correlated uncertainties and assuming known the true variances and correlations, one can extract the “best linear unbiased estimate” (unbiased, smallest variance)

Aitken, Proc. Roy. Soc. Edinburgh 55, 42 (1935); Lyons et al., NIM A 270, 110 (1988); Valassi, NIM A 500, 391 (2003)

Two unbiased measurements of m $m = \hat{m}_1 \pm \sigma_1$ $m = \hat{m}_2 \pm \sigma_2$ with **known** variances σ_i and correlation ρ

The BLUE estimate follows by finding the weights α, β that keep the linear combination $\hat{m} = \alpha\,\hat{m}_1 + \beta\,\hat{m}_2$ unbiased while minimizing its variance

weight weight

$$\hat{m} = \frac{\hat{m}_1(\sigma_2^2 - \rho\sigma_1\sigma_2) + \hat{m}_2(\sigma_1^2 - \rho\sigma_1\sigma_2)}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2}$$

$$\sigma_{\hat{m}}^2 = \frac{\sigma_1\sigma_2(1-\rho^2)}{\sigma_1^2 - 2\rho\sigma_1\sigma_2 + \sigma_2^2}$$

Depending on correlations, weights could be negative, which might appear counterintuitive (without that measurement I’d get a better result).
Generalizes to many measurements and multiple parameters

Similar methods

JHEP

Published for SISSA by Springer

Received: November 5, 2009
Accepted: December 19, 2009
Published: January 25, 2010

Combined measurement and QCD analysis of the inclusive $e^{\pm}p$ scattering cross sections at HERA

H1 and ZEUS collaborations

Averaging of DIS Cross Section Data

A. Glazov

DESY, Notkestrasse 85, Hamburg, D 22603 Germany

Eur. Phys. J. C (2009) 63: 625–678
DOI 10.1140/epjc/s10052-009-1128-6

THE EUROPEAN PHYSICAL JOURNAL C

Regular Article - Experimental Physics

Measurement of the inclusive *ep* scattering cross section at low Q^2 and x at HERA

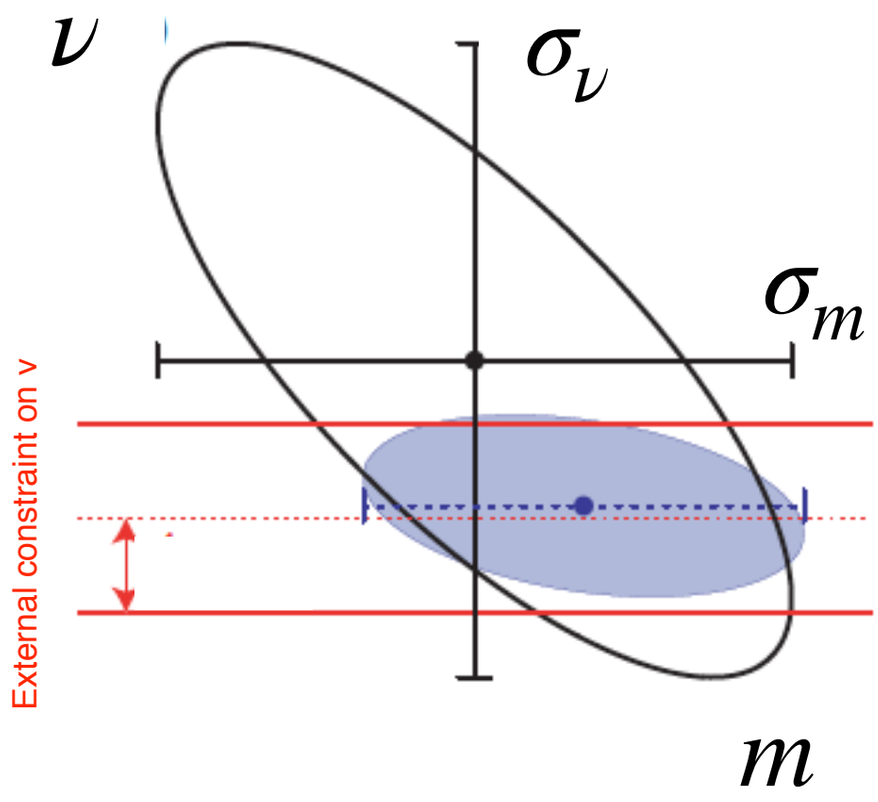
The H1 Collaboration

HFLAV

Some approximations are possible if distributions are or approximate Gaussians.

