Contents

1 Introduction 2
  1.1 Preliminaries ................................................. 2
  1.2 Generalizations and Use-Cases .............................. 3

2 The Likelihood Template 4
  2.1 Index Convention ............................................. 4
  2.2 The Template ................................................ 4
  2.3 Normalization Conventions ................................ 5

3 Interpolation & Constraints 6
  3.1 Interpolation Options ....................................... 7
    3.1.1 Defaults in ROOT 5.32 ................................ 9
  3.2 Constraint Terms (+ global observables and nuisance parameter priors) 10
    3.2.1 Consistent Bayesian and Frequentist modeling ........ 10
    3.2.2 Options for Constraint Terms .......................... 11

4 Usage of the HistFactory 15

5 Usage with RooStats tools 17

6 Examples 18
  6.1 A Simple Example ............................................ 18
  6.2 Multiple signal channels ................................... 19
  6.3 ABCD .......................................................... 19

7 The HistFactory XML Schema 21

8 Manual entries 25
1 Introduction

The HistFactory is a tool to build parametrized probability density functions (pdfs) in the RooFit/RooStats framework based on simple ROOT histograms organized in an XML file. The pdf has a restricted form, but it is sufficiently flexible to describe many analyses based on template histograms. The tool takes a modular approach to build complex pdfs from more primitive conceptual building blocks. The resulting PDF is stored in a RooWorkspace which can be saved to and read from a ROOT file.

1.1 Preliminaries

Let us begin by considering the simple case of a single channel with one signal and one background contribution and no systematics based on the discriminating variable is \( x \). While we will not continue with this notation, let us start with the familiar convention where the number of signal events is denoted as \( S \) and the number of background events as \( B \). Similarly, denote the signal and background “shapes” as \( f_S(x) \) and \( f_B(x) \) and note the these are probability density functions normalized so that \( \int dx f(x) = 1 \). It is common to introduce a “signal strength” parameter \( \mu \) such that \( \mu = 0 \) corresponds to the background-only hypothesis and \( \mu = 1 \) corresponds to the nominal signal+background hypothesis. This continuous parameter \( \mu \) is our parameter of interest.

Now we ask what the probability model is for obtaining \( n \) events in the data where the discriminating variable for event \( e \) has a value \( x_e \); thus the full dataset will be denoted \( \{x_1 \ldots x_n\} \). First one must include the Poisson probability of obtaining \( n \) events when \( \mu_S + B \) are expected. Secondly, one must take into account the probability density of obtaining \( x_e \) based on the relative mixture \( f_S(x) \) and \( f_B(x) \) for a given value of \( \mu \). Putting those two ingredients together one obtains what statisticians call a “marked Poisson model”:

\[
\mathcal{P}(\{x_1 \ldots x_n\}|\mu) = \text{Pois}(n|\mu_S + B) \left[ \prod_{e=1}^{n} \frac{\mu_S f_S(x_e) + B f_B(x_e)}{\mu_S + B} \right].
\]  

(1)

If one imagines the data as being fixed, then this equation depends on \( \mu \) and is called the likelihood function \( L(\mu) \). Simply taking the logarithm of the equation above and remembering that \( \text{Pois}(n|\nu) = \nu^n e^{-\nu}/n! \) gives us a familiar formula referred to by physicists as an “extended maximum likelihood fit”:

\[
-\ln L(\mu) = -n \ln(\mu_S + B) + (\mu S + B) + \ln n! - \sum_{e=1}^{n} \ln \left[ \frac{\mu_S f_S(x_e) + B f_B(x_e)}{\mu S + B} \right]
\]

\[
= (\mu S + B) + \ln n! - \sum_{e=1}^{n} \ln [\mu S f_S(x_e) + B f_B(x_e)].
\]  

(2)

Since HistFactory is based on histograms, it is natural to think of the binned equivalent of the probability model above. Denoted the signal and background histograms as \( \nu^\text{sig}_b \) and \( \nu^\text{bkg}_b \), where \( b \) is the bin index and the histograms contents correspond to the number of events expected in the data. We can relate the bin \( \nu_b \) and the shape \( f(x) \) via

\[
f_S(x_e) = \frac{\nu^\text{sig}_b}{S \Delta b_e} \quad \text{and} \quad f_B(x_e) = \frac{\nu^\text{bkg}_b}{B \Delta b_e},
\]  

(3)

where \( b_e \) is the index of the bin containing \( x_e \) and \( \Delta b_e \) is the width of that same bins. Note, because the \( f(x) \) are normalized to unity we have \( S = \sum_b \nu^\text{sig}_b \) and \( B = \sum_b \nu^\text{bkg}_b \).
Formally one can either write the probability model in terms of a product over Poisson distributions for each bin of the histogram, or one can also continue to use the unbinned expression above recognizing that the shapes \( f(x) \) look like histograms (ie. they are discontinuous at the bin boundaries and constant between them). Technically, the HistFactory makes a model that looks more like the unbinned expression with a single RooAbsPdf that is “extended” with a discontinuous shape in \( x \). Nevertheless, it can be more convenient to express the model in terms of the individual bins. Then we have

\[
P(n_b|\mu) = \text{Pois}(n_{\text{tot}}|\mu S + B) \prod_{b \in \text{bins}} \frac{\mu \nu_{\text{sig}} + \nu_{\text{bkg}}}{\mu S + B} = N_{\text{comb}} \prod_{b \in \text{bins}} \text{Pois}(n_b|\mu \nu_{\text{sig}} + \nu_{\text{bkg}}), \tag{4}
\]

where \( n_b \) is the data histogram and \( N_{\text{comb}} \) is a combinatorial factor that can be neglected since it is constant. Similarly, denote the data histogram is \( n_b \).

1.2 Generalizations and Use-Cases

Based on the discussion above, we want to generalize the model in the following ways:

- Ability to include multiple signal and background samples
- Ability to include unconstrained scaling of the normalization of any sample (as was done with \( \mu \))
- Ability to parametrize variation in the normalization of any sample due to some systematic effect
- Ability to parameterize variations in the shape of any sample due to some systematic effect
- Ability to include bin-by-bin statistical uncertainty on the normalization of any sample
- Ability to incorporate an arbitrary contribution where each bin’s content is parametrized individually
- Ability to combine multiple channels (regions of the data defined by disjoint event selections) and correlate the parameters across the various channels
- Ability to use the combination infrastructure to incorporate control samples for data-driven background estimation techniques
- Ability to reparametrize the model

<table>
<thead>
<tr>
<th>Normalization Variation</th>
<th>Constrained</th>
<th>Unconstrained</th>
</tr>
</thead>
<tbody>
<tr>
<td>OverallSys (( \eta_{cs} ))</td>
<td>NormFactor (( \phi_p ))</td>
<td></td>
</tr>
<tr>
<td>Coherent Shape Variation</td>
<td>HistoSys ( \sigma_{cab} )</td>
<td>-</td>
</tr>
<tr>
<td>Bin-by-bin variation</td>
<td>ShapeSys &amp; StatError ( \gamma_{cb} )</td>
<td>ShapeFactor ( \gamma_{csb} )</td>
</tr>
</tbody>
</table>

Table 1: Conceptual building blocks for constructing more complicated PDFs: parameters.
2  The Likelihood Template

2.1 Index Convention

We will use the following mnemonic index conventions:

- \( e \in \) events
- \( b \in \) bins
- \( c \in \) channels
- \( s \in \) samples
- \( p \in \) parameters

We define the following subsets of parameters

\[ N = \{ \phi_p \} \]

the unconstrained normalization factors (i.e. \texttt{NormFactor}),

\[ S = \{ \alpha_p \} \]

the parameters associated to systematic that have external constraints (i.e. \texttt{OverallSys} and \texttt{HistoSys}),

\[ \Gamma = \{ \gamma_{cb} \} \]

the bin-by-bin uncertainties with constraints (statistical errors, \texttt{ShapeSys} but not those associated to an unconstrained \texttt{ShapeFactor}). We also use greek symbols for parameters of the model and roman symbols for observable quantities with a frequentist notion of probability.

