HAverager

Mikhail Karnevskiy[1]

[1]DESY

Dubna, 19.02.2016

Introduction

- Presented tool is based on HERAaverager (also known as H1averager, f2ave).
- The tool was originally developed to combined H1-data, later H1 and Zeus data. Currently it is widely used in ATLAS experiment.
- HAverager is a stand-alone programme for combination (averaging) of several statistically-independent measurements using χ^2-minimization approach.
- The combination shows the compatibility of the measurements and deliver the combined value with smaller uncertainties.

Early papers and presentation are here: [1] – [4].

Basic χ^2

For a measurement μ with uncertainty Δ, assuming a Gaussian shape of the uncertainty, the measurement can be considered as a probability distribution function for a "true" quantity m:

$$P(m) = \frac{1}{\sqrt{2\pi}\Delta} \exp\left(-\frac{(m-\mu)^2}{2\Delta^2}\right) \quad (1)$$

This can be written as a χ^2 function by taking $-2\log$ (constant term was skipped):

$$\chi^2(m) = \frac{(m-\mu)^2}{\Delta^2} \quad (2)$$

χ^2 of two measurements

In case of two statistically independent measurements of the case quantity m: μ_1, Δ_1 and μ_2, Δ_2, the probability distribution function of m is given by the product of two:

$$P(m) \sim exp\left(-\frac{(m-\mu_1)^2}{2\Delta_1^2}\right) exp\left(-\frac{(m-\mu_2)^2}{2\Delta_2^2}\right), \qquad (3)$$

which corresponds to χ^2 that is given by the sum of the two: $\chi^2_{sum} = \chi^2_1 + \chi^2_2$.

$$\chi^2(m) = \frac{(m-\mu_1)^2}{\Delta_1^2} + \frac{(m-\mu_2)^2}{\Delta_2^2} \qquad (4)$$

Basic χ^2 minimization

Since χ^2_{sum} is a positive definite quadratic form it can be re-written in the form of Eq. 2. In this case μ is replaced by average μ_{ave} and Δ is replaced by the uncertainty on this average:

$$\chi^2_{sum}(m) = \frac{(m - \mu_{ave})^2}{\Delta^2_{ave}} + \chi^2_0, \qquad (5)$$

where the value of χ^2_0 measures consistency of the measurements, $\chi^2/N_{DoF} \sim 1$ for consistent measurements.
The value of μ_{ave} can be found by minimizing χ^2_{sum} with respect to m (this leads to a usual weighted averaging).

Basic χ^2 for binned measurement

Many experiments measures a number of independent quantities μ_i which correspond to the underlying physics values m_i (e.g. cross-section measurement in bins of $p_{T,Z}$, where i refers to a bin number). In this case the χ^2 function is a simple sum over the measurements (bins):

$$\chi^2_{exp}(m_i) = \sum_i \frac{(m_i - \mu_i)^2}{\Delta_i^2}, \quad (6)$$

Where:

- μ_i - the measurement in the bin i
- m_i "truth" value in the bin i
- Δ_i statistical uncertainty in bin i

Bin-to-bin correlated uncertainties

The systematic effects, which affects the measurement μ_i, are often correlated across bins. Let's consider measurement binned in a certain variable, which is affected by up/down shift of certain parameter:

$$\mu_i \to \mu_i + \Gamma_i^+, \ \ \mu_i \to \mu_i - \Gamma_i^-, \tag{7}$$

where $\Gamma_i^{\pm}$ correspond to the variation up/down.
If the correlated systematic uncertainty is approximately symmetric, one can symmetrize them:

$$\Gamma_i = max(|\Gamma_i^+|, |\Gamma_i^-|)\frac{\Gamma_i^+}{|\Gamma_i^+|}, \tag{8}$$

i.e. the size of the uncertainty is taken as maximal of up and down variations and the sign from one of the variations.

