Dissertation

Automated optimization of sensitivity in a search for boosted VBF Higgs pair production in the $b\bar{b}b\bar{b}$ quark final state with the ATLAS detector

For the attainment of the academic degree doctor rerum naturalium

(Dr. rer. nat.) in the subject: Physics

Frederic Renner

Berlin, 06.12.2023

Faculty of Mathematics and Natural Sciences of the Humboldt University of Berlin

1st Supervisor: Dr. Clara Elisabeth Leitgeb

2nd Supervisor: Prof. Dr. Cigdem Issever

(Only after the disputation for publication in the university library according to § 15 of the doctoral regulations enter the names and the date):

Reviewers:

1st:

2nd:

3rd:

Date of the oral examination:

Abstract

I am an abstract.

Contents

Chapter 1

Systematic Uncertainties

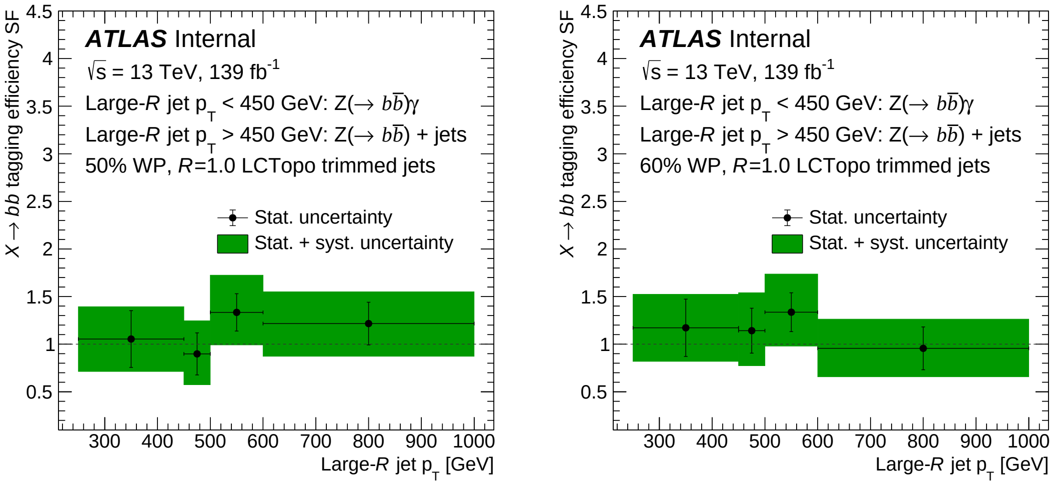
Any measurement needs to consider uncertainties in order to determine its validity. In this analysis they can be divided into systematic errors for the reconstructed objects, uncertainties from theoretical calculations, methodological errors and statistical uncertainties and are described in the following.

1.1 Luminosity

The combined integrated luminosity for the years 2015-2018 has an uncertainty of 0.83 % determined with the LUCID-2 detector and is applied. It is applied to the Higgs Pair Signal process and has minimal impact on the analysis.

1.2 Jet Uncertainties

Jets are calibrated using well known reference objects as described in section **??**. These corrections are themselves subject to uncertainties related to detector effects, modeling and statistics leading to corrections of the jet energy and are collectively referred to as Jet Energy Scale (JES) [1, 2]. Since simulations of jets have a higher accuracy than observed jets the uncertainties of the simulated jets are broadened to be consistent with the jets observed in the data. These uncertainties are known as Jet Energy Resolution (JER). Furthermore large-R jets are additionally corrected



(a) (b)

Figure 1.1: Derived scale factors in large-R jet p_{T} for the **(a)** 50 % and **(b)** 60 % working point (WP) from the calibration of the $X \to bb$ tagger.

for their mass. The uncer tainties related to this procedure are called Jet Mass Resolution (JMR) [3].

1.3 $X \to bb$ Tagger Uncertainties

The Neural Network (NN) of the $X \to bb$ tagger was trained using simulations leading to potential discrepancies in selection efficiencies between observed data and simulation. Calibration is conducted with $Z(\to b\bar{b})+$ jets and $Z(\to b\bar{b}) + \gamma$ applying the same methodology as in [4]. However as of this analysis the b-tagging algorithm for the variable radius (VR) track jets has been updated to the DL1d algorithm described in section **??**. The differences between Monte Carlo (MC) and data are measured in large-R jet p_{T} and the extracted scale factors and their corresponding combined systematic and statistical uncertainties are shown in figure 1.1.

1.4 Theory Uncertainties

The cross-section calculation for some process initiated by a proton proton collision calculated at n-th order has a functional form [5]

$$\sigma^{(n)} = PDF(x_1, \mu_F) PDF(x_2, \mu_F) \hat{\sigma}^{(n)}(x_1, x_2, \mu_R), \tag{1.4.1}$$

with the Parton Density Functions (PDFs) carrying momentum fraction $x_{1,2}$ of the partons and the factorization scale μ_F. This scale is named after the assumption that the cross-sections of the initial particle can be calculated by factorizing it in its parton contributions [6]. The term $\hat{\sigma}^{(n)}$ in equation 1.4.1 is the calculable part of the cross-section at renormalization scale μ_R as described in section **??** and is expanded to a desired order n in the strong coupling constant α_s with the usual Quantum Field Theory (QFT) ansatz outlined in section **??**

$$\hat{\sigma}^{(n)} = \alpha_s \hat{\sigma}^{(0)} + \alpha_s^2 \hat{\sigma}^{(1)} + \ldots + \alpha_s^n \hat{\sigma}^{(n)} + \mathcal{O}(\alpha_s^{n+1}). \tag{1.4.2}$$

Modelling cross-sections via PDFs is necessary since the approximation of the perturbation ansatz of section **??** breaks down for low energy scales Q^2 as described in section **??** which is the energy scale for which the approximation would need to hold to describe the partons inside a proton. However similar to renormalization a scaling behavior can be derived which allows to deduce an estimate of the PDFs by measuring it at a some energy scale Q^2 to extrapolate it to another. The equations enabling this are also expanded in α_s to a desired order and are knwon as DGLAP equations [6]. Three main sources of uncertainty arise in this calculation described in the following.