2.2 The Template

The parametrized probability density function constructed by the \texttt{HistFactory} is of a concrete form, but sufficiently flexible to describe many analyses based on template histograms. In general, the \texttt{HistFactory} produces probability density functions of the form

\[
P(n_c, x_e, a_p | \phi_p, \alpha_p, \gamma_b) = \prod_{c \in \text{channels}} \text{Pois}(n_c | \nu_c) \prod_{e=1}^{n_c} f_c(x_e | \alpha) \cdot G(L_0 | \lambda, \Delta_L) \cdot \prod_{p \in S + \Gamma} P_p(a_p | \alpha_p) \quad (5)
\]

where \( P_p(a_p | \alpha_p) \) is a constraint term describing an auxiliary measurement \( a_p \) that constrains the nuisance parameter \( \alpha_p \) (see Section 3.2). Denote the bin containing \( x_e \) as \( b_e \). We have the following expression for the expected (mean) number of events in a given bin

\[
\nu_{cb}(\phi_p, \alpha_p, \gamma_b) = \lambda_{cs} \gamma_{cb} \phi_{cs}(\alpha) \eta_{cs}(\alpha) \sigma_{csb}(\alpha), \quad (6)
\]

where the meaning of the various terms is described below and the specific interpolation algorithms are described in Section 3.1. The mean number of events in each bin implies the following probability density

\[
f_c(x_e | \phi_p, \alpha_p, \gamma_b) = \frac{\nu_{cb}}{\nu_c} \quad \text{with} \quad \nu_c = \sum_{b \in \text{bins of channel } c} \nu_{cb} \quad (7)
\]

It is perhaps more convenient to think of the likelihood as a product over bins

\[
P(n_{cb}, a_p | \phi_p, \alpha_p, \gamma_b) = \prod_{c \in \text{channels}} \prod_{b \in \text{bins}} \text{Pois}(n_{cb} | \nu_{cb}) \cdot G(L_0 | \lambda, \Delta_L) \cdot \prod_{p \in S + \Gamma} P_p(a_p | \alpha_p)
\]

- \( \lambda_{cs} \) - luminosity parameter for a given channel and sample. Within a given channel this parameter is a common luminosity parameter for all the samples that include luminosity uncertainty (i.e. \texttt{NormalizeByTheory}="True"). For all the samples with \texttt{NormalizeByTheory}="False" it is fixed to the nominal luminosity \( \lambda_{cs} = L_0 \).
- $\gamma_{cb}e$ - Bin-by-bin scale factor used for statistical uncertainties, bin-by-bin shape systematics (ShapeSys), and data-driven shape extrapolations (ShapeFactor). For statistical errors, the $\gamma_{cb}e$ is shared for all the samples in the channel (ie. subscript $s$ can be omitted). For samples that do not have any bin-by-bin scale factors $\gamma_{cb}e = 1$.

- $\phi_{cs}$ - Product of unconstrained normalization factors for a given sample within a given channel. These typically include the parameter of interest, eg. the signal cross-section or branching ratio.

$$\phi_{cs} = \prod_{p \in N_c} \phi_p$$

- $\eta_{cs}(\alpha)$ - The parametrized normalization uncertainties (ie. OverallSys) for a given sample within a given channel (a factor around 1).

- $\sigma_{csb}e$ - The parametrized histogram (ie. the nominal histogram and the HistoSys) for a given sample within a given channel.

### 2.3 Normalization Conventions

The nominal and variational histograms should all have the same normalization convention. There are a few conventions possible:

**Option 1:**
- Lumi="XXX" in the measurement XML’s element, where XXX is in fb$^{-1}$
- Histograms bins have units of fb
- Some samples have NormFactor that are all relative to prediction (eg. 1 is the nominal prediction)

**Option 2:**
- Lumi="1." in the measurement XML’s element
- Histograms are normalized to unity
- Each sample has a NormFactor that is the expected number of events in data

**Option 3:**
- Lumi="1" in the measurement XML’s element
- Histograms bins have units of number of events expected in data
- Some samples have NormFactor that are all relative to prediction (eg. 1 is the nominal prediction)

It’s up to you. In the end, the expected number is the product $Lumi \times NormFactor(s) \times BinContent$ corresponding to $\lambda_{cs} \phi_{cs} \sigma_{csb}$. 
3 Interpolation & Constraints

The treatment of systematic uncertainties is subtle, particularly when one wishes to take into account the correlated effect of multiple sources of systematic uncertainty across many signal and background samples. The most important conceptual issue is that we separate the source of the uncertainty (for instance the uncertainty in the calorimeter’s response to jets) from its effect on an individual signal or background sample (e.g. the change in the acceptance and shape of a W+jets background). In particular, the same source of uncertainty has a different effect on the various signal and background samples. The most important conceptual issue is that we separate the source of the systematic, not the resulting effect.

Once one has tabulated the effects of the individual sources of systematic uncertainty as above, one must address two related issues to form a likelihood parametrized with continuous nuisance parameters. First, one must provide an interpolation algorithm to interpolate to define \( \eta_s(\alpha) \) and \( \sigma_{sb}(\alpha) \). Secondly, one must incorporate constraint terms on the \( \alpha_p \) to reflect that the uncertain parameter has been estimated with some uncertainty by an auxiliary measurement. A strength of the histogram template based approach (compared to parametrized analytic functions) is that the effect of individual systematics are tracked explicitly; however, the ambiguities associated to the interpolation and constraints are a weakness.

\[ 1 \text{Here we suppress the channel index } c \text{ on } \eta_{sc} \text{ and } \sigma_{scb} \]
3.1 Interpolation Options

For each sample, one can interpolate and extrapolate from the nominal prediction \( \eta_s^0 = 1 \) and the variations \( \eta_{ps}^+ \) to produce a parametrized \( \eta_s(\alpha) \). Similarly, one can interpolate and extrapolate from the nominal shape \( \sigma_{sb}^0 \) and the variations \( \sigma_{psb}^\pm \) to produce a parametrized \( \sigma_{sb}(\alpha) \). We choose to parametrize \( \alpha_p \) such that \( \alpha_p = 0 \) is the nominal value of this parameter, \( \alpha_p = \pm 1 \) are the "\( \pm 1 \sigma \) variations". Needless to say, there is a significant amount of ambiguity in these interpolation and extrapolation procedures and they must be handled with care. In the future the HistFactory may support other types of shape interpolation, but as of ROOT 5.32 the shape interpolation is a 'vertical' style interpolation that is treated independently per-bin. Four interpolation strategies are described below and can be compared in Fig 1.