χ^2 for correlated uncertainties

Systematic uncertainties, like energy scale, can be also viewed as a result of an experiment (e.g. measurement of the calibration): there is a "true" detector energy scale α, measured detector calibration α_0 and its uncertainty Δ_α. Therefore, it is natural to add term

$$\chi^2_{syst}(\alpha) = \frac{(\alpha - \alpha_0)^2}{\Delta^2_\alpha} \equiv b^2 \tag{9}$$

to the χ^2 function. The nuisance parameter b, defined as $b = (\alpha - \alpha_0)/\Delta_\alpha$ corresponds to a coherent change of measurements $\mu_i \to \mu_i + b\Gamma_i$. E.g. taking one σ deviation of the detector energy scale we will back to $\mu_i \to \mu_i + \Gamma_i$.

Resulted χ^2

$$\chi^2(\vec{m}, \vec{b})_{exp} = \sum_i \frac{(m_i - \mu_i - \sum_j \Gamma_i^j b_j)^2}{\Delta_i^2} + \sum_j b_j^2, \tag{10}$$

- $\vec{b}$ defines a vector of nuisance parameters b_j corresponding to each source of systematic uncertainty,
- summation over i runs over all data points, and summation over j runs over all correlated sources of systematic uncertainty,
- Γ_i^j is the absolute correlated systematic uncertainty,
- Δ_i is the uncorrelated (statistical) uncertainty.

With this definition minimum χ^2 is obtained for all $m_i = \mu_i$ and $b_j = 0$. If $b_j = 0$ for all j except $j = k$, $b_k = 1$, then χ^2 minimum is archived at $m_i = \mu_i + \Gamma_i^k$ and it is equal to 1.
Total uncertainty for a parameter m_i defined by $\Delta\chi^2_{exp} = 1$ rule corresponds to the sum of correlated and uncorrelated uncertainties in quadrature:
$\Delta^2_{i,tot} = \Delta_i^2 + \sum(\Gamma_i^j)^2$.

Correction and new features of HAverager

- Multiplicative uncertainties
- Correction of statistics bias
- Asymmetrical uncertainties
- Post-rotation of the systematic uncertainties

Everything is now documented in the manual

Multiplicative uncertainties

Consider two measurements μ_1 and μ_2 of m. Let's assume, that $\mu_1 = m + mb$, $\mu_2 = m - mb$. Both measurements are performed with the same relative uncertainty δ. A weighted average of the two measurements returns

$$\mu_{ave} = m\frac{1 - b^2}{1 + b^2}, \tag{11}$$

which for $b = 5\%$ corresponds to 0.5% bias.
The bias occurs because the measurement at smaller value μ_2 got smaller absolute uncertainty $\delta(m - mb)$.
Measurements with multiplicative uncertainties can be combined bias-free using expected values m_i instead of measured μ_i to translate relative to absolute uncertainties. In this case Eq. 10 takes form:

$$\chi^2(\vec{m}, \vec{b})_{exp,mult} = \sum_i \left(\frac{m_i[1 - \sum_j \gamma_i^j b_j] - \mu_i}{\delta_i m_i} \right)^2 + \sum_j b_j^2, \tag{12}$$

Stat bias

Let's consider the counting of number of arbitrary events. Two measurements μ_1 and μ_2 gives $\mu_1 = N_1$, $\mu_2 = N_2$. Statistical uncertainties of the measurement are estimated as a square root of number of counts. Weighted average for these measurement returns:

$$\mu_{ave} = \frac{2N_1 N_2}{N_1 + N_2}, \tag{13}$$

instead of

$$\mu_{ave} = \frac{N_1 + N_2}{2} \tag{14}$$

Stat bias correction

Bias for statistical average can be removed by using expected instead of measured number of events. If statistical uncertainty for a measurement is quoted based on square root of number of measured events, then estimated unbiased relative statistical uncertainty $\delta_{stat,cor} = \frac{\sqrt{m}}{\mu} = \delta_{stat}\sqrt{\frac{m}{\mu}}$. Absolute unbiased statistical uncertainty can be expressed as:

