Kyle Cranmer
Draft 4, February 9, 2012

HistFactory User Guide (ROOT 5.32)

Contents

1 Introduction

The HistFactory is a tool to build parametrized probability density functions (pdfs) in the RooFit/RooStats framework based 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 primative 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 $f_{\rm S}(x)$ and $f_{\rm B}(x)$ 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 μ such that $\mu = 0$ corresponds to the background-only hypothesis and $\mu = 1$ corresponds to the nominal signal+background hypothesis. This continuous parameter μ 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 $\{x_1 \dots x_n\}$. 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 $f_{\rm S}(x)$ and $f_{\rm B}(x)$ for a given value of μ. Putting those two ingredients together one obtains what statisticians call a "marked Poisson model":

$$\mathcal{P}(\{x_1 \dots x_n\}|\mu) = {\rm Pois}(n|\mu S + B) \left[\prod_{e=1}^{n} \frac{\mu S f_{\rm S}(x_e) + B f_{\rm B}(x_e)}{\mu S + B} \right] . \tag{1}$$

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

$$\begin{aligned} -\ln L(\mu) &= -n \ln(\mu S + B) + (\mu S + B) + \ln n! - \sum_{e=1}^{n} \ln \left[\frac{\mu S f_{\rm S}(x_e) + B f_{\rm B}(x_e)}{\mu S + B} \right] \\ &= (\mu S + B) + \ln n! - \sum_{e=1}^{n} \ln \left[\mu S f_{\rm S}(x_e) + B f_{\rm B}(x_e) \right] . \end{aligned} \tag{2}$$

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 $\nu_b^{\rm sig}$ and $\nu_b^{\rm bkg}$, 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 ν_b and the shape $f(x)$ via

$$f_S(x_e) = \frac{\nu_{b_e}^{\rm sig}}{S \Delta_{b_e}} \qquad \text{and} \qquad f_B(x_e) = \frac{\nu_{b_e}^{\rm bkg}}{B \Delta_{b_e}} , \tag{3}$$

where b_e is the index of the bin containing x_e and Δ_{b_e} is the width of that same bins. Note, because the $f(x)$ are normalized to unity we have $S = \sum_b \nu_{\rm b}^{\rm sig}$ and $B = \sum_b \nu_{\rm b}^{\rm bkg}$.

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

$$\mathcal{P}(n_b|\mu) = \text{Pois}(n_{\text{tot}}|\mu S + B) \left[\prod_{b \in \text{bins}} \frac{\mu \nu_b^{\text{sig}} + \nu_b^{\text{bkg}}}{\mu S + B} \right] = \mathcal{N}_{\text{comb}} \prod_{b \in \text{bins}} \text{Pois}(n_b|\mu \nu_b^{\text{sig}} + \nu_b^{\text{bkg}}) \ , \quad (4)$$

where n_b is the data histogram and $\mathcal{N}_{\text{comb}}$ 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 μ)
- 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 a 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

	Constrained	Unconstrained
Normalization Variation	OverallSys (η_{csp})	NormFactor (f_p)
Coherent Shape Variation	HistoSys σ_{csbp}	–
Bin-by-bin variation	StatError γ_{cb}	ShapeFactor γ_{sb}

Table 1: Conceptual building blocks for constructing more complicated PDFs: parameters.

2 The Likelihood Template

2.1 Index Convention

I will do my best to stick to the following mnemonic index conventions.

$e \in$ events, eg. x_e

$b \in$ bins when binned, eg n_b

$c \in$ channels, e.g. n_c for number of events in channel c or n_{bc} for number of events in bin b of channel c.

$s \in$ samples

$p \in$ parameters $\mathbb{N}$ = NormFactors $\mathbb{S}$ =Systematics with External Constraints $\mathbf{\Gamma}$ =bin-by-bin

Sometimes a quantity will carry multiple indices, such as n_{bc} for number of events in bin b of channel c; x_{ec} for the value of the observable x for event e in channel c.