**Piecewise Linear (InterpCode=0)**

The piecewise-linear interpolation strategy is defined as

\[
\eta_s(\alpha) = 1 + \sum_{p \in \text{Syst}} I_{\text{lin}}(\alpha_p; 1, \eta_{ps}^+, \eta_{ps}^-)
\]

and for shape interpolation it is

\[
\sigma_{sb}(\alpha) = \sigma_{sb}^0 + \sum_{p \in \text{Syst}} I_{\text{lin}}(\alpha_p; \sigma_{sb}^0, \sigma_{psb}^+, \sigma_{psb}^-)
\]

with

\[
I_{\text{lin}}(\alpha; I^0, I^+, I^-) = \begin{cases} 
\alpha(I^+ - I^0) & \alpha \geq 0 \\
\alpha(I^0 - I^-) & \alpha < 0 
\end{cases}
\]

**Pros:** This approach is the most straightforward of the interpolation strategies.

**Cons:** It has two negative features. First, there is a kink (discontinuous first derivative) at \( \alpha = 0 \) (see Fig 1(b-d)), which can cause some difficulties for numerical minimization packages such as Minuit. Second, the interpolation factor can extrapolate to negative values. For instance, if \( \eta^\pm = 0.5 \) then we have \( \eta(\alpha) < 0 \) when \( \alpha < -2 \) (see Fig 1(c)).

Note that one could have considered the simultaneous variation of \( \alpha_p \) and \( \alpha_{p}' \) in a multiplicative way (see for example, Fig 2). The multiplicative accumulation is not an option currently.

Note that this is the default convention for \( \sigma_{sb}(\alpha) \) (ie. HistSys).

**Piecewise Exponential (InterpCode=1)**

The piecewise exponential interpolation strategy is defined as

\[
\eta_s(\alpha) = \prod_{p \in \text{Syst}} I_{\text{exp}}(\alpha_p; 1, \eta_{ps}^+, \eta_{ps}^-)
\]

and for shape interpolation it is

\[
\sigma_{sb}(\alpha) = \sigma_{sb}^0 \prod_{p \in \text{Syst}} I_{\text{exp}}(\alpha_p; \sigma_{sb}^0, \sigma_{psb}^+, \sigma_{psb}^-)
\]

with

\[
I_{\text{exp}}(\alpha; I^0, I^+, I^-) = \begin{cases} 
(I^+ / I^0)^\alpha & \alpha \geq 0 \\
(I^- / I^0)^{-\alpha} & \alpha < 0 
\end{cases}
\]
Pros: This approach ensures that $\eta(\alpha) \geq 0$ (see Fig 1(c)) and for small response to the uncertainties it has the same linear behavior near $\alpha \sim 0$ as the piecewise linear interpolation (see Fig 1(a)).

Cons: It has two negative features. First, there is a kink (discontinuous first derivative) at $\alpha = 0$, which can cause some difficulties for numerical minimization packages such as Minuit. Second, for large uncertainties it develops a different linear behavior compared to the piecewise linear interpolation. In particular, even if the systematic has a symmetric response (ie. $\eta^+ - 1 = 1 - \eta^-$) the interpolated response will develop a kink for large response to the uncertainties (see Fig 1(c)).

Note that the one could have considered the simultaneous variation of $\alpha_p$ and $\alpha_p'$ in an additive way, but this is not an option currently.

Note, that when paired with a Gaussian constraint on $\alpha$ this is equivalent to linear interpolation and a log-normal constraint in $\ln(\alpha)$. This is the default strategy for normalization uncertainties $\eta_s(\alpha)$ (ie. OverallSys) and is the standard convention for normalization uncertainties in the LHC Higgs Combination Group. In the future, the default may change to the Polynomial Interpolation and Exponential Extrapolation described below.

Quadratic Interpolation and Linear Extrapolation (InterpCode=2)

The quadratic interpolation and linear extrapolation strategy is defined as

$$\eta_s(\alpha) = 1 + \sum_{p \in \text{Syst}} I_{\text{quad.\,lin}.}(\alpha_p; 1, \eta^+_s, \eta^-_s)$$

and for shape interpolation it is

$$\sigma_{sb}(\alpha) = \sigma^0_{sb} + \sum_{p \in \text{Syst}} I_{\text{quad.\,lin}.}(\alpha_p; \sigma^0_{sb}, \sigma^+_p, \sigma^-_p)$$

with

$$I_{\text{quad.\,lin}.}(\alpha; I^0, I^+, I^-) = \begin{cases} (b + 2a)(\alpha - 1) & \alpha > 1 \\ a\alpha^2 + b\alpha & |\alpha| \leq 1 \\ (b - 2a)(\alpha + 1) & \alpha < -1 \end{cases}$$

and

$$a = \frac{1}{2}(I^+ + I^-) - I^0 \quad \text{and} \quad b = \frac{1}{2}(I^+ - I^-).$$

Pros: This approach avoids the kink (discontinuous first derivative) at $\alpha = 0$ (see middle panel of Fig 1), which can cause some difficulties for numerical minimization packages such as Minuit.

Cons: It has a few negative features. First, in the case that both the response to both positive and negative variations have the same sign of effect relative to the nominal (ie. $(\eta^+ - 1)(\eta^- - 1) > 0$), the quadratic interpolation can lead to an an intermediate value with the opposite effect. For example, Fig 1(b) shows a case where $\eta(\alpha = -0.3) < 1$ while $\eta^\mp > 0$. Second, when the positive and negative variations have opposite signs, the extrapolation can reverse the trend. For example, Fig 1(d) shows an example for $\eta^- = 0.95$ and $\eta^+ = 1.5$ where for $\alpha \lesssim 1.5$ we have the reversal $\eta(\alpha) > 1$. Third, the interpolation factor can extrapolate to negative values. For instance, if $\eta^- = 0.5$ then we have $\eta(\alpha) < 0$ when $\alpha < -2$ (see Fig 1(c)).
Note that one could have considered the simultaneous variation of $\alpha_p$ and $\alpha_p'$ in a multiplicative way (see for example, Fig 2). The multiplicative accumulation is not an option currently.

**Polynomial Interpolation and Exponential Extrapolation (InterpCode=4)**

The strategy of this interpolation option is to use the piecewise exponential extrapolation as above with a polynomial interpolation that matches $\eta(\alpha = \pm \alpha_0)$, $d\eta/d\alpha|_{\alpha=\pm\alpha_0}$, and $d^2\eta/d\alpha^2|_{\alpha=\pm\alpha_0}$ and the boundary $\pm \alpha_0$ is defined by the user (with default $\alpha_0 = 1$).

\[
\eta_s(\alpha) = \prod_{p \in \text{Syst}} I_{\text{poly|exp.}}(\alpha_p; 1, \eta_{\alpha_p}^+, \eta_{\alpha_p}^-, \alpha_0)
\]

with

\[
I_{\text{poly|exp.}}(\alpha; I^0, I^+, I^-, \alpha_0) = \begin{cases} 
(I^+/I_0)^{\alpha} & \alpha \geq \alpha_0 \\
1 + \sum_{i=1}^{6} a_i \alpha^i & |\alpha| < \alpha_0 \\
(I^-/I_0)^{-\alpha} & \alpha \leq -\alpha_0
\end{cases}
\]

and the $a_i$ are fixed by the boundary conditions described above.

**Pros:** This approach avoids the kink (discontinuous first and second derivatives) at $\alpha = 0$ (see Fig 1(b-d)), which can cause some difficulties for numerical minimization packages such as Minuit. This approach ensures that $\eta(\alpha) \geq 0$ (see Fig 1(c)).