Scale Variations

α_s is expanded to some order n in the cross-section calculation and as well in estimating the PDFs. To account for missing higher orders corrections of theses expansions scale variations of the renormalization and factorization scales are performed pairwise $\{\mu_\mathrm{r}, \mu_\mathrm{f}\} \times \{0.5, 0.5\}, \{1, 0.5\}, \{0.5, 1\}, \{1, 1\}, \{2, 1\}, \{1, 2\}, \{2, 2\}$. For the cross-section calculation this accounts essentially for the term $\mathcal{O}(\alpha_s^{n+1})$ in equation 1.4.2. The envelope that gives the largest variation is taken as the scale uncertainty.

PDF Uncertainties

PDFs need to be deduced from experiment and thus come by themselves with experimental uncertainties. Further uncertainties arise from the functional forms assumed for the PDFs.

α_s Uncertainties

α_s is also experimentally deduced at the scale of the Z mass which is subject to uncertainties. In all perturbative calculations it is truncated at some order that needs to be accounted for.

The uncertainties on α_s and the PDFs are both estimated by varying α_s. Even though there correlation is not strong they are usually applied combined [5].

1.4.1 Uncertainty on HH cross section

The cross-section calculation for the vector-boson fusion (VBF) Higgs pair production process has associated uncertainties for the scale variations $^{+0.03\%}_{-0.04\%}$ and the combined PDF+α_s uncertainty is $\pm 2.1\%$ [7].

1.4.2 Uncertainty on Acceptance

Theoretical uncertainties on the final acceptance are evaluated on MC simulations for scale variations and PDFs + α_s TODO, although shouldnt matter...

1.4.3 Parton Shower

Uncertainties related to the parton showering are estimated using different modelings from Pythia 8 and Herwig 7. The largest deviations from the nominal are used as uncertainties on the Higgs pair process. TODO

1.4.4 Branching Ratio Uncertainty

The error estimate for the branching ratio takes into account theoretical uncertainties (THU) and parametric uncertainties (PU) that are included in the Standard Model (SM) calculations. The theoretical uncertainties mainly considers missing higher orders while for the parameters p the four leading non-negligible contributions of $p = \alpha_s, m_c, m_b, m_t$ are considered.

Parametric uncertainties are Gaussian errors and are added in quadrature which ensures unity in the Branching Ratio calculation. Theoretical uncertainties in turn are not Gaussian and would lead to underestimated errors and are therefore added linearly [7]. By assuming a Higgs mass of 125 GeV and considering that there are two Higgs decaying to two b-quarks the error on the branching is

$$\Delta\mathrm{BR} = 2 \times \left(\Delta\mathrm{BR}(\mathrm{THU}) + \sqrt{\textstyle\sum_p \Delta\mathrm{BR}(\mathrm{PU}_p)^2}\right) =^{+3.4\%}_{-3.5\%} . \tag{1.4.3}$$

1.5 Statistical Uncertainties

As discussed in the chapter on statistics 2 the bin content for histograms in this work follows a Poisson distribution. Therefore the standard error for N events is the square root of the variance $\sigma = \sqrt{\mathrm{Var}} = \sqrt{N}$. Since histograms are filled weighted $\sum_i w_i N_i$ this needs to be taken into account. By making use of the additive property and invariance with respect to constants of the variance a bin filled with weights w_i can be written as

$$\begin{aligned}\sigma^2_{\mathrm{stat}} &= \mathrm{Var}_{\mathrm{bin}}\left(\sum_i w_i\right) = \underbrace{\sum_i \mathrm{Var}(w_i \times 1 \text{ event})}_{\mathrm{Var}(i+j)=\mathrm{Var}(i)+\mathrm{Var}(j)} = \underbrace{\sum_i w_i^2 \mathrm{Var}(1 \text{ event})}_{\mathrm{Var}(aX)=a^2\mathrm{Var}(X)} \\ &= \sum_i w_i^2 \sqrt{(1 \text{ event})},\end{aligned} \tag{1.5.1}$$

so that the statistical error reads

$$\sigma^{\mathrm{bin}}_{\mathrm{stat}} = \sqrt{\sum_i w_i^2}. \tag{1.5.2}$$

1.6 Background Derivation Uncertainties

The Quantum Chromodynamics (QCD) background is estimated with the ABCD method from the control region as detailed in section **??**. The uncertainties are assessed through error propagation of the statistical uncertainty meaning the statistical errors of the event yields used to retrieve the weight factor result in an uncertainty for the weight factor $\Delta w_{\mathrm{CR}} = 0.0\,\%$. To estimate a bin-wise statistical uncertainty of the NN score histogram the ABCD procedure is applied in the VR by also propagating the error Δw_{CR} of the applied weight factor.

Chapter 2

Statistics

Every scientific investigation starts with a hypothesis that is to be tested empirically. The main objective is to evaluate if the proposed hypothesis agrees or disagrees with observed data, to either accept or reject it against the null-hypothesis which represents a baseline scenario where only known phenomena are presumed to occur.