2.2 The Template

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

$$\mathcal{P}(n_m, a_p|\mu, \alpha_p) = \prod_{c\in\text{channels}} \text{Pois}(n_c|\nu_c) \left[\prod_{e=1}^{N_c} f_c(x_e|\boldsymbol{\alpha})\right] \cdot G(L_0|L, \Delta_L) \cdot \prod_{p\in\mathbb{S}+\Gamma} P_p(a_p|\alpha_p) \quad (5)$$

where c is an index over distinct subsets of the data 'channels', p is an index over systematic effects, $P_p(a_p|\alpha_p)$ is a constraint term describing an auxiliary measurement a_p that constrains the nuisance parameter α_p.

$$f_c(x_e|\boldsymbol{\alpha}) = \frac{\nu_{b_e}}{\nu_c} \qquad \text{with} \qquad \nu_c = \sum_{b\in\text{bins of channel } c} \nu_b \quad (6)$$

$$f_c(x_e|\boldsymbol{\alpha}) = \frac{1}{\nu_c}\left[\sum_{s\in\text{w/o S.E.}} L_{cs}F_{cs}\eta_{cs}(\boldsymbol{\alpha})\sigma_{csb_e} + \gamma_{cb_e}\left(\sum_{s\in\text{w/S.E.}} L_{cs}F_{cs}\eta_{cs}(\boldsymbol{\alpha})\sigma_{csb_e}\right)\right] \quad (7)$$

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

$$\mathcal{P}(n_m, a_p|\mu, \alpha_p) = \prod_{c\in\text{channels}} \prod_{b\in\text{bins}_c} \text{Pois}(n_{cb}|\nu_{cb}) \cdot G(L_0|L, \Delta_L) \cdot \prod_{p\in\text{Syst}} P_p(a_p|\alpha_p)$$

where b is an index over the bins of the template histograms, p is an index over systematic effects, n_{cb} is the observed number of events in bin b of channel c, $N(a_p|\alpha_p, 1)$ is the normal distribution for the auxiliary measurement a_p that constrains the nuisance parameter α_p and ν_{cb} is the expected number of events in bin b of channel c given by

$$\nu_{cb} = \sum_{s\in\text{w/o S.E.}} L_{cs}F_{cs}\eta_{cs}(\boldsymbol{\alpha})\ \sigma_{csb}(\boldsymbol{\alpha}) + \gamma_{cb}\left(\sum_{s\in\text{w/ S.E.}} L_{cs}F_{cs}\eta_{cs}(\boldsymbol{\alpha})\ \sigma_{csb}(\boldsymbol{\alpha})\right), \quad (8)$$

where F_{cs} is a product of unconstrained normalization factors for sample s for channel c

$$F_{cs} = \prod_{p \in \text{NormFact}_{cs}} f_p \tag{9}$$

that typically include the parameter of interest (eg. $\mu = \sigma/\sigma_{SM}$). The term $\eta_s(\boldsymbol{\alpha})$ parametrizes relative changes in the overall normalization, and $\sigma_{sb}(\boldsymbol{\alpha})$ contains the nominal normalization and parametrizes uncertainties in the shape of the distribution of the discriminating variable. Here s is an index of contributions from different processes with $s = 1$ being the signal process. The nuisance parameters α_p are associated to the source of the systematic effect (e.g. the muon momentum resolution uncertainty), while $\eta_s(\boldsymbol{\alpha})$ and $\sigma_{sb}(\boldsymbol{\alpha})$ represent the effect of that uncertainty. The α_p are scaled so that $\alpha_p = 0$ corresponds to the nominal expectation and $\alpha_p = \pm 1$ correspond to the $\pm 1\sigma$ variations of the source, thus $N(\alpha_p)$ is the standard normal distribution.

3 Interpolation Options

In this section we discuss the impact of systematic uncertainties in our ability to predict the normalization and shape of various 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 (eg. 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 [1]. The effect of these variations about the nominal predictions $\eta_s^0 = 1$ and σ_{sb}^0 is quantified by dedicated studies that provide $\eta_{sp}^{\pm}$ and $\sigma_{spb}^{\pm}$. The result of these studies can be arranged in tables like those below. The main purpose of the `HistFactory` XML schema is to represent these tables.

Syst	Sample 1	...	Sample N
p=OverallSys 1	$\eta^+_{p=1,s=1}$, $\eta^-_{p=1,s=1}$	...	$\eta^+_{p=1,s=N}$, $\eta^-_{p=1,s=N}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
p=OverallSys M	$\eta^+_{p=M,s=1}$, $\eta^-_{p=M,s=1}$	...	$\eta^+_{p=M,s=N}$, $\eta^-_{p=M,s=N}$
Net Effect	$\eta_{s=1}(\boldsymbol{\alpha})$	...	$\eta_{s=N}(\boldsymbol{\alpha})$

Table 2: Tabular representation of sources of uncertainties that produce a correlated effect in the normalization individual samples (eg. OverallSys). The η^+_{pu} represent histogram when $\alpha_s = 1$ and are inserted into the `High` attribute of the `OverallSys` XML element. Similarly, the η^-_{pu} 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 $+/-$ correspond to the variation in the source of the systematic, not the resulting effect.

