DISSERTATION

End to end optimization in a search for boosted Higgs boson pair production in the *bbbb* final state via vector-boson-fusion (VBF) production using the run 2 dataset with the ATLAS detector

---

For the attainment of the academic degree doctor rerum naturalium

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

M.Sc. Frederic Renner

Berlin, 07.07.2023

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

List of Figures

List of abbreviations

ATLAS   The Atlas Experiment
pdf   probability density function

0.1 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. The metric at hand to do so is the p-value that arises within hypothesis testing.

In the field of high-energy physics, a framework based on likelihood statistics 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. The following is based on [1–3].

0.1.1 Building the likelihood

The statistical model must take into account how compatible the observed collision events are with the theoretical predictions. This can be described by a likelihood $L(\boldsymbol{x}|\boldsymbol{\phi})$ that is just a probability for an observation $\boldsymbol{x}$ under a given set of parameters $\boldsymbol{\phi}$. Since we are dealing with a counting experiment the tool at hand will be bins of a histogram $\boldsymbol{h} = (h_1, ..., h_N)$.

It is useful to subdivide a measurement $\boldsymbol{x} = (\boldsymbol{n}, \boldsymbol{a})$ further into a quantity of interest $\boldsymbol{n}$, (e.g. the invariant mass of a particle) and *auxiliary* measurement histograms $\boldsymbol{a}$ that help to constrain the model. Additionally in the context of hypothesis testing it is useful to split 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 the parameter of interest will be only one parameter, the signal strength μ.

The bin heights (counts) can then be expressed in terms of the amount of signal $s_i(\boldsymbol{\Theta})$ and background $b_i(\boldsymbol{\Theta})$ in them depending in some way on the nuisance parameters. The prediction (expectation value) of the n_i can then be expressed as

$$\langle n_i(\mu, \boldsymbol{\Theta}) \rangle = \mu s_i(\boldsymbol{\Theta}) + b_i(\boldsymbol{\Theta}), \tag{0.1.1}$$

and similar for the a_i

$$\langle a_i(\boldsymbol{\Theta}) \rangle = u_i(\boldsymbol{\Theta}). \tag{0.1.2}$$

As we are expecting the bin counts to occur with a constant mean rate and independent of time compared to the last event, each bin follows a Poisson distribution

$$\frac{r^k e^{-r}}{k!}. \tag{0.1.3}$$

r is the expected rate of occurrences, which translates as our prediction, whereas k are the actual measured occurrences. Accounting for all the bins N by multiplying them together yields

$$L(\mu, \boldsymbol{\Theta}) = \prod_{j=1}^{N} \frac{(\mu s_j(\boldsymbol{\Theta}) + b_j(\boldsymbol{\Theta}))^{n_j}}{n_j!} e^{-(\mu s_j(\boldsymbol{\Theta}) + b_j(\boldsymbol{\Theta}))} \prod_{k=1}^{M} \frac{u_k(\boldsymbol{\Theta})^{a_k}}{a_k!} e^{-u_k(\boldsymbol{\Theta})}. \tag{0.1.4}$$

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

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\boldsymbol{\Theta}}})}{L(\hat{\mu}, \hat{\boldsymbol{\Theta}})} \tag{0.1.5}$$

The denominator is the unconditional maximum likelihood estimation so that $\hat{\mu}$ and $\hat{\boldsymbol{\Theta}}$ both are free to vary to maximize L, whereas the numerator is the found maximum likelihood conditioned on some chosen μ and the set nuisance parameters $\hat{\hat{\boldsymbol{\Theta}}}$ that maximize the likelihood for that particular μ. This definition gives $0 \leq \lambda \leq 1$. For a $\lambda \approx 1$ the hypothesized value of μ shows good agreement to the Poissonian model.

0.1.2   From test statistic to p-value

Transforming the profile likelihood into a test statistic t_μ is practical to calculate p-values

$$t_\mu = -2 \log \lambda(\mu). \tag{0.1.6}$$

This translates as $t_\mu \to 0$ as good agreement, $t_\mu \to \infty$ as bad agreement to the model. A right-tail p-value can then be calculated from the probability density

function of t_μ: $\text{pdf}(t_\mu) = f(t_\mu \mid \mu)$

$$p_\mu = \int_{t_{\mu,obs}}^{\infty} f(t_\mu \mid \mu)\mathrm{d}t_\mu \tag{0.1.7}$$

$t_{\mu,obs}$ is the test statistic t_μ evaluated at the observed data. This is like plugging into the Poisson distributions the same values for r as for k in eq. 0.1.3. Just like a probability density function for a standard normal distribution, intuitively the pdf is just, how probable a particular value of the test statistic t_μ is under a fixed value of the signal strength (how often it occurs compared to all other values t_μ can have).