A key metric that quantifies this is the p-value that arises within hypothesis testing. Test results of an experiment follow some probability density function. Assuming some hypothesis the p-value is the integrated probability for test results compatible with this hypothesis and ergo measuring the compatibility of the observation to the assumption. In other words if the experiment where to be repeated it gives the probability that the result favors the proposed hypothesis.

In the field of high energy physics a framework has been developed specifically for this task. This section begins to lay out the mathematical fundamentals of the approach and explains its implementation in PYHF [8, 9]. The following is based on [8, 10, 11].

2.1 Building the likelihood

The statistical model needs to reflect the compatibility of predictions with the observed collision events. This can be quantified by a likelihood $L(\boldsymbol{x}|\boldsymbol{\phi})$ which is a probability for an observation $\boldsymbol{x}$ under a given set of parameters $\boldsymbol{\phi}$ that govern the predictions. Given that this is a counting experiment bins of a histogram $\boldsymbol{h} = (h_1, ..., h_N)$ are the main tool of analysis.

The observation can be subdivided $\boldsymbol{x} = (\boldsymbol{n}, \boldsymbol{a})$ into observable histograms $\boldsymbol{n}$ and auxiliary measurements. Observable histograms could be the invariant mass of a particle on the other hand auxiliary measurement histograms $\boldsymbol{a}$ can be any additional observable that assist in constraining the model. For instance they can be a measurement of a kinematic variable in a phase space region where only background is expected. It should be noted that these auxiliary measurements are not equivalent to the inclusion of uncertainties into the model. The treatment of uncertainties is addressed separately in section 2.4.

Another useful splitting for the set of parameters $\boldsymbol{\phi} = (\boldsymbol{\psi}, \boldsymbol{\Theta})$ into so-called parameters of interest $\boldsymbol{\psi}$ and nuisance parameters $\boldsymbol{\Theta}$. For this section only one parameter of interest is considered, the signal strength μ.

The bin contents can then be expressed in terms of the amount of signal $s_i(\boldsymbol{\Theta})$ and background $b_i(\boldsymbol{\Theta})$ in bin i that depend on the nuisance parameters. The prediction (expectation value) of the histogram bins of the observable n_i can then be expressed as

$$\langle n_i(\mu, \boldsymbol{\Theta}) \rangle = \mu s_i(\boldsymbol{\Theta}) + b_i(\boldsymbol{\Theta}). \tag{2.1.1}$$

Similarly for auxiliary measurement bins a_i their expectation value are calculable from some function $u_i(\boldsymbol{\Theta})$ modeling some observable and is also dependent on the nuisance parameters

$$\langle a_i(\boldsymbol{\Theta}) \rangle = u_i(\boldsymbol{\Theta}). \tag{2.1.2}$$

Since this is a counting experiment in which events occur at a constant mean rate and independently of time each bin follows a Poisson distribution

$$P(r, k) = \frac{r^k e^{-r}}{k!}. \tag{2.1.3}$$

r is the expected rate of occurrences which translates as the prediction whereas k are the actual measured occurrences. A likelihood can then be constructed from a product of Poisson probabilities

$$L(\mu, \mathbf{\Theta}) = \prod_{j=1}^{N} \frac{(\mu s_j(\mathbf{\Theta}) + b_j(\mathbf{\Theta}))^{n_j}}{n_j!} e^{-(\mu s_j(\mathbf{\Theta}) + b_j(\mathbf{\Theta}))} \prod_{k=1}^{M} \frac{u_k(\mathbf{\Theta})^{a_k}}{a_k!} e^{-u_k(\mathbf{\Theta})}. \tag{2.1.4}$$

The last product can also be thought of penalizing the likelihood if e.g. an auxiliary measurement displays a very improbable value for some quantity. To test for a hypothesized value of μ, the best choice according to the Neyman-Pearson lemma [11], is the profile likelihood ratio that reduces the dependence to the parameter(s) of interest

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\mathbf{\Theta}}})}{L(\hat{\mu}, \hat{\mathbf{\Theta}})}. \tag{2.1.5}$$

The denominator is the unconditional maximum likelihood estimate so that $\hat{\mu}$ and $\hat{\mathbf{\Theta}}$ both are free to vary to maximize L, whereas the numerator is the found maximum likelihood conditioned on some chosen μ and the set of nuisance parameters $\hat{\hat{\mathbf{\Theta}}}$ that maximize the likelihood for that particular μ. This definition gives $0 \leq \lambda \leq 1$ where $\lambda = 1$ corresponds to perfect agreement of the hypothesized value of μ to the model.

2.2 From test statistic to p-value

To test for alternative hypotheses it is useful to transform the profile likelihood into a test statistic

$$t(\mu) = -2 \ln \lambda(\mu). \tag{2.2.1}$$

This translates to $t \to 0$ as increasing agreement and $t \to \infty$ as decreasing agreement to the model. A right-tail p-value can then be calculated from the probability density function of the test statistic: $\mathrm{PDF}(t) = f(t|\mu)$

$$p = \int_{t_{\mathrm{obs}}}^{\infty} f(t|\mu)\mathrm{d}t \tag{2.2.2}$$

t_{obs} is the test statistic t evaluated at the observed data which means replacing the predictions in the likelihood of the numerator of equation 2.1.5 with the values observed in data. Similar to the PDF of a standard normal distribution the PDF in this context quantifies how probable a particular value of the test statistic t is under a fixed value of the signal strength. This essentially measures how frequently a particular value of t occurs in comparison to all other possible values that t can have.