Syst	Sample 1	...	Sample N
p=HistoSys 1	$\sigma^+_{p=1,s=1,b}$, $\sigma^-_{p=1,s=1,b}$	...	$\sigma^+_{p=1,s=N,b}$, $\sigma^-_{p=1,s=N,b}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
p=HistoSys M	$\sigma^+_{p=M,s=1,b}$, $\sigma^-_{p=M,s=1,b}$	...	$\sigma^+_{p=M,s=N,b}$, $\sigma^-_{p=M,s=N,b}$
Net Effect	$\sigma_{s=1,b}(\boldsymbol{\alpha})$	...	$\sigma_{s=N,b}(\boldsymbol{\alpha})$

Table 3: Tabular representation of sources of uncertainties that produce a correlated effect in the normalization and shape individual samples (eg. HistoSys). The σ^+_{psb} represent histogram when $\alpha_s = 1$ and are inserted into the `HighHist` attribute of the `HistoSys` XML element. Similarly, the σ^-_{psb} represent histogram when $\alpha_s = -1$ and are inserted into the `LowHist` attribute of the `HistoSys` XML element.

For each sample, one can interpolate and extrapolate from the nominal prediction $\eta_s^0 = 1$ and the variations $\eta_{ps}^{\pm}$ to produce a parametrized $\eta_s(\boldsymbol{\alpha})$. Similarly, one can interpolate and extrapolate from the nominal shape σ_{sb} and the variations $\sigma_{psb}^{\pm}$ to produce a parametrized $\sigma_{sb}(\boldsymbol{\alpha})$. 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.

Three interpolation strategies are described below and can be compared in Fig **??**.

[1] Here we suppress the channel index c on η_{cs} and σ_{cab}

Piecewise Linear (InterpCode=0)

The piecewise-linear interpolation strategy is defined as

$$\eta_s(\boldsymbol{\alpha}) = 1 + \sum_{p \in \text{Syst}} I_{\text{lin.}}(\alpha_p; 1; \eta^+_{sp},\ \eta^-_{sp}) \tag{10}$$

and for shape interpolation it is

$$\sigma_{sb}(\boldsymbol{\alpha}) = \sigma^0_{sb} + \sum_{p \in \text{Syst}} I_{\text{lin.}}(\alpha_p; \sigma^0_{sb}, \sigma^+_{psb},\ \sigma^-_{psb}) \tag{11}$$

with

$$I_{\text{lin.}}(\alpha; I^0, I^+, I^-) = \begin{cases} \alpha(I^+ - I^0) & \alpha \geq 0 \\ \alpha(I^0 - I^-) & \alpha < 0 \end{cases} \tag{12}$$

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 middle panel of Fig 1), which can cause some difficulties for numerical minimization packages such as `Minuit`. Second, the interpolation factor can extrapolate to negative values. For instance, if $\eta^- = 0.5$ then we have $\eta(\alpha) < 0$ when $\alpha < -2$ (see right panel of Fig 1).

Note that one could have considered the simultaneous variation of α_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 $\sigma_{sb}(\boldsymbol{\alpha})$ (ie. `HistoSys`).

Piecewise Exponential (InterpCode=1)

The piecewise exponential interpolation strategy is defined as

$$\eta_s(\boldsymbol{\alpha}) = \prod_{p \in \text{Syst}} I_{\text{exp.}}(\alpha_p; 1, \eta^+_{sp},\ \eta^-_{sp}) \tag{13}$$

and for shape interpolation it is

$$\sigma_{sb}(\boldsymbol{\alpha}) = \sigma^0_{sb} \prod_{p \in \text{Syst}} I_{\text{exp.}}(\alpha_p; \sigma^0_{sb}, \sigma^+_{psb},\ \sigma^-_{psb}) \tag{14}$$

with

$$I_{\text{exp.}}(\alpha; I^0, I^+, I^-) = \begin{cases} (I^+)^\alpha & \alpha \geq 0 \\ (I^-)^{-\alpha} & \alpha < 0 \end{cases} \tag{15}$$

Pros: This approach ensures that $\eta(\alpha) \geq 0$ (see left panel of Fig 1). and for small response to the uncertainties it has the same linear behavior near $alpha \sim 0$ as the piecewise linear interpolation (see left panel of Fig 1).