HFLAV chooses to explicitly include the input dependence on the external parameters ν by increasing the dimensionality of the problem and interpreting it as a joint measurement of m and ν, p(x|m) ==> p(x|m ν)

This way, any additional external information or constrain on ν is applied only once during the combination (typically assuming it Gaussian)



Satisfactory approximation if things are effectively Gaussian.
May be very poor in other cases.

Complication #2: venturing in non-Gaussian-land

Venturing in non-Gaussian-land

Our findings so far assumed -2lnL(m) distributed as a χ^2 distribution:

$$-2\ln L(m) \propto \sum_i^N \frac{(x_i - m)^2}{\sigma_i^2} \approx \chi^2(m)$$

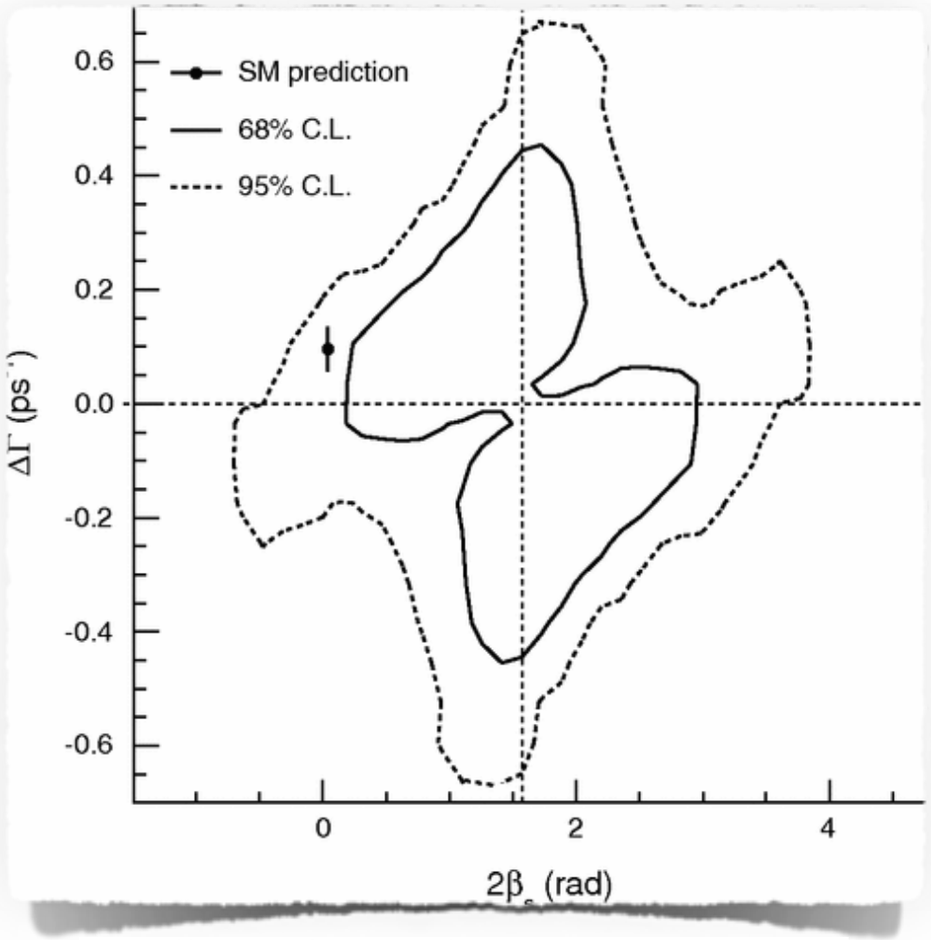
which requires that the numerator is distributed as a Gaussian and that the observed variances σ_i^2 approximate well the true variances

Likelihood theory tells us that maximum likelihood estimators become Gaussian only *asymptotically* (N = ∞, that is, *never*).