$$\Delta_{stat,cor} = \delta_{stat}\sqrt{m\mu} \tag{15}$$

Finally, the number of observed events can be modified by the correlated systematic uncertainties. This modification can be taken into account by using

$$m(1 - \sum_j \gamma^j b_j) \tag{16}$$

instead of m in Eq. 15. This brings us to the χ^2 formula:

$$\chi^2(\vec{m}, \vec{b})_{exp,cor} = \sum_i \frac{(m_i[1 - \sum_j \gamma_i^j b_j] - \mu_i)^2}{\delta^2_{i,stat}\mu_i m_i[1 - \sum_j \gamma_i^j b_j] + \delta^2_{i,uncorr} m_i^2} + \sum_j b_j^2, \tag{17}$$

Asymmetrical uncertainties

In case if assumption of the symmetric systematic uncertainty (expressed by Eq.8) is not valid for performed measurement, the χ^2 Eq. 10 can be written in more general form:

$$\chi^2(\vec{m}, \vec{b})_{exp} = \sum_i \frac{(m_i - \mu_i - \sum_j f_i(b_j))^2}{\Delta_i^2} + \sum_j b_j^2. \tag{18}$$

If $f_i(b_j) = \Gamma_i^j b_j$, Eq. 18 again back to Eq. 10. Asymmetric systematic uncertainties can be approximated as:

$$f_i(b_j) = \Gamma_i^j b_j + \omega_i^j b_j^2,\ \omega_i = \frac{\Gamma_i^{j+} + \Gamma_i^{j-}}{2}. \tag{19}$$

χ^2 definition in this case become non-linear.

Iterative combination

For most practical situations the bias is small. Therefore, the expectation m_i can be estimated in an iterative procedure starting from linear formula Eq. 10 and using $m_i = \mu_{i,ave}$. The key for the unbiased result is that the same expectation is used for all measurements.

Similar iterative approach is applied to combine the measurements with asymmetric systematics uncertainties. The first iteration is performed with linearised χ^2 Eq.10 using symmetrised uncertainties Γ (linear part of $f(b_j)$). The next iterations are performed with corrected uncertainties $\Gamma' = \Gamma + \omega * \beta_{ave}$, e.g. correction depends on the systematic shift.

Post-rotation of the systematics

- Sources of the systematic uncertainties before the combination are assumed to be uncorrelated: one nuisance parameter corresponds to one systematic source, total uncertainty is a quadratic sum of all sources. $\Gamma_j b_j$, $\delta^2_{\mathrm{total}} = \sum \Gamma_j^2$
- After the combination systematic sources are not independent: total uncertainty is smaller as quadratic sum of all sources. $\delta^2_{\mathrm{total}} > \sum (\Gamma_j b_j)^2$
- Corresponding correlation matrix can be decomposed to deliver new systematic sources. The number of sources keeped constant, but new systematics are the linear combination of initial systematics. One can not say, which source is which. $\delta^2_{\mathrm{total}} = \sum \Gamma^2_{new,j} b_{new,j}$, $\Gamma_{new} = \sum \Gamma_k l_k$
- The correlation matrix can be rotated such a way, that at least one systematic source have the same origination (e.g. energy scale).

Steering parameters. HAverager

```
&HAverager
  OutputMode   = 'ORTH'
  OutputPrefix = 'Ave'
  OutputFolder = '../output'
  IDebug = 0
  WriteOriginal = .false.
  WriteSysTexTable = .false.
  PostRotateSyst = .true.
&End
```

- `OutputMode` - is the output options for the systematics uncertainties.
  - `'ORTH'` - orthogonal representation
  - `'ORIG'` - original structure of the systematic uncertainties.
- `IDebug` - Debug level. Higher value corresponds to more debug messages.
- `WriteOriginal` - include original information to the output summary (file `ave Ave.dat`)
- `WriteSysTexTable` - write output information about systematic uncertainties in tex format (file `sys.tex`)
- `PostRotateSyst = .true.` keep output systematic uncertainties align to the original sources as much as possible.