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 right panel of Fig 1).

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

Note, that when paired with a Gaussian constraint on α this is equivalent to linear interpolation and a log-normal constraint in $\ln(\alpha)$. This is the default strategy for normalization uncertainties $\eta_s(\boldsymbol{\alpha})$ (ie. `OverallSys`) and is the standard convention for normalization uncertainties in the LHC Higgs Combination Group..

Quadratic Interpolation and Linear Extrapolation (InterpCode=2)

The quadratic interpolation and linear extrapolation strategy is defined as

$$\eta_s(\boldsymbol{\alpha}) = 1 + \sum_{p \in \text{Syst}} I_{\text{quad.|lin.}}(\alpha_p; \eta^+_{sp},\ \eta^-_{sp}) \tag{16}$$

and for shape interpolation it is

$$\sigma_{sb}(\boldsymbol{\alpha}) = \sigma^0_{sb} + \sum_{p \in \text{Syst}} I_{\text{quad.|lin.}}(\alpha_p; \sigma^0_{sb}, \sigma^+_{psb},\ \sigma^-_{psb}) \tag{17}$$

with

$$I_{\text{quad.|lin.}}(\alpha; I^0, I^+, I^-) = \begin{cases} (b+2a)(\alpha-1) & \alpha > 1 \\ a\alpha^2 + b\alpha & |\alpha| \leq 1 \\ (b-2a)(\alpha+1) & \alpha < -1 \end{cases} \tag{18}$$

and

$$a = \tfrac{1}{2}(I^+ + I^-) - I^0 \qquad \text{and} \qquad b = \tfrac{1}{2}(I^+ - I^-)\ . \tag{19}$$

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 two 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 the middle panel of Fig 1 shows a case where $\eta(\alpha = -0.3) < 1$ while $\eta^\pm > 0$. Second, the interpolation factor can extrapolate to negative values. For instance, if $\eta^- = 0.5$ then we have $\eta(\alpha) < 0$ when $\alpha < -2$ (see right panel of Fig 1).

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

3.1 Defaults in ROOT 5.32

The default strategy for normalization uncertainties $\eta_s(\boldsymbol{\alpha})$ (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 $\sigma_{sb}(\boldsymbol{\alpha})$ (ie. `HistoSys`) is the piecewise linear option.

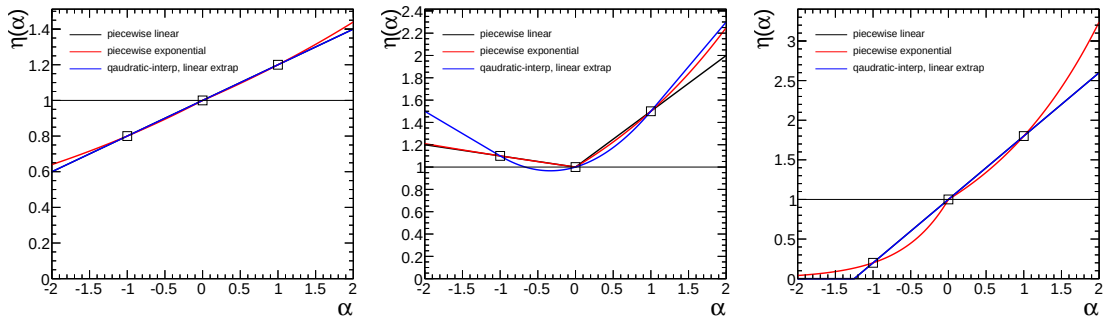

Figure 1: Comparison of the three interpolation options for different $\eta^{\pm}$. Left: $\eta^{-} = 0.8$, $\eta^{+} = 1.2$. Middle: $\eta^{-} = 1.1$, $\eta^{+} = 1.5$. Right: $\eta^{-} = 0.2$, $\eta^{+} = 1.8$.