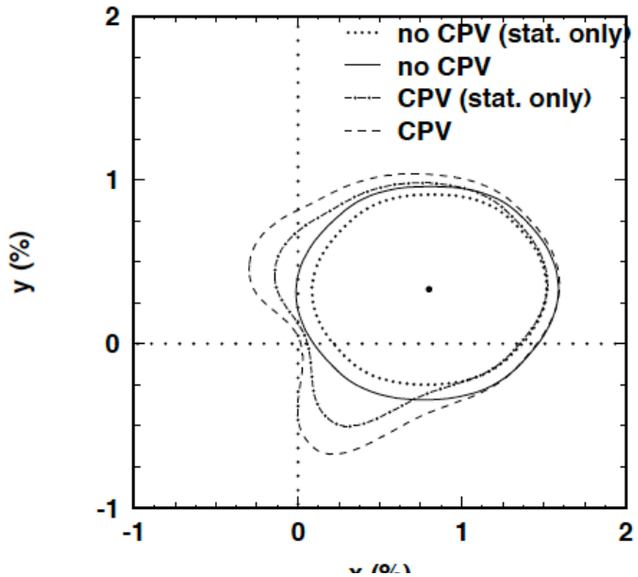
No theorem tells us when we have sufficient data for the likelihood in our problem to approximate the asymptotic regime. How do we check?

First obvious test is to look at the likelihood in your data: not uncommon to find any kind of wild non-Gaussianities: multiple minima, minima that modify the likelihood dimensionality, minima that approach the physical boundaries, you name it..

Venturing in non-Gaussian-land

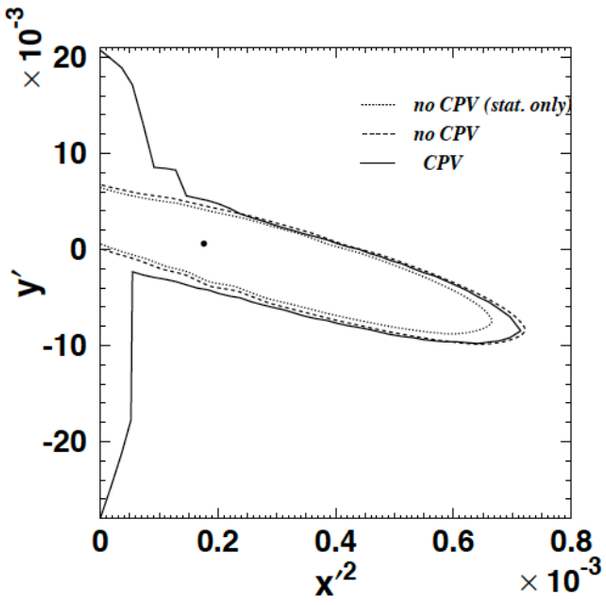


https://doi.org/10.1103/PhysRevLett.100.161802



https://doi.org/10.1103/PhysRevLett.99.131803

https://doi.org/10.1103/PhysRevLett.96.151801



If the previous methods are applied in such cases, a number of unpleasant consequences are in order

- unreliable uncertainties (no coverage)
- biased combined central values

Is my likelihood Gaussian?

So, if the likelihood in our data appears to be non Gaussian we are certainly in trouble.

But even if that looks Gaussian, it is not guaranteed that it will remain so for other sets of data or choice of true values (the ML estimator is a statistics, and as such it undergo fluctuations)

When in doubt, use Monte Carlo: generate many ensembles of data each with a different choice a true values and study the properties of your estimator.

When the distributions of the estimators of my inputs are clearly non-Gaussian, there are no shortcuts to:

- find parametrizations that make your inputs more Gaussian
  (e.g., Appendix A in https://doi.org/10.1103/PhysRevD.99.012007)
- Perform a combined analysis

Combined analysis

The only reliable way of handling such cases is to embark into a full-fledged combined analysis of the input data sets.

This is done effectively by analyzing jointly/simultaneously the data.