This form is particularly useful due to existing approximations for $f(t|\mu)$ [10]. Wald [12] proved that for the null hypothesis in the large sample limit, the test statistic follows a normalized sum of squared distances between the tested parameters of interest μ and its maximum likelihood estimate $\hat{\mu}$.

$$t(\mu) = -2\ln\lambda(\mu) = \left(\frac{\mu - \hat{\mu}}{\sigma_{\hat{\mu}}}\right)^2 + \mathcal{O}(\frac{1}{\sqrt{N}}). \tag{2.2.3}$$

The maximum likelihood estimate $\hat{\mu}$ is in the large sample limit normally distributed around their true values μ' with standard deviation $\sigma_{\hat{\mu}}$. This is the definition of a non-central χ-squared distribution with one degree of freedom. It can be shown that the PDF of t can be described by cite

$$f(t|\mu) = \frac{1}{2\sqrt{t}}\frac{1}{\sqrt{2\pi}}\left[\exp\left(-\frac{1}{2}\left(\sqrt{t}+\sqrt{\Lambda_\mu}\right)\right) + \exp\left(-\frac{1}{2}\left(\sqrt{t}-\sqrt{\Lambda_\mu}\right)\right)\right], \tag{2.2.4}$$

with non-centrality parameter as the normalized distance between tested and true parameter of interest value

$$\Lambda_\mu = \frac{(\mu - \mu')^2}{\sigma^2}. \tag{2.2.5}$$

Figure 2.1 illustrates the different steps. Being able to calculate p-values allows now to state how likely it is that the proposed hypothesis is reflected by the observed data. In other words, the p-value represents the probability, how incompatible the proposed hypothesis (prediction) is with the observation.

In the scientific community a widely accepted threshold for this is a p-value of 0.05. Though particle physicists only claim discovery of a new phenomenon for $p < 2.87 \times 10^{-7}$ corresponding to 5 standard deviations of the standard normal distribution and exclude hypotheses if the p-value is not below 2 standard deviations of the standard normal distribution $p \lesssim 0.05$. One caveat here is that this particular form of t assumes μ can also be negative, which can be non-physical depending

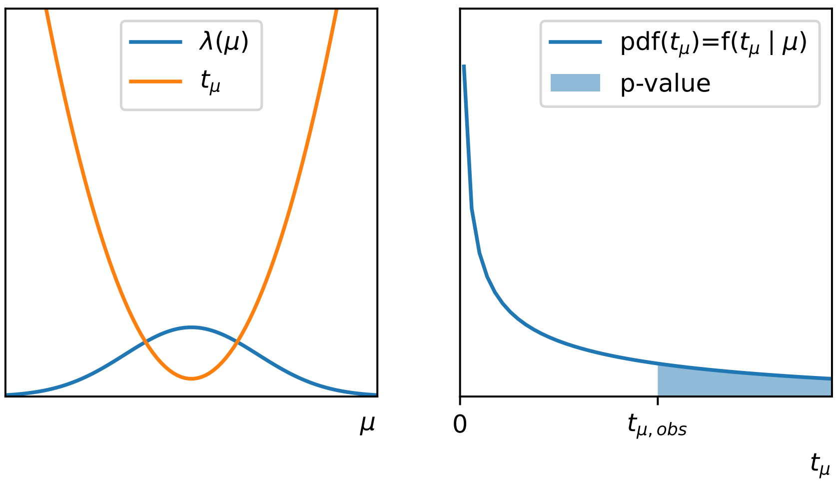


Figure 2.1: A sketch to follow the steps to calculate p-values. (**left**) The profile likelihood (■) has essentially some hill-like form with a maximum at $\lambda(\hat{\mu}, \hat{\mathbf{\Theta}})$, t (■) is $-2\ln(\lambda)$. (**right**) For one parameter of interest in the large sample limit $f(t|\mu)$ follows a non-central chi-squared distribution with one degree of freedom, equation 2.2.4. The blue shaded area under the PDFs is a right hand sided p-value.

on the impact of a new process. Test statistics and their PDFs approximations considering the different cases are covered in [10].

2.3 The CL$_s$ value

Particle physicists are usually interested in two things when making statistical tests for the discovery of new phenomena: how well is the modeling of backgrounds (things we know) and whether there is evidence in the observations for a new phenomenon. This means one needs to test two hypotheses: a background only (b) and a signal plus background ($s+b$) hypothesis. Each will result in a p-value of their own. For example, $p_b = 0$ would mean that the backgrounds are perfectly reflected by the observations and a $p_{s+b} < 0.05$ could be a sign of e.g. new physics. To combine these two metrics into a single score, particle physicists came up with the pseudo Confidence Level/p-value called CL$_s$ incorporating also the goodness of