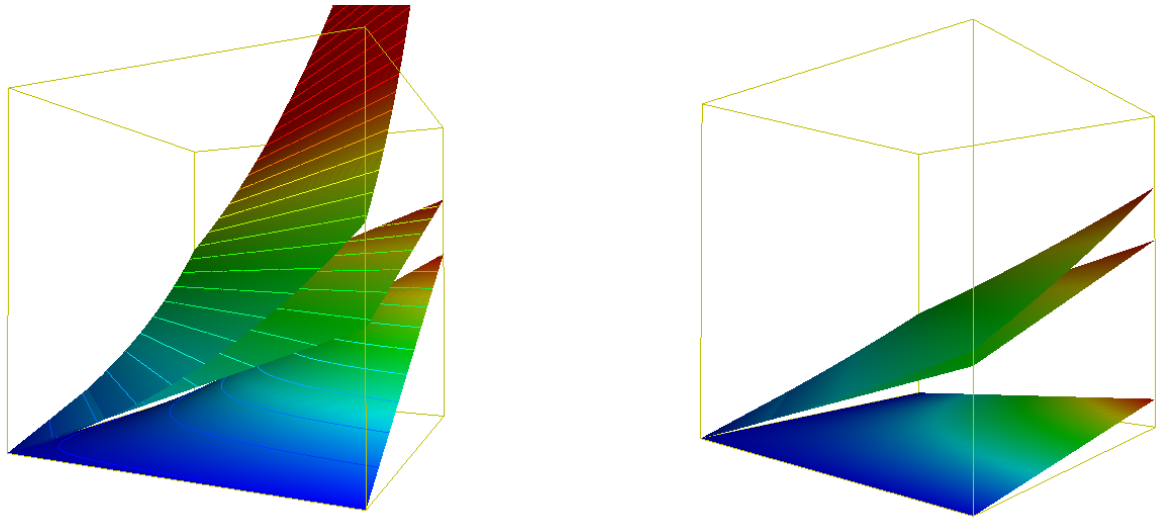

Figure 2: The upper-most curve corresponds to $\eta = (\eta_1^+)^{\alpha_1}(\eta_2^+)^{\alpha_2}$ (as in the exponential interpolation option). The middle surface corresponds to $\eta = 1 + \eta_1^+\alpha_1 + \eta_2^+\alpha_2$ (as in the linear interpolation option). The lowest surface corresponds to $\eta = 1 + \eta_1^+\alpha_1 \cdot \eta_2^+\alpha_2$ (currently not an opiton). 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).

4 Constraint Term Options

5 Examples

5.1 Single Channel

5.2 Multiple signal channels

5.3 ABCD

```
<!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="" />

    <!-- This is the signal (eg. mu)-->
    <Sample Name="A_signal" HistoPath="" HistoName="unit_histogram">
      <!-- now mu is number of events-->
      <NormFactor Name="mu" Val="1" Low="0" High="200" />
      <OverallSys Name="syst1" High="1.01" Low="0.99" />
    </Sample>

    <!-- This bkg is estimated from MC (eg. mu_A^K) -->
    <Sample Name="A_backgroundMC" HistoPath="" NormalizeByTheory="True" HistoName="unit_histogram"  >
      <NormFactor Name="mu_K_A" Val="100" Low="0" High="200" />
    </Sample>

    <!-- Background 2 is completely Data-Driven -->
    <Sample Name="A_backgroundDD" HistoPath="" NormalizeByTheory="False" HistoName="unit_histogram" >
      <NormFactor Name="mu_D_U" Val="100" Low="24500" High="26000" />
      <NormFactor Name="etaB" Val="1" Low="0." High="0.02" Const="False" />
      <NormFactor Name="etaC" Val="1" Low="0." High="0.3" Const="False" />
      <!-- NormFactor and ShapeFactor same for a 1-bin histogram. But we can name NormFactor-->
    </Sample>

  </Channel>
```

```
<!DOCTYPE Channel  SYSTEM 'HistFactorySchema.dtd'>

  <Channel Name="B" InputFile="./data/ABCD.root"  >
    <Data HistoName="B_data" HistoPath="" />

    <!-- This is the signal contamination in B (eg. b*mu)-->
    <Sample Name="B_signal" HistoPath="" HistoName="unit_histogram">
      <NormFactor Name="mu" Val="1" Low=".2" High="1.5" />
      <NormFactor Name="b_acceptance" Val="0.1" Low="0." High="1.5" Const="True"/>
    </Sample>

    <!-- This bkg is estimated from MC (eg. mu_B^K) -->
    <Sample Name="B_backgroundMC" HistoPath="" NormalizeByTheory="True" HistoName="unit_histogram"  >
      <NormFactor Name="mu_K_B" Val="100" Low="0" High="200" />
    </Sample>

    <!-- Background 2 is completely Data-Driven -->
    <Sample Name="B_backgroundDD" HistoPath="" NormalizeByTheory="False" HistoName="unit_histogram" >
      <!-- Note, need some reasonable guess for the range of tauB -->
      <NormFactor Name="etaB" Val="10" Low="5" High="15" Const="False" />
      <NormFactor Name="mu_D_U" Val="100" Low="0" High="200" />
    </Sample>

  </Channel>
```

```
<!DOCTYPE Channel  SYSTEM 'HistFactorySchema.dtd'>

  <Channel Name="C" InputFile="./data/ABCD.root"  >
    <Data HistoName="C_data" HistoPath="" />

    <!-- This is the signal contamination in C (eg. c*mu)-->
    <Sample Name="C_signal" HistoPath="" HistoName="unit_histogram">
      <NormFactor Name="mu" Val="1" Low=".2" High="1.5" />
      <NormFactor Name="c_acceptance" Val="0.1" Low="0." High="1.5" Const="True"/>
    </Sample>

    <!-- This bkg is estimated from MC (eg. mu_C^K) -->
    <Sample Name="C_backgroundMC" HistoPath="" NormalizeByTheory="True" HistoName="unit_histogram"  >
      <NormFactor Name="mu_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="mu_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="mu_D_U" Val="100" Low="0" High="200" />
    </Sample>

  </Channel>
```