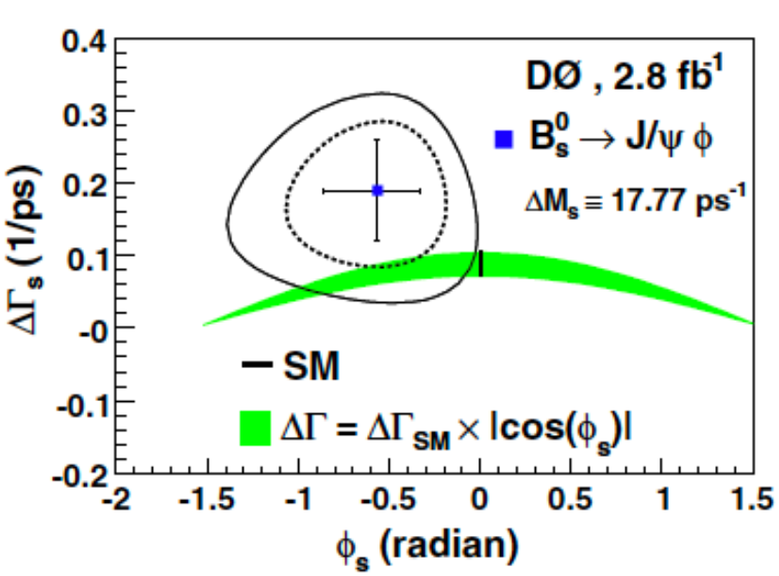
Alternatively, combine (multiply) the ***likelihood*** functions from each analysis, provided that they are written as functions of the same set of physics and common nuisance parameters (e.g, avoiding $\sin(\alpha)$ vs α, m^2 vs m, τ vs Γ), and sampled on common ranges with same granularity.

This approach takes also care naturally of inputs expressed in terms of confidence intervals (as opposed to point estimates).

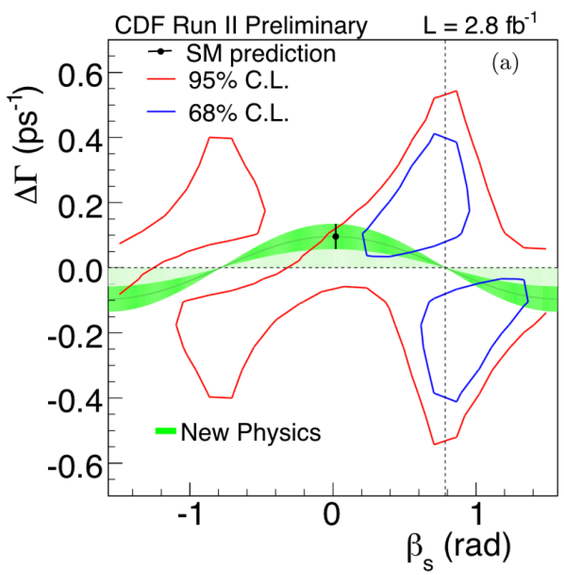
It is important to combine real likelihoods, as opposed to profiled-likelihood (lower-dimensional mathematical functions obtained by deriving the likelihood wrt some parameters) or posterior densities (lower-dimensional mathematical functions obtained by integrating the product of the likelihoods with other functions)

Is it possible?

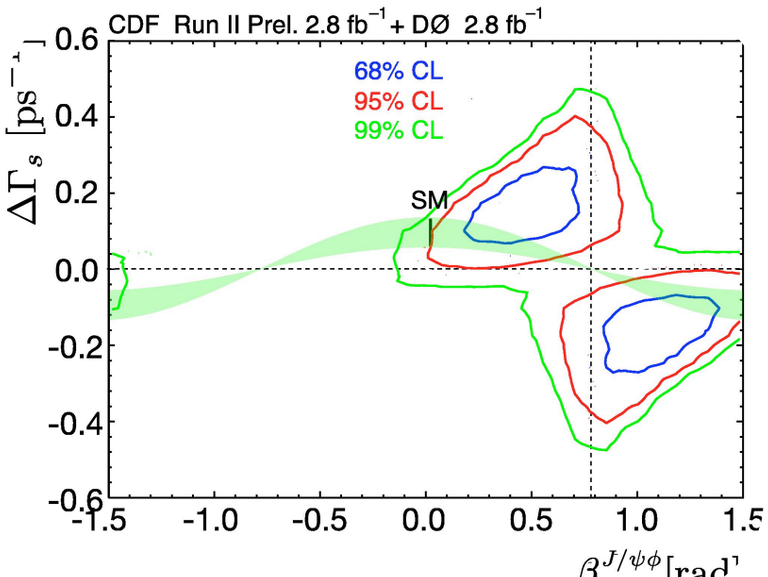
Yes. And usually the properties of the combined likelihood typically improve over those of input likelihoods:



X



=



https://doi.org/10.1103/PhysRevLett.101.241801 arXiv:0810.3229

…although it is unlikely that the combination of two non-Gaussian likelihoods yields a Gaussian result.