Steering parameters. BiasCorrection

```
&BiasCorrection
  AverageType = 'MIXED'
  Iteration = 10
 ! Rescale the stat and uncorr uncertainties separately:
  RescaleStatSep = .false.
 ! Correction of the syst bias for stat errors
  CorrectStatBias = .false.
 ! Keeping the stat errors fixed'
  FixStat = .false.
&End
```

- `AverageType` define the type of the systematic uncertainties
  - `'ADD'` - all systematic uncertainties are processed as additive
  - `'MULT'` - all systematic uncertainties are processed as multiplicative
  - `'MIXED'` - type of systematic uncertainties is taken from the data file

Correlation of systematic uncertainties

- χ^2 definition assume the knowledge about the correlation between sources of systematic uncertainties and across bins. Also it assumes, that uncertainties are Gaussian.
- With respect of these assumptions a certain behaviour of combined results is expected (e.g. Gaussian behaviour of pulls). Significant deviation in this behaviour indicate some problems with considered uncertainties or central values.
- These requirements on the input data can be fulfilled only with intensive communication between fitting and analyses groups.
    - Exact requirements on the data have to be clarified
    - Instruments and instructions, which helps to fulfil these requirements have to be provided

New prospects of HAverager

- The code of the HAverager was cleaned from HERA scripts ("swimming")
- Few parts of the code were optimized with respect of the timing profile. Filling arrays is a bottle-neck
- HAverager as a python library
  - The data-reading and filling of internal variables can be implemented using existing python libraries (pandas, numpy)
  - The fortran code of the HAverager-kernel can be easily compiled as a python library
  - Advantages: still very fast, suitable for the very large data sets (allocatable arrays, no hand-written limitations on the array-length), easily-readable scripts

Python HAverager

```
import numpy as np
import pandas as pd

# Get averager
from  haverage import *

# Read data:
a = pd.read_csv("bla1.csv")
b = pd.read_csv("bla2.csv")

# Prepare arrays of data and uncertainties
........................

# Perform averaging
t,u,f = haverage(d,e,s)

# Print results
print (t)
print (u)
print (f)
```

- Reading data, creating a grid of bins, preparing arrays and printing results are performed with python.

Summary/Plans

- New version of the averaging package is nearly ready. Several new features, such as asymmetric uncertainties, post-rotation of the systematic sources are implemented. HAverager will be available as an open-source package.
- HAverager is a standalone package based on a Fortran code, which can be used as a python library. An example is given.
- HAverager have a certain requirements on the input data, which can be fulfilled only with intensive communication between data-post-processing and data-extractor groups.
- It would be interesting to compare HAverager with much-slower likelihood-based programs. Can we pretend to be suitable for low-statistics data?

Backup

Pulls χ^2

The pull of the central values:

$$p^{i,e} = \frac{\mu^{i,e} - \mu^{i,ave}(1 - \sum_j \gamma^j_{i,e}\beta_{j,ave})}{\sqrt{\Delta^2_{i,e} - \Delta^2_{i,ave}}}, \quad (20)$$

where $\gamma^j_{i,e} = \Gamma^j_{i,e}/\mu_{i,ave}$. This definition is similar to the χ^2 definition, but not summed over bins. These pulls show how the average measurement are shifted compare to individual measurement and also have two contributions, similar to χ^2. The pull for systematic uncertainties can be defined as:

$$p^i = \frac{\beta_{i,ave}}{\sqrt{1 - D^2_{ii}}}. \quad (21)$$

This value shown, how significant was the systematic shifted due to the combination. The large systematic pull suggests, that systematic uncertainty was not correctly estimated.