6 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 σ^0_{ijm}. 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 α_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 η^+_{IJ}, eg when the source of the systematic is at $+1\sigma$ and $\alpha_I = 1$. Similarly, the `Low` attribute corresponds to η^-_{IJ}, eg when the source of the systematic is at -1σ and $\alpha_I = -1$. The nominal value is $\eta^0_{IJ} = 1$ for the ovreall systematics. The distinction between the sign of the source α and the effect η 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 σ_J.
- A `HistoSys` element is used to introduce shape systematics and the `HistoNameHigh` and `HistoNameLow` attributes have the variational histograms σ^+_{ijm} and σ^-_{ijm} corresponding to $\alpha_i = +1$ and $\alpha = -1$, respectively.

Below is an example XML file for the electron channel.

```
<!DOCTYPE Channel  SYSTEM 'HistFactorySchema.dtd'>
  <Channel Name="channelEle" InputFile="./data/central_Ele_5jet_inc_35invpb.root" HistoName="" >
    <!--<Data HistoName="data" HistoPath="" />-->
    <Sample Name="bbAtautau120" HistoPath="" NormalizeByTheory="True" HistoName="bbAtautau120All">
      <OverallSys Name="JES" High="1.05" Low="0.95"/>
      <OverallSys Name="EVTEFF" High="1.122" Low="0.878"/>
      <OverallSys Name="bbAtautau" High="1.15" Low="0.85"/>
      <NormFactor Name="NEle_bbAtautau120" Val=".83202" Low=".83202" High=".83202" Const="True" />
      <NormFactor Name="SigXsecOverSM" Val="0" Low="-10." High="30." Const="True" />
    </Sample>
    <Sample Name="Atautau120" HistoPath="" NormalizeByTheory="True" HistoName="Atautau120All">
      <OverallSys Name="JES" High="1.05" Low="0.95"/>
      <OverallSys Name="EVTEFF" High="1.122" Low="0.878"/>
      <OverallSys Name="Atautau" High="1.15" Low="0.85"/>
      <NormFactor Name="NEle_Atautau120" Val=".24224" Low=".24224" High=".24224" Const="True" />
      <NormFactor Name="SigXsecOverSM" Val="0" Low="-10." High="30." Const="True" />
    </Sample>
    <Sample Name="bbAtautau130" HistoPath="" NormalizeByTheory="True" HistoName="bbAtautau130All">
      <OverallSys Name="JES" High="1.05" Low="0.95"/>
      <OverallSys Name="EVTEFF" High="1.122" Low="0.878"/>
      <OverallSys Name="bbAtautau" High="1.15" Low="0.85"/>
      <NormFactor Name="NEle_bbAtautau130" Val=".01767" Low=".01767" High=".01767" Const="True" />
      <NormFactor Name="SigXsecOverSM" Val="0" Low="-10." High="30." Const="True" />
    </Sample>
    <Sample Name="Atautau130" HistoPath="" NormalizeByTheory="True" HistoName="Atautau130All">
      <OverallSys Name="JES" High="1.05" Low="0.95"/>
      <OverallSys Name="EVTEFF" High="1.122" Low="0.878"/>
      <OverallSys Name="Atautau" High="1.15" Low="0.85"/>
      <NormFactor Name="NEle_Atautau130" Val=".02441" Low=".02441" High=".02441" Const="True" />
      <NormFactor Name="SigXsecOverSM" Val="0" Low="-10." High="30." Const="True" />
    </Sample>
    <Sample Name="Ztautau" HistoPath="" NormalizeByTheory="True" HistoName="ZtautauAll">
      <OverallSys Name="JES" High="1.05" Low="0.95"/>
      <OverallSys Name="EVTEFF" High="1.122" Low="0.878"/>
      <OverallSys Name="Alpgen" High="1.131" Low="0.869"/>
      <OverallSys Name="Ztautau" High="1.15" Low="0.85"/>
      <NormFactor Name="NEle_Ztautau" Val="1.26818" Low="1.26818" High="1.26818" Const="True" />
    </Sample>
   <Sample Name="AddOn" HistoPath="" NormalizeByTheory="False" HistoName="AddOnAll">
      <OverallSys Name="AddOn" High="1.173" Low=".827"/>
      <NormFactor Name="NEle_AddOn" Val=".88267" Low=".88267" High=".88267" Const="True" />
```

```
    </Sample>