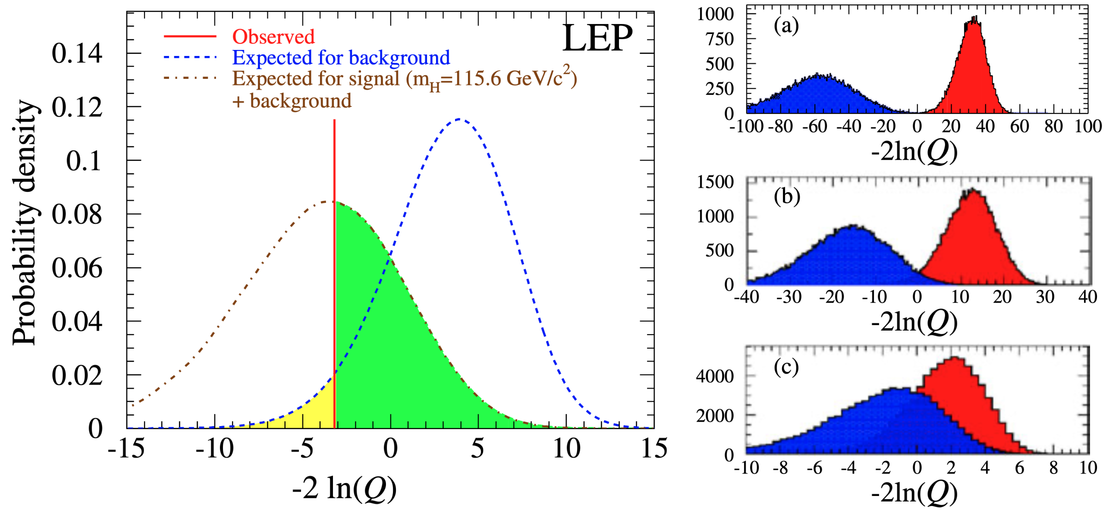


Figure 2.2: Probability density functions of test statistics from a Higgs search at LEP illustrating the calculation of p-values (λ becomes Q). (**left**) The PDFs's of the test statistic $f(t|\mu)$ of the signal + background (/) and background (/) only hypotheses. The p-value is calculated by integration from t_{obs} (the red observed line (/)) to infinity (see eq. 2.2.2). The green shaded area (■) corresponds to p_{s+b} whereas the yellow area (■) corresponds to $1 - p_b$ since the integral over one whole PDFs is 1. (**right**) Degradation of search sensitivity from (a) to (c). Note that the colors of the PDFs's change here to signal + background (■) and background only (■). For example putting the observation (t_{obs}) on the x-axis at 0 in these plots, one would get for plot (a) $p_b \approx 1$ and $p_{s+b} \approx 0$ resulting in a $\mathrm{CL}_s \approx 0$, whereas with increasing overlap the CL_s value increases and the sensitivity decreases. Taken from [13].

the modeling of the backgrounds

$$\mathrm{CL}_s = \frac{p_{s+b}}{1 - p_b} = \frac{\int_{t_{\text{obs}}}^{\infty} f(t_{s+b}|\mu)\mathrm{d}t}{1 - \int_{t_{\text{obs}}}^{\infty} f(t_b|\mu)\mathrm{d}t}. \tag{2.3.1}$$

Intuitively the numerator is again just the value for the alternative hypothesis whereas the denominator penalizes CL_s if the modeling of the backgrounds is not reflected in the observations. This can also be understood visually from the first figure of the paper that introduced the CL_s quantity [13] (see description of fig. 2.2).

2.4 The HistFactory model

A model used widely to build a likelihood as described in section 2.1 is called HistFactory [14] and is implemented within PYHF [8]. This section is based on the introduction to the model within the documentation of PYHF. HistFactory reduces the building of a likelihood into a small number of basic components. For this purpose, it is again useful to think of another splitting of the model parameters $\boldsymbol{\phi}$ into

$$L(\boldsymbol{x}|\boldsymbol{\phi}) = L(\boldsymbol{x}|\overbrace{\boldsymbol{\psi}}^{\text{parameters of interest}}, \underbrace{\boldsymbol{\theta}}_{\text{nuisance parameters}}) = L(\boldsymbol{x}|\overbrace{\boldsymbol{\eta}}^{\text{free}}, \underbrace{\boldsymbol{\chi}}_{\text{constrained}}) \tag{2.4.1}$$

free parameters $\boldsymbol{\eta}$ and constrained parameters $\boldsymbol{\chi}$. Free parameters are free to choose in the model and can be for example a cross-section of a process. Constrained parameters are used to incorporate uncertainties into the likelihood to constrain it. Further there might be several histograms of an observable, for example measured in orthogonal kinematic regions, that are called channels c. Bins have the index b here and constraint terms are denoted c_χ. With that the likelihood can be described by

$$L(\boldsymbol{n}, \boldsymbol{a}\,|\,\boldsymbol{\eta}, \boldsymbol{\chi}) = \underbrace{\prod_{c \in \text{channels}} \prod_{b \in \text{bins}_c} \text{Pois}\left(n_{cb}\,|\,\nu_{cb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right)\right)}_{\substack{\text{Simultaneous measurement}\\ \text{of multiple channels}}} \underbrace{\prod_{\chi \in \boldsymbol{\chi}} c_\chi(a_\chi|\,\chi)}_{\substack{\text{constraint terms}\\ \text{for auxiliary measurements}}} \;. \tag{2.4.2}$$

The n_{cb} is the observation and $\nu_{cb}(\boldsymbol{\eta}, \boldsymbol{\chi})$ the prediction. The $c_\chi(a_\chi|\,\chi)$ are calculated from auxiliary measurements a_χ (the uncertainties) to constrain the parameter χ and can be any function (e.g. Gaussian, Poissonian,...) the parameter/uncertainty is believed to be distributed.