Sum of several χ^2

The sum of two χ^2:

$$\chi^2(\vec{m},\vec{b})_{sum} = \sum_e \sum_i^{N_M} \frac{\left(m_i - \mu_{i,e} - \sum_j^{N_S} \Gamma^j_{i,e} b_j\right)^2}{\Delta^2_{i,e}} W_{i,e} + \sum_j^{N_S} b_j^2, \qquad (22)$$

- i runs over all measured points N_M
- j runs over all sources of systematic uncertainties N_S
- symbol $W_{i,e}$ is equal to 1 if data set e contributes to a measurement at the point i, otherwise it is 0.
- $\Gamma^i_{j,e}$ equals to 0 if the measurement i from the data set e is insensitive to the systematic source j.

The dimension of $\vec{m_{ave}}$ is equal to dimension of union set of $\vec{m_1}$ and $\vec{m_2}$. e.g. if both experiments measure for the same binning, $N_{M1} = N_{M2} = N_{M,ave} = N_M$. Similarly, for the systematic uncertainties $N_{S,ave} = N_{S1} + N_{S2} - N_{S,common} = N_S$, where $N_{S,common}$ is the number of common systematic error sources for the two measurements.

χ^2-minimization

Since χ^2_{sum} is a quadratic form of $\vec{m}$ and $\vec{b}$, it may be rearranged such that it takes a form similar to Eq. 5.

$$\chi^2(\vec{m},\vec{b}) = \chi^2_{min} + \sum_i^{N_{M,ave}} \frac{\left(m_i - \mu_{i,ave} - \sum_j^{N_{S,ave}} \Gamma^j_{i,ave}(\alpha_j - \alpha_{j,ave})\right)^2}{\Delta^2_{i,ave}} + \\ + \sum_j^{N_{S,ave}} \sum_k^{N_{S,ave}} (\alpha_j - \alpha_{j,ave})(\alpha_k - \alpha_{k,ave})(A'_S)_{jk}, \quad (23)$$

where

- $\mu_{i,ave}$ are average values of measured quantities
- $\Delta_{i,ave}$ are their uncorrelated uncertainties

The values of $\alpha_{j,ave}$, $\Delta_{i,ave}$, $\mu_{i,ave}$ and matrix A'_S are determined by minimization of χ^2 function in Eq. 22 with respect to m_i and b_j.

minimized χ^2

After diagonalizability of matrix A'_S, χ^2 function in Eq. 23 can be re-written in form, similar to Eq. 10:

$$\chi^2(\vec{m}, \vec{b'})_{tot} = \chi^2_{min} + \sum_i^{N_M} \frac{(m_i - \mu_{i,ave} - \sum_j^{N_S} \Gamma^j_{i,ave} b'_j)^2}{\Delta^2_{i,ave}} + \sum_j^{N_S} (b'_j)^2, \qquad (24)$$

where $b'_j = \sum_k U_{jk}(b_k - \beta_{k,ave})D_{jj}$.
The orthogonal matrix U connecting the systematic sources before and after averaging with Eq. 29. Diagonal elements of matrix D shows, how the uncertainties of combined measurement $\Gamma^j_{i,ave}$ are reduced, compared to initial systematic uncertainties.

χ^2-minimization

The minimum of Eq. 22 is found by solving a system of linear equations obtained by requiring $\partial\chi^2/\partial m_i = 0$ and $\partial\chi^2/\partial b_j = 0$ which can be written in matrix form

$$\begin{pmatrix} A_M & A_{SM} \\ (A_{SM})^T & A_S \end{pmatrix} \begin{pmatrix} M_{ave} \\ B_{ave} \end{pmatrix} = \begin{pmatrix} C_M \\ C_S \end{pmatrix} \tag{25}$$