    <Sample Name="SameSign" HistoPath="" NormalizeByTheory="False" HistoName="SameSignAll">
      <OverallSys Name="SameSign" High="1.06828" Low=".93172"/>
      <NormFactor Name="NEle_SameSign" Val="4.00568" Low="4.00568" High="4.00568" Const="True" />
    </Sample>
    <Sample Name="Others" HistoPath="" NormalizeByTheory="True" HistoName="OthersAll">
      <OverallSys Name="JES" High="1.05" Low="0.95"/>
      <OverallSys Name="EVTEFF" High="1.122" Low="0.878"/>
      <OverallSys Name="QFAC" High="1.03" Low="0.97"/>
      <OverallSys Name="Alpgen" High="1.131" Low="0.869"/>
      <OverallSys Name="Others" High="1.15" Low="0.85"/>
      <NormFactor Name="NEle_Others" Val=".17949" Low=".17949" High=".17949" Const="True" />
    </Sample>
  </Channel>
```

One can convert this Gaussian constraints into a Poisson/Gamma systematic by adding lines like

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

to the `Measurement` element. For example:

```
  <Measurement Name="AllSYS" Lumi="35.2" LumiRelErr="0.11" BinLow="0" BinHigh="20" Mode="comb"  ExportOnly="True"↩
      >
    <POI>SigXsecOverSM</POI>
    <ParamSetting Const="True">NEle_AddOn,NEle_Atautau120,NEle_Atautau130,NEle_Others,
      NEle_SameSign,NEle_Ztautau,NEle_bbAtautau120,NEle_bbAtautau130,NMuo_AddOn,
      NMuo_Atautau120,NMuo_Atautau130,NMuo_Others,NMuo_SameSign,NMuo_Ztautau,NMuo_bbAtautau120,
      NMuo_bbAtautau130
    </ParamSetting>
    <ConstraintTerm Type="Gamma" RelativeUncertainty="0.1">JES</ConstraintTerm>
    <!--<ConstraintTerm Type="LogNormal" RelativeUncertainty="0.1">JES</ConstraintTerm>-->
  </Measurement>
```

7 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.

8 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

```
results/[Prefix]_[Channel]_[Measurement]_model.root
```

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

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

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

```
<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

```
<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`

```
$ hist2workspace config/example.xml
$ 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
 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
---------
...

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

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

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

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

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



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
```

9 Manual entries

```
man prepareHistFactory
PREPAREHISTFACTORY(1)                                                PREPAREHISTFACTORY(1)

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  con-
       fig/  directory relative to the given path.  It also copies the HistFactorySchema.dtd and exam-
       ple 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; 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                PREPAREHISTFACTORY(1)
```

```
man hist2workspace
HISTTOWORKSPACE(1)                                                HISTTOWORKSPACE(1)

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-
       pares 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
             - what is (are) the parameter(s) of interest that will be measured
             - which parameters should be fixed/floating (eg. nuisance parameters)
             -  which type of constriants 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 named '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                HISTTOWORKSPACE(1)
```