**Note:** This option is not available in ROOT 5.32.00, but is available for normalization uncertainties (OverallSys) in the subsequent patch releases. In future releases, this may become the default.

### 3.1.1 Defaults in ROOT 5.32

The default strategy for normalization uncertainties $\eta_s(\alpha)$ (ie. **OverallSys**) is the piecewise exponential option and it is the standard convention for normalization uncertainties in the LHC Higgs Combination Group.

The default convention for $\sigma_{sb}(\alpha)$ (ie. **HistoSys**) is the piecewise linear option.

The code for $\eta_s(\alpha)$ can be found here:


The code for $\sigma_{sb}(\alpha)$ can be found here:

http://root.cern.ch/root/html532/src/PiecewiseInterpolation.cxx.html
3.2 Constraint Terms (+ global observables and nuisance parameter priors)

3.2.1 Consistent Bayesian and Frequentist modeling

The variational estimates $\eta^\pm$ and $\sigma^\pm$ correspond to so-called “±1σ variations” in the source of the uncertainty. Here we are focusing on the source of the uncertainty, not its affect on rates and shapes. For instance, we might say that the jet energy scale has a 10% uncertainty.\(^2\) This is common jargon, but what does it mean? The most common interpretation of this statement is that the uncertain parameter $\alpha_p$ (e.g. the jet energy scale) has a Gaussian distribution. However, this way of thinking is manifestly bayesian. If the parameter was estimated from an auxiliary measurement, then it is the PDF for that measurement that we wish to include in our probability model. In the frequentist way of thinking, the jet energy scale has an unknown true value and upon repeating the experiment many times the auxiliary measurements estimating the jet energy scale would fluctuate randomly about this true value. To aid in this subtle distinction, we use greek letters for the parameters (e.g. $\alpha_p$) and roman letters for the auxiliary measurements $a_p$.

\(^2\)Without loss of generality, we choose to parametrize $\alpha_p$ such that $\alpha_p = 0$ is the nominal value of this parameter, $\alpha_p = \pm 1$ are the “±1σ variations”.
Figure 2: The upper-most curve corresponds to $\eta_1^+ = (\eta_1^+)^{\alpha_1} (\eta_2^+)^{\alpha_2}$ (as in the exponential interpolation option). The middle surface corresponds to $\eta_1 = 1 + \eta_1^+ \alpha_1 + \eta_2^+ \alpha_2$ (as in the linear interpolation option). The lowest surface corresponds to $\eta_1 = 1 + \eta_1^+ \alpha_1 \cdot \eta_2^+ \alpha_2$ (currently not an option). The left frame has limits correspond to $\alpha_{1,2} \in [0, 3]$ and $\eta(\alpha_1, \alpha_2) \in [0, 5]$ and $\eta_1^+ = \eta_2^+ = 1.1$ (eg. a 10% relative uncertainty). The right frame has limits correspond to $\alpha_{1,2} \in [0, 3]$ and $\eta(\alpha_1, \alpha_2) \in [0, 5]$ and $\eta_1^+ = \eta_2^+ = 1.5$ (eg. a 50% relative uncertainty).

Furthermore, we interpret the “±1σ” variation in the frequentist sense, which leads to the constraint term $G(a_{\mu}|\alpha_{\mu}, 1)$. Then, we can pair the resulting likelihood with some prior on $\alpha_{\mu}$ to form a bayesian posterior if we wish.

It is worth mentioning here that the constraint terms are idealized versions of the auxiliary measurements. In reality, the measurements that were used to estimate the uncertainty in a quantity such as the jet energy scale are actually quite complex. Ideally, one would include the full likelihood function for those auxiliary measurements into the probability model, but that is often impractical. To the extent that the likelihood resulting from the auxiliary measurement is in the Gaussian regime, then this idealization is not a bad approximation.

It is often advocated that a “log-normal” or “gamma” distribution for $\alpha_{\mu}$ is more appropriate. Here we must take some care to build a probability model that can maintain a consistent interpretation in bayesian a frequentist settings. This will be discussed in the subsections below. Table 4 summarizes a few consistent treatments of the frequentist pdf, the likelihood function, a prior, and the resulting posterior.

Finally, it is worth mentioning that the uncertainty on some parameters is not the result of an auxiliary measurement – so the constraint term idealization, it is not just a convenience, but a real conceptual leap. This is particularly true for theoretical uncertainties from higher-order corrections or renormalization and factorization scale dependence. In these cases a formal frequentist analysis would not include a constraint term for these parameters, and the result would simply depend on their assumed values. As this is not the norm, we can think of reading Table 4 from right-to-left with a subjective Bayesian prior $\pi(\alpha)$ being interpreted as coming from a fictional auxiliary measurement.

3.2.2 Options for Constraint Terms

Gaussian Constraint

The Gaussian constraint for $\alpha_{\mu}$ corresponds to the familiar situation. It is a good approximation of the auxiliary measurement when the likelihood function for $\alpha_{\mu}$ from that
For instance, energy scales, reconstruction efficiencies, and background normalizations must be attributed some positive probability to the unphysical regions. Technically, RooFit’s PDF classes (and thus \( a \)) are more directly related – in particular \( \eta(\alpha) \) is a linear function. Consider a background that is estimated with 50% uncertainty, then for \( \alpha \) normalization uncertainties where \( \eta(\alpha) = 1 + n_p \) would correspond to an unphysical negative jet energy scale. One can also consider normalization uncertainties where \( \alpha \) is a number that is extracted from the data and thus its distribution has a frequentist interpretation. In the RooFit workspace produced by HistFactory, this variable has a name like \texttt{nom\_alpha\_<name>} and it is included in the \	exttt{ModelConfig}’s list of GlobalObservables. We chose to scale \( \alpha \) (and thus \( a \)) so that the distribution has unit variance: \( G(a \mid \alpha_p, 1) \). Note that if we assume the true value \( \alpha_p \neq 0 \) and we sample \( a \) via (toy) Monte Carlo techniques, the distribution of \( a \) will not have a mean of 0.

\[
G(a \mid \alpha_p, \sigma_p) = \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left[-\frac{(a - \alpha_p)^2}{2\sigma_p^2}\right]
\]  

(21)

with \( \sigma_p = 1 \) by default.

Note that the PDF of \( a \) and the likelihood for \( \alpha_p \) are positive for all values. Thus if \( \alpha_p \) represents a shifted and rescaled version of a more physical parameter that is bounded, then the Gaussian distribution is attributing some positive probability to the unphysical regions. For instance, energy scales, reconstruction efficiencies, and background normalizations must be \( \geq 0 \). Consider a jet energy scale that is estimated with 25% uncertainty, then \( \alpha < -4 \) would correspond to an unphysical negative jet energy scale. One can also consider normalization uncertainties where \( \alpha \) and \( \eta(\alpha) \) are more directly related – in particular \( \eta(\alpha) \) is a linear function. Consider a background that is estimated with 50% uncertainty, then for \( \alpha < -2 \) will correspond to a negative background estimate, and we will have \( \eta(\alpha < 2) < 0 \).

Technically, RooFit’s PDF classes (\texttt{RooGaussian} in this case) make sure that the PDF is normalized to unity within the range of the observable (in this case \( a \)). So the technical implementation will actually correspond to a truncated and renormalized Gaussian (the default range for \( a \) is \([-5, 5]\)).