The prediction is a sum of nominal bin counts[1] ν^0_{scb} over all samples s (e.g. $t\bar{t}$, multijet-background, etc.). These nominal bin counts are subject to uncertainties. Therefore the bin counts can be varied within the bounds of these uncertainties. However the effect of this modification to the likelihood must be taken into account which is through the constraint terms. These penalize the likelihood the larger the

[1] also called rates, like in the definition of a Poisson distribution

modification to a nominal value becomes. The ν^0_{scb} are varied with multiplicative κ_{scb} and additive modifiers Δ_{scb}

$$\nu_{cb}\left(\boldsymbol{\phi}\right) = \sum_{s \in \text{samples}} \nu_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right) \tag{2.4.3}$$

$$= \sum_{s \in \text{samples}} \underbrace{\left(\prod_{\kappa \in \boldsymbol{\kappa}} \kappa_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right)\right)}_{\text{multiplicative modifiers}} \left(\nu^0_{scb} + \underbrace{\sum_{\Delta \in \boldsymbol{\Delta}} \Delta_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right)}_{\text{additive modifiers}}\right). \tag{2.4.4}$$

The different types of modifiers are explained in section 2.5 and the constraint terms c_χ in section 2.6.

Why this is useful can be seen by considering one uncertainty to a nominal bin count estimate ν^0_{scb}. By modifying ν^0_{scb} with a factor κ in a way that increases the Poisson probability while the corresponding constraint term $c_\kappa(\kappa)$ stays around 1, it can be beneficial for the goal of maximizing the likelihood. This means the most likely/compatible value to the observed data within the modeling of the uncertainties can be found.

2.5 The Modifiers

In HistFactory there are by convention four types $\{\lambda, \mu, \gamma, \alpha\}$ of such multiplicative rate modifiers that are explained in this section. There are **free rate modifiers** λ **and** μ that affect all bins equally, like the cross-section of a process or the luminosity

$$\nu_{scb}(\mu) = \mu \nu^0_{scb}. \tag{2.5.1}$$

These are bin-independent normalization factors and preserve the shape of the histogram. Further there are **bin-wise modifiers** γ_b (uncorrelated shape)

$$\nu_{scb}(\gamma_b) = \gamma_b \nu^0_{scb}. \tag{2.5.2}$$

These are useful for example to include uncertainties of a per bin data-driven background estimate. This type without a constraint term is not of much use as if there is only one sample or channel, the fit would always match the data perfectly.

In addition there exist **interpolation parameters** α (shape factors) that enter the modeling through an interpolation function η instead of being the factor itself. They exist in multiplicative versions

$$\nu_{scb}(\alpha) = \eta(\alpha)\nu^0_{scb}, \tag{2.5.3}$$

and additive versions

$$\nu_{scb}(\alpha) = \nu^0_{scb} + \eta(\alpha). \tag{2.5.4}$$

This is useful to include systematic uncertainties. In a typical ATLAS analysis usually one knows the one standard deviation of a bin count $\eta_{-1} = \nu^{\text{1down}}_{scb}$ and $\eta_1 = \nu^{\text{1up}}_{scb}$ to the nominal value ν^0_{scb} of an uncertainty. These are used to construct interpolation functions that modify the nominal value with a nuisance parameter that is also used to apply a penalization c_α according to the modeling of the uncertainty.

In HistFactory there exists four of such interpolation functions. For those exist an identity operator

$$\eta_0 = \eta(\alpha = 0) = \begin{cases} 1, & \text{multiplicative modifier, } (\kappa) \\ 0, & \text{additive modifier, } (\lambda). \end{cases} \tag{2.5.5}$$

One example of these interpolation functions that scales the bin count linearly over the known deviations $\eta_{-1} = \nu^{\text{1down}}_{scb}$ and $\eta_1 = \nu^{\text{1up}}_{scb}$ is

$$\eta_{\text{linear}}(\alpha) = \begin{cases} \alpha(\eta_0 - \eta_1), & \alpha > 0 \\ \alpha(\eta_0 - \eta_{-1}), & \alpha < 0 \end{cases} \tag{2.5.6}$$

This is illustrated in fig. 2.3(a). For the other ones see e.g. [15]. It is noted that α is the nuisance parameter and not the function $\eta(\alpha)$ and there is an associated constraint term c_α to each α.