This particular form is handy because there exist approximations for $f(t_\mu \mid \mu)$ [1]. Wald [4] proved that in the large sample limit the test statistic follows a normalized sum of squared distances between the tested parameter of interest μ_i and its maximum likelihood estimate $\hat{\mu}_i$. The result was extended by Wilk [5] for any number of parameters of interest so the test statistic becomes

$$t_\mu = \sum_i \frac{(\mu_i - \hat{\mu}_i^2)}{\sigma_i^2} + \mathcal{O}(1/\sqrt{N}). \tag{0.1.8}$$

The $\hat{\mu}_i$ are in the large sample limit normally distributed with mean μ' (true values) and standard deviation σ_i. This is the definition of a non-central chi-squared distribution with degrees of freedom i equal to the parameters of interest (see section 3.1 in [1]). For one parameter of interest the distribution reads

$$f(t_\mu \mid \mu) = \frac{1}{2\sqrt{t_\mu}} \frac{1}{\sqrt{2\pi}} \left[\exp\left(-\frac{1}{2}\left(\sqrt{t_\mu} + \sqrt{\Lambda_\mu}\right)\right) + \exp\left(-\frac{1}{2}\left(\sqrt{t_\mu} - \sqrt{\Lambda_\mu}\right)\right) \right], \tag{0.1.9}$$

with non-centrality parameter

$$\Lambda_\mu = \frac{(\mu - \mu')^2}{\sigma^2}. \tag{0.1.10}$$

Figure 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. Put differently, if the experiment would be repeated the p-value represents
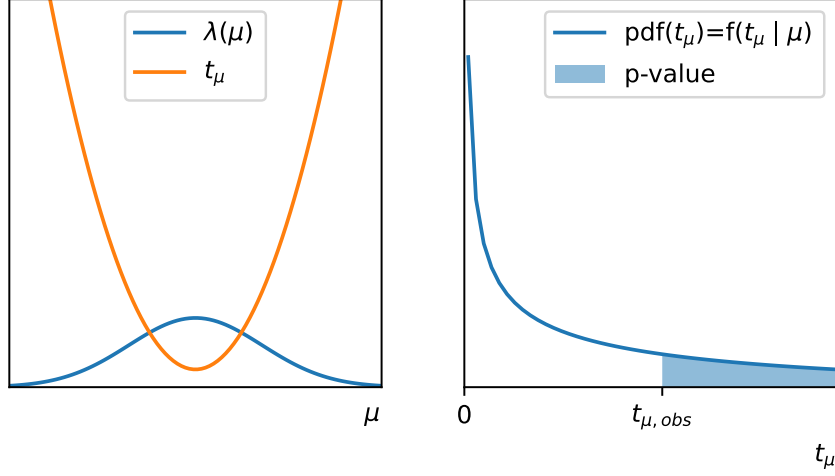
Figure 1: A sketch to follow the steps to calculate p-values. (**left**) The profile likelihood (■) has essentially some hill-like form with a maximum at $\lambda(\hat{\mu}, \hat{\boldsymbol{\Theta}})$, t_μ (■) is $-2\ln(\lambda)$. (**right**) For one parameter of interest in the large sample limit $f(t_\mu \mid \mu)$ follows a non-central chi-squared distribution with one degree of freedom, equation 0.1.9. The blue shaded area under the pdf is a right hand sided p-value.

the probability of that the outcome favors the alternative hypothesis over the null hypothesis.

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 $p < 2.87 \times 10^{-7}$ (5 standard deviations of the standard normal distribution). One caveat here is that this particular form of t_μ assumes μ can also be negative, which can be non-physical depending on the impact of a new process. Test statistics and their pdf approximations considering the different cases are covered in [1].

0.1.3 The CL$_s$ value

Particle physicists are usually interested in two things when making statistical tests for discovery of new phenomena: how well is the modeling of backgrounds (things we know) and if 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 on 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 the modeling of the backgrounds

$$\mathrm{CL}_s = \frac{p_{s+b}}{1 - p_b} = \frac{\int_{t_{\mu,obs}}^{\infty} f(t_\mu \mid \mu)\mathrm{d}t_\mu}{1 - \int_{t_{\mu,obs}}^{\infty} f(t_\mu \mid \mu)\mathrm{d}t_\mu}. \tag{0.1.11}$$