It is worth?

Featured in Physics | Editors' Suggestion | Access by KEK High Energy Acce

Observation of s-Channel Production of Single Top Quarks at the Tevatron

T. Aaltonen *et al.* (CDF Collaboration†, D0 Collaboration‡)
Phys. Rev. Lett. **112**, 231803 – Published 9 June 2014

Physics See Viewpoint: Top Quarks Go Solo in Rare Events

First Observation of CP Violation in $\bar{B}^0 \to D^{(*)}_{CP} h^0$ Decays by a Combined Time-Dependent Analysis of *BABAR* and Belle Data

A. Abdesselam *et al.* (BaBar Collaboration, Belle Collaboration)
Phys. Rev. Lett. **115**, 121604 – Published 16 September 2015

LETTER

OPEN

doi:10.1038/nature14474

Observation of the rare $B^0_s \to \mu^+\mu^-$ decay from the combined analysis of CMS and LHCb data

The CMS and LHCb collaborations*

Editors' Suggestion | Open Access | Access by KEK H

Measurement of $\cos 2\beta$ in $B^0 \to D^{(*)}h^0$ with $D \to K^0_S\pi^+\pi^-$ decays by a combined time-dependent Dalitz plot analysis of *BABAR* and Belle data

I. Adachi *et al.* (*BABAR* Collaboration, Belle Collaboration)
Phys. Rev. D **98**, 112012 – Published 26 December 2018

It is worth?

In the present measurement, the benefit is twofold: first, the combination of the *BABAR* and Belle data samples improves the achievable experimental precision by effectively doubling the statistics available for the measurement; second, the combined approach enables common assumptions and applies the same $D^0 \to K^0_S\pi^+\pi^-$ decay amplitude model simultaneously in the analysis of the data collected by both experiments. The approach of combining *BABAR* and Belle data enables unique experimental sensitivity beyond what would be possible by combining two independent measurements, in particular for $\cos 2\beta$.

In this Letter, the two sets of data are combined and analysed simultaneously to exploit fully the statistical power of the data and to account for the main correlations between them.

The use of these two results by both CMS and LHCb is the only significant source of correlation between their individual branching fraction measurements. The combined fit takes advantage of the larger data sample to increase the precision while properly accounting for the correlation.

A minor (?) point — blindness..

It’s hard to perform combinations while fully respecting the blinding philosophy.

While procedures can be established before knowing the combined result, one typically “senses” where the final result will be, since known inputs are used.

In addition, one is somewhat incentivized to combine if she suspects that the combined result will be more useful/exciting than the individual inputs (e.g, when both are close to a conventional threshold for “evidence/observation”)

(To some extent, this arbitrariness is similar to deciding what to do of the apparently “inconsistent” data in averages)

Aside: combining limits

No generally accepted and robust strategy exists to combine exclusion limits.

It is in fact, much more straightforward to combine the corresponding point-estimates and then derive the limit afterward, from the combined result.

Whenever the result of your measurement is an exclusion limit — please please please put the central value +/- uncertainty in the paper too. It will make combiners' lifes much simpler down the line.

Take home messages

Combining results can be good for you. You should be knowing what you are doing.

Tabulate, store, document, and **publish** the full *likelihood* (not the *profile likelihood* or the *posterior density*) of your analysis. It's a complete summary of the data. It will greatly facilitate future combinations, if any.

(Who knows? Maybe in future HEP collaborations will evolve to be less territorial and NASA-like data sharing will be easier)

The quality of the combination won't exceed the quality of the inputs: rather than neurotizing in constructing oversophisticated combination frameworks, do focus on making your analysis as much robust and accurate as possible — that's likely to become the most relevant contribution to any future combination.

Combining is not always a good idea. It's time consuming and less fun than doing the real analysis — only worth if there is an obvious scientific benefit.

Thanks for your attention