where

- vector M_{ave} corresponds to all measurements
- vector B_{ave} corresponds to all sources of the systematic uncertainties
- matrix A_M has a diagonal structure with $N_{M,ave}$ diagonal elements $A_M^{ii} = \sum_e \frac{W_{i,e}}{\Delta^2_{i,e}}$
- $A_{SM}^{ij} = -\sum_e \frac{\Gamma^j_{i,e}}{\Delta^2 i,e} W_{i,e}$
- $A_S^{ij} = \delta_{ij} + \sum_e \sum_k^{N_M} \frac{\Gamma^k_{i,e}\Gamma^k_{j,e}}{\Delta^2_{k,e}} W_{k,e}$
- $C_M^i = \sum_e \frac{\mu^i_e}{\Delta^2_{i,e}} W_{i,e}$
- $C_S^j = -\sum_e \sum_k^{N_M} \frac{\mu^k_e \Gamma^k_{j,e}}{\Delta^2_{k,e}} W_{k,e}$

Here δ_{ij} is the Kronecker symbol. The matrix A_{SM} has dimension $N_M \times N_S$ while the matrix A_S is quadratic with $N_S \times N_S$ elements.

Average values

Using the method of the Schur complement, the solution is found as:

$$
\begin{aligned}
A'_S &= A_S - (A_{SM})^T A_M^{-1} A_{SM} \\
B_{ave} &= (A'_S)^{-1}(C_S - (A_{SM})^T A_M^{-1} C_M) \\
M_{ave} &= A_M^{-1}(C_M - A_{SM} B_{ave})
\end{aligned}
\tag{26}
$$

Given the components of the vector B_{ave}, $\beta_{j,ave} = \alpha_{j,ave}/\Delta_{\alpha_j}$

Average values

The solution for $\mu_{i,ave}$ can be written in explicit form:

$$\mu_{i,ave} = \frac{\sum_e \left(\mu_{i,e} + \sum_j \Gamma^i_{j,e}\beta_{j,ave}\right) \frac{W_{i,e}}{\Delta^2 i,e}}{\sum_e \frac{W_{i,e}}{\Delta^2_{i,e}}} \tag{27}$$

The uncorrelated uncertainty squared is determined by the inverse of the elements of the diagonal matrix A_M:

$$\Delta^2_{i,ave} = \frac{1}{\sum_e \frac{W_{i,e}}{\Delta^2_{i,e}}} \tag{28}$$

Eq. 27 and 28 reproduce the standard formula for a statistically weighted average of several uncorrelated measurements when all shifts of the systematic error sources are set to zero. The values of $\beta_{i,ave}$ in Eq. 27 show, how the combined measurements $\mu_{i,ave}$ are shifted, compared to initial measurements $\mu_{i,e}$ in terms of systematic uncertainties $\Gamma^j_{i,e}$.

Diagonalization of A'_S

The non-diagonal nature of the matrix A'_S expresses the fact that the original sources of the systematic uncertainties are correlated with each other after averaging. The matrix A'_S can be decomposed to re-express Eq. 10 in terms of diagonalised sources of systematic uncertainties:

$$DD = UA'_S U^{-1} \quad \Gamma_{ave} = A_{SM} A_M^{-1} D^{-1} U^{-1} \tag{29}$$

Here U is an orthogonal matrix composed of the eigenvectors of A'_S, D is a diagonal matrix with corresponding square roots of eigenvalues as diagonal elements and Γ_{ave} represents the sensitivity of the average result to these new sources. Its elements are the $\Gamma^j_{i,ave}$.

The bibliography

- A. Glazov, “Averaging of DIS cross section data,” AIP Conf. Proc. 792 (2005) 237–240.
- H1 and ZEUS collaborations, “Combined measurement and QCD analysis of the inclusive $e^{\pm}p$ scattering cross sections at HERA”, JHEP01 (2010) 109
- H1 Collaboration, “Measurement of the Inclusive ep Scattering Cross Section at Low Q^2 and x at HERA”, Eur. Phys. J. C **63** (2009) 625
- S. Glazov, “Data Combination”, Statistics School, DESY, October 2011. Slides.