**Poisson (“Gamma”) constraint**

When the auxiliary measurement is actually based on counting events in a control region (eg. a Poisson process), a more accurate to describe the auxiliary measurement with a Poisson distribution. It has been shown that the truncated Gaussian constraint can lead to undercoverage (overly optimistic) results, which makes this issue practically relevant. Table 4 shows that a Poisson PDF together with a uniform prior leads to a gamma posterior, thus this type of constraint is often called a “gamma” constraint. This is a bit unfortunate since the gamma distribution is manifestly Bayesian and with a different choice of prior, one might not arrive at a gamma posterior. When dealing with the Poisson constraint, it is no longer convenient to work with our conventional scaling for \( \alpha_p \) which can be negative. Instead, it is more natural to think of the number of events measured in the auxiliary measurement \( n_p \) and the mean of the Poisson parameter. This information is not

<table>
<thead>
<tr>
<th>PDF</th>
<th>Likelihood ( \propto )</th>
<th>Prior ( \pi_0 )</th>
<th>Posterior ( \pi )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( G(a \mid \alpha_p, \sigma_p) )</td>
<td>( G(a \mid \alpha_p, \sigma_p) )</td>
<td>( \pi_0(a \mid \alpha_p) \propto \text{const} )</td>
<td>( G(a \mid \alpha_p, \sigma_p) )</td>
</tr>
<tr>
<td>( \text{Pois}(n \mid \tau_p \beta_p) )</td>
<td>( P_\Gamma(\beta_p \mid A = \tau_p; B = 1 + n_p) )</td>
<td>( \pi_0(\beta_p) \propto \text{const} )</td>
<td>( P_\Gamma(\beta_p \mid A = \tau_p; B = 1 + n_p) )</td>
</tr>
<tr>
<td>( P_{LN}(n \mid \beta_p, \sigma_p) )</td>
<td>( \beta_p \cdot P_{LN}(n \mid \beta_p, \sigma_p) )</td>
<td>( \pi_0(\beta_p) \propto \text{const} )</td>
<td>( P_{LN}(\beta_p \mid n, \sigma_p) )</td>
</tr>
<tr>
<td>( P_{LN}(n \mid \beta_p, \sigma_p) )</td>
<td>( \beta_p \cdot P_{LN}(n \mid \beta_p, \sigma_p) )</td>
<td>( \pi_0(\beta_p) \propto 1/\beta_p )</td>
<td>( P_{LN}(\beta_p \mid n, \sigma_p) )</td>
</tr>
</tbody>
</table>

Table 4: Table relating consistent treatments of PDF, likelihood, prior, and posterior for nuisance parameter constraint terms.
usually available, instead one usually has some notion of the relative uncertainty in the parameter \( \sigma_{rel} \) (e.g. a the jet energy scale is known to 10%). In order to give some uniformity to the different uncertainties of this type and think of relative uncertainty, the nominal rate is factored out into a constant \( \tau_p \) and the mean of the Poisson is given by \( \tau_p \beta_p \).

\[
\text{Pois}(n_p | \tau_p \beta_p) = \frac{(\tau_p \beta_p)^{n_p} e^{-\tau_p \beta_p}}{n_p!}
\]

Here we can use the fact that \( \text{Var}[n_p] = \sqrt{\tau_p \beta_p} \) and reverse engineer the nominal auxiliary measurement

\[
n_p^0 = \tau_p = \left( \frac{1}{\sigma_{rel}^2} \right)^2.
\]

where the superscript 0 is to remind us that \( n_p \) will fluctuate in repeated experiments but \( n_p^0 \) is the value of our measured estimate of the parameter.

Thus the nominal situation corresponds to \( \beta_p = 1 \) and the “\( \pm 1\sigma \) variations” (which is now ambiguous) conventionally correspond to \( \beta_p = 1 \pm \sigma_{rel}^2 = 1 \pm \tau_p^{-1/2} \). It is more convenient to modify the constraint term while keeping the interpolation \( \eta(\alpha) \) fixed, thus we introduce the linear relationship that satisfies \( \alpha(\beta = 1) = 0 \) and \( \alpha(\beta = 1 \pm \tau_p^{-1/2}) = \pm 1 \)

\[
\alpha_p(\beta_p) = \sqrt{\tau_p} (\beta_p - 1)
\]

One important thing to keep in mind is that there is only one constraint term per nuisance parameter, so there must be only one \( \sigma_{rel}^2 \) per nuisance parameter. This \( \sigma_{rel}^2 \) is related to the fundamental uncertainty in the source and we cannot infer this from the various response terms \( \eta_{ps} \) or \( \sigma_{pub} \). In the XML this is not a property of a channel, but of a measurement and it is encoded in a term like

```
<ConstraintTerm Type="Gamma" RelativeUncertainty="0.1">JES</ConstraintTerm>
```

Another technical difficulty is that the Poisson distribution is discrete. So if one were to say the relative uncertainty was 30%, then we would find \( n_p^0 = 11.11... \), which is not an integer. Rounding \( n_p \) to the nearest integer while maintaining \( \tau_p = (1/\sigma_{rel}^2)^2 \) will bias the maximum likelihood estimate of \( \beta_p \) away from 1. As of ROOT 5.32 the `ConstraintTerm Type="Gamma"` used the `RooGamma` (which generalizes more continuously) with

\[
P_{\Gamma}(\beta_p | A = \tau_p, B = n_p - 1) = A(A\beta_p)^B e^{-A\beta_p}/\Gamma(B)
\]

The implementation works fine for likelihood fits, bayesian calculations, and frequentist techniques based on asymptotic approximations, but it does not offer a consistent treatment of the pdf for the global observable \( n_p \) that is needed for techniques based on Monte Carlo techniques. In future versions of ROOT, the constraint will probably be replaced with `RooPoisson` with an option `setNoRounding(true)`.

**Log-normal constraint**

From Eadie et al., “The log-normal distribution represents a random variable whose logarithm follows a normal distribution. It provides a model for the error of a process involving many small multiplicative errors (from the Central Limit Theorem). It is also appropriate when the value of an observed variable is a random proportion of the previous observation.”

As in the case of the “Gamma” constraints we need to reparametrize to a nuisance parameter \( \beta_p \) that is positive and centered around 1. Again we use \( \alpha \) for the response of the systematics and relate the two via

\[
\alpha_p(\beta_p) = \sqrt{\tau_p} (\beta_p - 1)
\]
And the equivalent global observable is

\[ n_p^0 = \tau_p = \left(1/\sigma_{p}^{\text{rel}}\right)^2. \]  

(27)

\[ P_{\text{LN}}(n_p|\beta_p, \kappa_p) = \frac{1}{\sqrt{2\pi} \ln \kappa_{n_p}} \exp \left[ -\frac{\ln(n_p/\beta_p)^2}{2(\ln \kappa_p)^2} \right] \]  

(28)

(blue curve in Fig. 3(a)).

\[ L(\beta_p) = \frac{1}{\sqrt{2\pi} \ln \kappa_{n_p}} \exp \left[ -\frac{\ln(n_p/\beta_p)^2}{2(\ln \kappa_p)^2} \right] \]  

(29)

(red curve in Fig. 3(b)).

\[ \pi(\beta_p) \propto \pi_0(\beta_p) \frac{1}{\sqrt{2\pi} \ln \kappa_{n_p}} \exp \left[ -\frac{\ln(n_p/\beta_p)^2}{2(\ln \kappa_p)^2} \right] \]  

(30)

When paired with an “ur-prior” \( \pi_0(\beta_p) \propto 1/\beta_p \) (green curve in Fig. 3(b)), this results in a posterior distribution that is also of a log-normal form for \( \beta_p \) (blue curve in Fig. 3(b)).