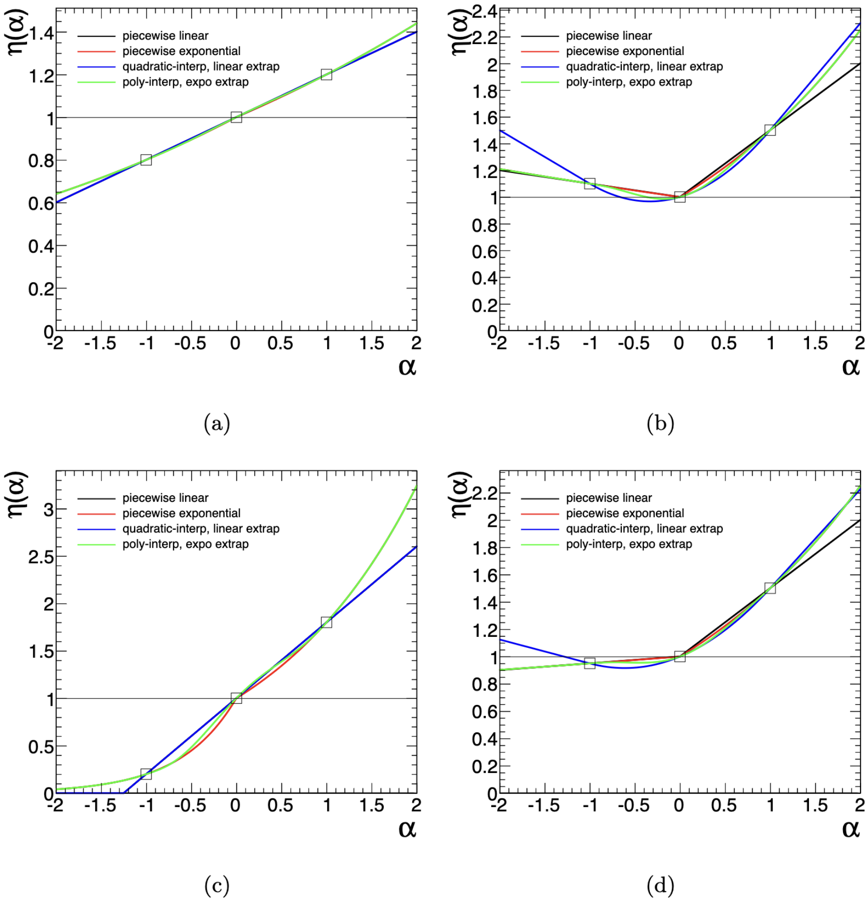


Figure 2.3: The four interpolation functions $\eta(\alpha)$ for different up and down standard deviation values. For example in (a) the bin count will be scaled with a factor of 0.8 for an $\alpha = -1$ (1.2 for an $\alpha = 1$). From [14].

Table 2.1: Modifiers and constraint terms used in HistFactory implemented by PYHF. Note that the interpolation functions are called f_p and g_p here instead of η as chosen in the full text. Taken from [8]

Description	Modification	Constraint Term c_χ	c_χ input
Uncorrelated Shape	$\kappa_{scb}(\gamma_b) = \gamma_b$	$\prod_b \text{Pois}\left(r_b = \sigma_b^{-2} \middle\| \rho_b = \sigma_b^{-2}\gamma_b\right)$	σ_b
Correlated Shape	$\Delta_{scb}(\alpha) = f_p\left(\alpha \middle\| \Delta_{scb,\alpha=-1}, \Delta_{scb,\alpha=1}\right)$	$\text{Gaus}\left(a = 0 \middle\| \alpha, \sigma = 1\right)$	$\Delta_{scb,\alpha=\pm1}$
Normalisation Unc.	$\kappa_{scb}(\alpha) = g_p\left(\alpha \middle\| \kappa_{scb,\alpha=-1}, \kappa_{scb,\alpha=1}\right)$	$\text{Gaus}\left(a = 0 \middle\| \alpha, \sigma = 1\right)$	$\kappa_{scb,\alpha=\pm1}$
MC Stat. Uncertainty	$\kappa_{scb}(\gamma_b) = \gamma_b$	$\prod_b \text{Gaus}\left(a_{\gamma_b} = 1 \middle\| \gamma_b, \delta_b\right)$	$\delta_b^2 = \sum_s \delta_{sb}^2$
Luminosity	$\kappa_{scb}(\lambda) = \lambda$	$\text{Gaus}\left(l = \lambda_0 \middle\| \lambda, \sigma_\lambda\right)$	$\lambda_0, \sigma_\lambda$
Normalisation	$\kappa_{scb}(\mu_b) = \mu_b$		
Data-driven Shape	$\kappa_{scb}(\gamma_b) = \gamma_b$		

2.6 The constraint terms

Uncertainties are modeled either Gaussian or Poissonian. The Gaussian implementation is straightforward as the uncertainty appears squared in the definition. For the interpolation function the nuisance parameter is scaled to the standard deviation values as described before $\text{Gauss}(\alpha|a, \sigma = 1)$.

For a Poissonian constraint to a multiplicative modifier γ_b, with a nominal (most probable) value $\gamma_0 = 1$, the Poisson distribution must be scaled with a factor f, so it reflects the original bin-count uncertainty σ. To find the corresponding Poisson distribution all parameters are multiplied by a factor f and is then solved for the one with the desired uncertainty. Since the variance of a Poissonian like eq. 2.1.3 is the rate parameter λ it follows

$$\text{Var}\left[\text{Pois}(k = f\gamma_0, \lambda = f\gamma)\right] = \lambda \overset{\gamma=\gamma_0}{=} f\gamma_0 = (f\sigma)^2 \quad \rightarrow \quad f = (1/\sigma^2). \quad (2.6.1)$$

This completes all the requirements needed for the creation of HistFactory models. The different types of modifiers and their constraint terms are summarized in table 2.1.

Part I

Results

Chapter 3

$HH \to 4b$ Results

3.1 Background validation

$$\frac{N(\text{CR, 2Xbb})}{N(\text{CR, 1Xbb})} = 0.0 \pm 0.0.$$

Appendices

Appendix A

Acronyms

CERN Organisation européenne pour la recherche nucléaire

ATLAS A Toroidal LHC Apparatus

SM Standard Model

QFT Quantum Field Theory

QCD Quantum Chromodynamics

QED Quantum Electrodynamics

EW Electroweak

EWSB Electroweak Symmetry Breaking

VEV Vacuum Expectation Value

CKM Cabibbo-Kobayashi-Maskawa

EM electromagnetic

IP impact parameter of tracks

ML Machine Learning

neos neural end-to-end-optimized summary statistics

HEP High Energy Physics

LHC Large Hadron Collider

HL-LHC High Luminosity LHC

ID Inner Detector

SCT semiconductor tracker

TRT transition radiation tracker

IBL insertable b-layer

HLT high level trigger

L1 Level-1

PFO Particle Flow Object

TCC Track CaloCluster

UFO Unified Flow Object

JES Jet Energy Scale

JER Jet Energy Resolution

JMR Jet Mass Resolution

GGF gluon-gluon fusion

VBF vector-boson fusion

NNLO next-to-next-to-leading order

N^{3}LO next-to-next-to-next-to-leading order

SR Signal Region

VR Validation Region

CR Control Region

KDE Kernel Density Estimation

bKDE binned Kernel Density Estimation

MC Monte Carlo

PDF Parton Density Function

PV primary vertex

JVT jet vertex tagger

NN Neural Network

ANN Artificial Neural Network

WP working point

VR variable radius

Appendix B

Cutflow

TODO, also fine like that?