Figure 3: The lognormal constraint term: (left) the pdf for the global observable \( a_p \) and (right) the likelihood function, the posterior based on a flat prior on \( \beta_p \), and the posterior based on a \( 1/\beta_p \) prior.

```cpp
void LogNormalExample(double relUncert = 0.8) {
  double kappa = 1 + relUncert;
  RooWorkspace w("w");

  // make the PDF (likelihood term) which has a log-normal distribution for x
  w.factory(Form("Lognormal::pdf(x[1,0,10],mu[1,0,10],kappa[%f])",kappa));

  // make the ur-prior that gives a log-normal posterior
  w.factory("EXPR::urprior('1/mu',mu)");

  // Multiply prior*likelihood to get posterior numerically
  w.factory("PROD::posterior_numerical(pdf,urprior)");

  // check against analytic derivation of posterior
  w.factory("Lognormal::posterior_analytical(mu,x,kappa)"));

  TCanvas* c1 = new TCanvas();
  c1->Divide(2);

  // The PDF for the measurement
  c1->cd(1);
  RooPlot* xframe = w.var("x")->frame();
```
4 Usage of the HistFactory

ROOT installation

Download, install, and setup ROOT v5.28 or greater. It is recommended to use one of the patch releases of v5.28 as the "standard form" described below was not available before the patch releases.

```
cd $ROOTSYS
source bin/thisroot.sh
```

This will setup your MANPATH environment variable so that you can use the command line help.

**prepareHistFactory**

```
man prepareHistFactory
prepareHistFactory
```

The command line executable `prepareHistFactory [dir_name]` is a simple script that prepares a working area (and creates the directory dir_name if specified). Within the directory dir_name, it creates a results/, data/, and config/ directory relative to the given path. It also copies the HistFactorySchema.dtd and example XML files into the config/ directory. Additionally, it copies a root file into the data/ directory for use with the examples. Once this is done, one is ready to run the example hist2workspace input.xml or edit the XML files for a new project.

**hist2workspace**

```
man hist2workspace
hist2workspace config/example.xml
```

The command line executable `hist2workspace [option] [input xml]` is a utility to create RooFit/RooStats workspace from histograms.
OPTIONS:

- **standard_form** default model (from v5.28.00a and beyond), which creates an extended PDF that interpolates between RooHistFuncs. This is much faster for models with many bins and uses significantly less memory.

- **number_counting_form** this was the original model in 5.28 (without patches). It uses a Poisson for each bin of the histogram. This can become slow and memory intensive when there are many bins.
5 Usage with RooStats tools

Once one runs `hist2workspace` on an XML file there will be output root and eps files in the results directory. The files are named

\[ \text{results/}[\text{Prefix}]_\text{[Channel]}_\text{[Measurement]}_\text{model.root} \]

where Prefix is specified in the `<Combination>` element in the top-level XML file, for example:

```xml
<Combination OutputFilePrefix="./results/example" Mode="comb" />
```

Measurement is specified in each of the `<Measurement>` elements in the top-level XML file

```xml
<Measurement Name="AllSYS" ...>
```

and Channel is ”combined” for the combined model, but a model file is exported for each individual channel as well using the name taken from the `<Channel>` element of the corresponding channel’s XML file

```xml
<Channel Name="channelEle" ...>
```

These root files have inside a RooWorkspace which contains a RooDataSet and a ModelConfig that can be used with standard RooStats tools (see for example

```
$ ROOTSYS/tutorials/RooStats/Standard*Demo.C
```

```bash
$ root.exe results/example_combined_GaussExample_model.root
root (0) Attaching file results/example_combined_GaussExample_model.root as _file0...
root (1) .ls TFile** results/example_combined_GaussExample_model.root
KEY: RooWorkspace combined;1 combined
KEY: TProcessID ProcessID0;1 1222429a-5b98-11e0-9717-0701a8c0beef

root (2) combined->Print()
RooWorkspace(combined) combined contents

variables
--------
...

d.p.f.s
-------
....

functions
---------
....

datasets
--------
RooDataSet::asimovData(channelCat,obs_channel1)
RooDataSet::obsData(channelCat,obs_channel1)

named sets
----------
...

generic objects
---------------
RooStats::ModelConfig::ModelConfig
```

```bash
root (3) using namespace RooStats
root (4) ModelConfig* mc = (ModelConfig*) combined->obj("ModelConfig")
root (5) mc->Print()

*** Using the following for ModelConfig ***

Observables: RooArgSet:: = (obs_channel1,weightVar,channelCat)
Parameters of Interest: RooArgSet:: = (SigXsecOverSM)

Nuisance Parameters: RooArgSet:: = (alpha_syst2,alpha_syst3)
Global Observables: RooArgSet:: = (nominalLumi,nom_alpha_syst1,nom_alpha_syst2,nom_alpha_syst3)

PDF: RooSimultaneous::simPdf[ indexCat=channelCat channel1=model_channel1 ] = 260.156
```
6 Examples

6.1 A Simple Example

Here we consider a single channel example with one signal and two backgrounds. All three samples histograms are based on theoretical predictions (aka. Monte Carlo), thus the luminosity uncertainty should propagate to these channels – this is accomplished by NormalizeByTheory="True". In this example, no shape uncertainties are included and statistical uncertainty on the histograms is not taken into account. The parameter of interest in this example is the signal cross section relative to the predicted value used to create the signal histogram – it is called SigXsecOverSM. Three systematic effects are considered that modify the normalization on the channels – here just named “syst1”, “syst2”, and “syst3”. In this example syst3 affects the normalization of both backgrounds (though in opposite directions).

<table>
<thead>
<tr>
<th>Syst</th>
<th>Signal (s=1)</th>
<th>Background1 (s=2)</th>
<th>Background2 (s=2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>syst1 (p = 1)</td>
<td>(\eta_{11}^+ = 1.05, \eta_{11}^- = 0.95)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>syst2 (p = 2)</td>
<td>–</td>
<td>(\eta_{22}^+ = 1.07, \eta_{22}^- = 0.93)</td>
<td>–</td>
</tr>
<tr>
<td>syst3 (p = 3)</td>
<td>–</td>
<td>(\eta_{32}^+ = 1.03, \eta_{32}^- = 0.95)</td>
<td>(\eta_{33}^+ = 0.97, \eta_{33}^- = 1.02)</td>
</tr>
</tbody>
</table>

Table 5: Tabular representation of sources of uncertainties that produce a correlated effect in the normalization individual samples (eg. OverallSys). The \(\eta_{ps}^+\) represent histogram when \(\alpha_s = 1\) and are inserted into the High attribute of the OverallSys XML element. Similarly, the \(\eta_{ps}^-\) represent histogram when \(\alpha_s = -1\) and are inserted into the Low attribute of the OverallSys XML element. Note, this does not imply that \(\eta^+ > \eta^-\), the ± superscript correspond to the variation in the source of the systematic, not the resulting effect.
6.2 Multiple signal channels

6.3 ABCD

```xml
<!DOCTYPE Combination SYSTEM 'HistFactorySchema.dtd'>
<Combination OutputFilePrefix="./results/ABCD" >
  <Input>./config/A.xml</Input>
  <Input>./config/B.xml</Input>
  <Input>./config/C.xml</Input>
  <Input>./config/D.xml</Input>
  <Measurement Name="ABCD" Lumi="1." LumiRelErr="0.1" ExportOnly="True">
    <POI>mu</POI>
    <ParamSetting Const="True">Lumi b_acceptance c_acceptance d_acceptance mu_K_A mu_K_B mu_K_C mu_K_D</ParamSetting>
  </Measurement>
</Combination>

<!DOCTYPE Channel SYSTEM 'HistFactorySchema.dtd'>
<Channel Name="A" InputFile="./data/ABCD.root">
  <Data HistoName="A_data" HistoPath="" />
  <Sample Name="A_signal" HistoPath="" HistName="unit_histogram">
    <!-- now mu is number of events -->
    <NormFactor Name="mu" Val="1" Low="0" High="200" />
    <OverallSys Name="syst" High="1.01" Low="0.99" />
  </Sample>
</Channel>

<!DOCTYPE Channel SYSTEM 'HistFactorySchema.dtd'>
<Channel Name="B" InputFile="./data/ABCD.root">
  <Data HistoName="B_data" HistoPath="" />
  <Sample Name="B_signal" HistoPath="" HistName="unit_histogram">
    <NormFactor Name="mu" Val="1" Low="0.2" High="1.5" />
    <NormFactor Name="b_acceptance" Val="0.1" Low="0" High="1.5" Const="True" />
  </Sample>
</Channel>

<!DOCTYPE Channel SYSTEM 'HistFactorySchema.dtd'>
<Channel Name="C" InputFile="./data/ABCD.root">
  <Data HistoName="C_data" HistoPath="" />
  <Sample Name="C_signal" HistoPath="" HistName="unit_histogram">
    <NormFactor Name="mu" Val="1" Low="0.2" High="1.5" />
    <NormFactor Name="c_acceptance" Val="0.1" Low="0" High="1.5" Const="True" />
  </Sample>
</Channel>
```

19
<Sample Name="C_backgroundMC" HistoPath="" NormalizeByTheory="True" HistoName="unit_histogram" >
  <NormFactor Name="nu_K_C" Val="100" Low="0" High="200" />
</Sample>

<!-- Background 2 is completely Data-Driven -->
<Sample Name="C_backgroundDD" HistoPath="" NormalizeByTheory="False" HistoName="unit_histogram" >
  <!-- Note, need some reasonable guess for the range of tauC -->
  <NormFactor Name="etaC" Val="100" Low="50" High="150" Const="False" />
  <NormFactor Name="mu_D_U" Val="100" Low="20000" High="30000" />
</Sample>
</Channel>

<!DOCTYPE Channel SYSTEM 'HistFactorySchema.dtd'>

<Channel Name="D" InputFile="/data/ABCD.root" >
  <Data HistoName="D_data" HistoPath="" />
  <!-- This is the signal contamination in D (eg. d*mu) -->
  <Sample Name="D_signal" HistoPath="" HistoName="unit_histogram" >
    <NormFactor Name="mu" Val="1" Low="2" High="1.5" />
    <NormFactor Name="d_acceptance" Val="0.1" Low="0." High="1.5" Const="True" />
  </Sample>
  <!-- This bkg is estimated from MC (eg. mu_D^K) -->
  <Sample Name="D_backgroundMC" HistoPath="" NormalizeByTheory="True" HistoName="unit_histogram" >
    <NormFactor Name="nu_K_D" Val="100" Low="0" High="200" />
  </Sample>
  <!-- Background 2 is completely Data-Driven -->
  <Sample Name="D_backgroundDD" HistoPath="" NormalizeByTheory="False" HistoName="unit_histogram" >
    <NormFactor Name="tauB" Val="1" Low="2" High="1.5" Const="False" />
    <NormFactor Name="tauC" Val="1" Low="2" High="1.5" Const="False" />
    <!--
    <NormFactor Name="nu_B_D" Val="100" Low="0" High="200" />
    </Sample>
  </Sample>
</Channel>
7 The HistFactory XML Schema

Note, when using the HistFactory the production modes \( l \) and backgrounds \( j \) correspond to a single XML Sample element. The HistoName attribute inside each sample element specifies the histogram with the \( \sigma_{ijm}^0 \). The index \( j = J \) is set by the Name attribute of the Sample element (eg. <Sample Name="J">). Between the open <Sample> and close </Sample> one can add

- An OverallSys element where the Name="I" attribute identifies which \( \alpha_I \) is the source of the systematic and implies that the Gaussian constraint \( N(a_i | \alpha_I, 1) \) is present. The High attribute corresponds to \( \eta_{IJ}^+ \), eg when the source of the systematic is at +1\( \sigma \) and \( \alpha_I = 1 \). Similarly, the Low attribute corresponds to \( \eta_{IJ}^- \), eg when the source of the systematic is at -1\( \sigma \) and \( \alpha_I = -1 \). The nominal value is \( \eta_{IJ}^0 = 1 \) for the overall systematics. The distinction between the sign of the source \( \alpha \) and the effect \( \eta \) allows one to have anti-correlated systematics. The HistFactory is able to deal with asymmetric uncertainties as well, by using a piece-wise linear interpolation for the \( \alpha_I > 0 \) and \( \alpha_I < 0 \) regions.

- A NormFactor element is used to introduce an overall constant factor into the expected number of events. In the example below, the term \( \mu = \sigma/\sigma_{SM} \) corresponds to the line <NormFactor Name="SigXsecOverSM">. In this case, the histograms were normalized to unity, so additional NormFactor elements were used to give the overall cross-sections \( \sigma_J \).

- A HistoSys element is used to introduce shape systematics and the HistoNameHigh and HistoNameLow attributes have the variational histograms \( \sigma_{ijm}^+ \) and \( \sigma_{ijm}^- \) corresponding to \( \alpha_i = +1 \) and \( \alpha = -1 \), respectively.

```xml
<!-- The top level combination spec -->
<!-- OutputFilePrefix: Prefix to the output root file to be created (inspection ← histograms) -->
<!-- Mode: Type of the analysis -->
<!ELEMENT Combination (Function*, Input*, Measurement*)>
<!ATTLIST Combination
OutputFilePrefix CDATA # REQUIRED
Mode CDATA # IMPLIED>

<!-- Input files detailing the channels. One channel per file -->
<!ELEMENT Function EMPTY>
<!ATTLIST Function
Name CDATA # REQUIRED
Expression CDATA # REQUIRED
Dependents CDATA # REQUIRED>

<!-- Configuration for each measurement -->
<!-- Name: to be used as the heading in the table -->
<!-- Lumi: the luminosity of the measurement -->
<!-- LumiRelErr: the relative error known for the lumi -->
<!-- BinLow: the lowest bin number used for the measurement (inclusive) -->
<!-- BinHigh: the highest bin number used for the measurement (exclusive) -->
<!-- Mode: type of the measurement (a closed list of ...) -->
<!-- ExportOnly: if "True" skip fit, only export model -->
<!ELEMENT Measurement (POI, ParamSetting*, ConstraintTerm*)>
<!ATTLIST Measurement
Name CDATA # REQUIRED>
```
Lumi CDATA #REQUIRED
LumiRelErr CDATA #REQUIRED
BinLow CDATA #IMPLIED
BinHigh CDATA #IMPLIED
Mode CDATA #IMPLIED
ExportOnly CDATA #IMPLIED