Selection	Event	Fraction [%]	Total Fraction [%]
Initial	16854036422.000		
Preselections (MNT + Jet Cleaning)	670573995.000	100.000	100.000
PassTrigBoosted	63944638.000	9.536	9.536
PassTwoFatJets	57510800.000	89.938	8.576
PassTwoHbbJets	12875.000	0.0223	<0.001
PassVBFJets	5762.000	44.753	<0.001
PassFatJetPt	3902.000	67.720	<0.001
PassVBFCut	314.000	8.047	<0.001

Table B.1: Cut-flow table for data before signal region cut

Selection	Event	Fraction [%]	Total Fraction [%]
Initial	1475.226		
Preselections (MNT + Jet Cleaning)	547.960	100.000	100.000
PassTrigBoosted	20.926	3.819	3.819
PassTwoFatJets	14.141	67.576	2.581
PassTwoHbbJets	5.353	37.852	0.977
PassVBFJets	2.243	41.903	0.409
PassFatJetPt	1.408	62.793	0.257
PassVBFCut	0.148	10.539	0.027
PassSR	0.097	65.484	0.018
OverlapRemoval	0.059	61.200	0.011

Table B.2: Cut-flow table for DSID = 600463

Bibliography

[1] ATLAS Collaboration atlas. publications@ cern. ch, Georges Aad, B Abbott, DC Abbott, A Abed Abud, K Abeling, DK Abhayasinghe, SH Abidi, OS AbouZeid, NL Abraham, et al. Jet energy scale and resolution measured in proton–proton collisions at s= 13 tev with the atlas detector. *The European Physical Journal C*, 81(8):689, 2021. doi:10.48550/arXiv.2007.02645.

[2] The ATLAS collaboration. In situ calibration of large-radius jet energy and mass in 13 tev proton–proton collisions with the atlas detector. *The European Physical Journal C*, 79(2):135, 2019. doi:10.1140/epjc/s10052-019-6632-8. URL `https://doi.org/10.1140/epjc/s10052-019-6632-8`.

[3] ATLAS Collaboration. Measurement of the ATLAS Detector Jet Mass Response using Forward Folding with $80\,\mathrm{fb}^{-1}$ of $\sqrt{s} = 13\,\mathrm{TeV}$ *pp* data. ATLAS-CONF-2020-022, 2020. URL `https://cds.cern.ch/record/2724442`.

[4] ATLAS Collaboration. Efficiency corrections for a tagger for boosted $H \to b\bar{b}$ decays in *pp* collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. Technical report, CERN, Geneva, 2021. URL `https://cds.cern.ch/record/2777811`.

[5] The ATLAS collaboration. Particle Modeling Group systematic uncertainty recipes. URL `https://twiki.cern.ch/twiki/bin/view/AtlasProtected/PmgSystematicUncertaintyRecipes`. Last accessed: 2023-11-29.

[6] Francis Halzen, A Martin, and Leptons Quarks. An introductory course in modern particle physics. *John and Wiley*, 1984.

[7] Daniel de Florian, D Fontes, J Quevillon, M Schumacher, FJ Llanes-Estrada, AV Gritsan, E Vryonidou, A Signer, P de Castro Manzano, D Pagani, et al. *arXiv: Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*. Number arXiv: 1610.07922. Cern, 2016.

[8] Lukas Heinrich, Matthew Feickert, and Giordon Stark. pyhf: v0.7.2. URL `https://doi.org/10.5281/zenodo.1169739`. https://github.com/scikit-hep/pyhf/releases/tag/v0.7.2.

[9] Lukas Heinrich, Matthew Feickert, Giordon Stark, and Kyle Cranmer. pyhf: pure-python implementation of histfactory statistical models. *Journal of Open Source Software*, 6(58):2823, 2021. doi:10.21105/joss.02823. URL `https://doi.org/10.21105/joss.02823`.

[10] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *The European Physical Journal C*, 71:1–19, 2011.

[11] Olaf Behnke, Kevin Kröninger, Grégory Schott, and Thomas Schörner-Sadenius. *Data analysis in high energy physics: a practical guide to statistical methods*. John Wiley & Sons, 2013.

[12] Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1): 60–62, 1938.

[13] Alexander L Read. Presentation of search results: the cls technique. *Journal of Physics G: Nuclear and Particle Physics*, 28(10):2693, 2002.

[14] Kyle Cranmer, George Lewis, Lorenzo Moneta, Akira Shibata, and Wouter Verkerke. HistFactory: A tool for creating statistical models for use with RooFit and RooStats. Technical report, New York U., New York, 2012. URL `https://cds.cern.ch/record/1456844`.

[15] Lukas Heinrich. *Searches for Supersymmetry, RECAST, and Contributions to Computational High Energy Physics*. PhD thesis, New York University, 2019.

Statutory Declaration - Eidesstattliche Erklärung

I declare that I have authored this thesis independently, that I have not used other than the declared sources/ resources and that I have explicitly marked all materials which has been quoted either literally or by content form the used sources.

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Berlin, 06.12.2023

Frederic Renner