<!-- Specify what you are measuring. Corresponds to the name specified in the ←
construction
of the model in the channel setup. Typically the NormFactor for xsec measurements --→
>
<!ELEMENT POI (#PCDATA)>

<!-- Specify what parameters are fixed, or have particular value --→
<!ELEMENT ParamSetting (#PCDATA)>
<!ELEMENT ConstraintTerm (#PCDATA)>
<!ELEMENT Channel ( Data*, StatErrorConfig*, Sample*)>
<!ELEMENT Data EMPTY>
<!ELEMENT StatErrorConfig EMPTY>
<!ELEMENT Sample (StatError | HistoSys | OverallSys | ShapeSys | NormFactor | ←
ShapeFactor)>
One can convert this Gaussian constraints into a Poisson/Gamma systematic by adding lines like

```xml
<!ELEMENT StatError EMPTY>
<!ATTLIST StatError
  Activate CDATA #REQUIRED
  HistName CDATA #IMPLIED
  InputFile CDATA #IMPLIED
  HistPath CDATA #IMPLIED>
```

```xml
<!ELEMENT HistoSys EMPTY>
<!ATTLIST HistoSys
  Name CDATA #REQUIRED
  InputFile CDATA #IMPLIED
  HistFileHigh CDATA #IMPLIED
  HistPathHigh CDATA #IMPLIED
  HistNameHigh CDATA #IMPLIED
  HistFileLow CDATA #IMPLIED
  HistPathLow CDATA #IMPLIED
  HistNameLow CDATA #IMPLIED
  InputFileLow CDATA #IMPLIED
  InputFileHigh CDATA #IMPLIED>
```

```xml
<!ELEMENT OverallSys EMPTY>
<!ATTLIST OverallSys
  Name CDATA #REQUIRED
  High CDATA #REQUIRED
  Low CDATA #REQUIRED>
```

```xml
<!ELEMENT ShapeSys EMPTY>
<!ATTLIST ShapeSys
  Name CDATA #REQUIRED
  HistName CDATA #REQUIRED
  HistPath CDATA #IMPLIED
  InputFile CDATA #IMPLIED
  ConstraintType CDATA #IMPLIED>
```

```xml
<!ELEMENT NormFactor EMPTY>
<!ATTLIST NormFactor
  Name CDATA #REQUIRED
  Val CDATA #REQUIRED
  High CDATA #REQUIRED
  Low CDATA #REQUIRED
  Const CDATA #IMPLIED>
```

```xml
<!ELEMENT ShapeFactor EMPTY>
<!ATTLIST ShapeFactor
  Name CDATA #REQUIRED>
```
to the Measurement element.
8 Manual entries

NAME
prepareHistFactory - create a working directory for the HistFactory tools

SYNOPSIS
prepareHistFactory [dir_name]

DESCRIPTION
prepareHistFactory is a simple script that prepares a working area (and creates the directory dir_name if specified). Within the directory dir_name, it creates a results/, data/, and config/ directory relative to the given path. It also copies the HistFactorySchema.dtd and example XML files into the config/ directory. Additionally, it copies a root file into the data/ directory for use with the examples. Once this is done, one is ready to run the example hist2workspace input.xml or edit the XML files for a new project.

ORIGINAL AUTHORS
Dominique Tardif
and Kyle Cranmer

COPYRIGHT
This library is free software; you can redistribute it and/or modify it under the terms of the GNU Lesser General Public License as published by the Free Software Foundation; either version 2.1 of the License, or (at your option) any later version.

This library is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; without even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the GNU Lesser General Public License for more details.

You should have received a copy of the GNU Lesser General Public License along with this library; if not, write to the Free Software Foundation, Inc., 51 Franklin St, Fifth Floor, Boston, MA 02110-1301 USA
NAME
hist2workspace - utility to create RooFit/RooStats workspace from histograms

SYNOPSIS
hist2workspace [option] input.xml

DESCRIPTION
hist2workspace is a utility to create RooFit/RooStats workspace from histograms

OPTIONS
- standard_form default model, which creates an extended PDF that interpolates between RooHist-
 Funcs. This is much faster for models with many bins and uses significantly less memory.
- number_counting_form this was the original model in 5.28 (without patches). It uses a Poisson
  for each bin of the histogram. This can become slow and memory intensive when there are many
  bins.

Prepare working area
The ROOT release ships with a script prepareHistFactory in the $ROOTSYS/bin directory that pre-
prepares a working area. It creates a results/, data/, and config/ directory. It also copies the
HistFactorySchema.dtd and example XML files into the config/ directory. Additionally, it
copies a root file into the data/ directory for use with the examples.

HistFactorySchema.dtd
This file is located in $ROOTSYS/etc/ specifies the XML schema. It is typically placed in the
config/ directory of a working area together with the top-level XML file and the individual
channel XML files. The user should not modify this file.

The HistFactorySchema.dtd is commented to specify exactly the meaning of the various options.

Top-Level XML File
(see for example $ROOTSYS/tutorials/histfactory/example.xml) This file is edited by the user.
It specifies
- A top level 'Combination' that is composed of:
  - several 'Channels', which are described in separate XML files.
  - several 'Measurements' (corresponding to a full fit of the model) each of which specifies
    - a name for this measurement to be used in tables and files
    - what is the luminosity associated to the measurement in picobarns
    - which bins of the histogram should be used
    - what is the relative uncertainty on the luminosity
    - which parameters should be fixed/floating (eg nuisance parameters)
    - which type of constraints are desired - Gaussian by default - Gamma, LogNor-
    mal, and Uniform are also supported
    - if the tool should export the model only and skip the default fit

Channel XML Files
(see for example $ROOTSYS/tutorials/histfactory/example_channel.xml) This file is edited by the
user. It specifies for each channel
- observed data
- if absent the tool will use the expectation, which is useful for expected sensitivity
- several 'Samples' (eg. signal, bkg1, bkg2, ...), each of which has:
  - a name
  - if the sample is normalized by theory (eg N = L*sigma) or not (eg. data driven)
  - a nominal expectation histogram
  - a name 'Normalization Factor' (which can be fixed or allowed to float in a fit)
  - several 'Overall Systematics' in normalization with:
    - a name
    - +/- 1 sigma variations (eg. 1.05 and 0.95 for a 5% uncertainty)
  - several 'Histogram Systematics' in shape with:
    - a name (which can be shared with the OverallSyst if correlated)
    - +/- 1 sigma variational histograms

ORIGINAL AUTHORS
Kyle Cranmer , Akira Shibata , and Dominique Tardif

COPYRIGHT
This library is free software; you can redistribute it and/or modify it under the terms of the
GNU Lesser General Public License as published by the Free Software Foundation; either version
2.1 of the License, or (at your option) any later version.

This library is distributed in the hope that it will be useful, but WITHOUT ANY WARRANTY; with-
out even the implied warranty of MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE. See the
GNU Lesser General Public License for more details.

You should have received a copy of the GNU Lesser General Public License along with this
library; if not, write to the Free Software Foundation, Inc., 51 Franklin St, Fifth Floor,
Boston, MA 02110-1301 USA

DEC. 2010