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 heavily cited CL_s paper that introduced the quantity [6] (see description of fig. 2).
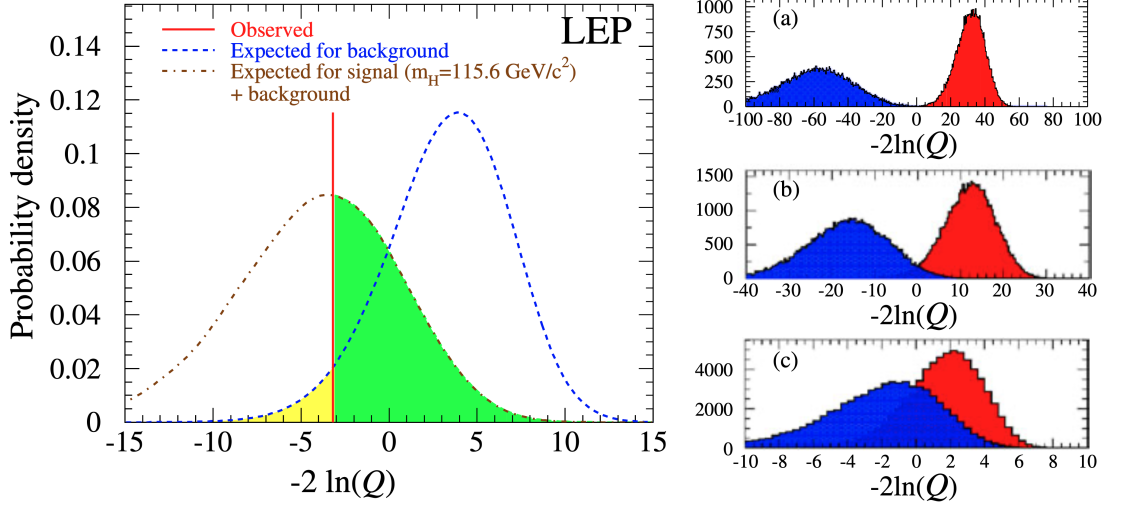
Figure 2: Probability density functions of test statistics from a Higgs search at LEP illustrating the calculation of p-values (λ becomes Q). (**left**) The pdf's of the test statistic $f(t_\mu \mid \mu)$ of the signal + background ($\diagup$) and background ($\diagup$) only hypotheses. The p-value is calculated by integration from $t_{\mu,obs}$ (the red observed line ($\diagup$)) to infinity (see eq. 0.1.7). The green shaded area ($\blacksquare$) corresponds to p_{s+b} whereas the yellow area ($\blacksquare$) corresponds to $1 - p_b$ since the integral over one whole pdf is 1. (**right**) Degradation of search sensitivity from (a) to (c). Note that the colors of the pdf's change here to signal + background ($\blacksquare$) and background only ($\blacksquare$). For example putting the observation ($t_{\mu,obs}$) 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. From [6].

0.1.4   The HistFactory model

A model used widely to achieve the preceding for a typical ATLAS analysis is called HistFactory [7] and is implemented within PYHF [8]. This section is based on the introduction to the model within the documenation of PYHF [3]. HistFactory draws the name from the idea that the problem is reducible to a small number of basic components. In order to build a likelihood like eq. 0.1.4 it is again useful to think of another splitting of the nuisance parameters $\boldsymbol{\phi}$ into

$$L(\boldsymbol{x}|\boldsymbol{\phi}) \quad = \quad L(\boldsymbol{x}|\overbrace{\boldsymbol{\psi}}^{\text{parameters of interest}}, \underbrace{\boldsymbol{\theta}}_{\text{nuisance parameters}}) \quad = \quad L(\boldsymbol{x}|\overbrace{\boldsymbol{\eta}}^{\text{free}}, \underbrace{\boldsymbol{\chi}}_{\text{constrained}}) \qquad (0.1.12)$$

6

free parameters $\boldsymbol{\eta}$, for example a cross section of a process and constrained parameters $\boldsymbol{\chi}$, that incorporate systematic uncertainties. The likelihood can then be modeled with Poissonians for each of the channels c (e.g. signal regions), bins b and constraint terms c_χ that again penalize the likelihood

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

(0.1.13)

The n_{cb} is the observation and $\nu_{cb}(\boldsymbol{\eta}, \boldsymbol{\chi})$ the prediction. The c_χ are calculated from auxiliary measurements a_χ to constrain χ and can be any function (e.g. Gaussian, Poissonian,...) we believe the parameter to be distributed.

The prediction is a sum of nominal bin height estimates[1] ν_{scb}^0 over all samples s (e.g. $t\bar{t}$, multijet-background etc.). These nominal bin height estimates have associated errors so we could modify the bin heights according to some belief we think this affects the likelihood. To modify ν_{scb}^0 we use multiplicative κ_{scb} and additive factors Δ_{scb}

$$\nu_{cb}\left(\boldsymbol{\phi}\right) = \sum_{s \in \text{ samples}} \nu_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right) \tag{0.1.14}$$

$$= \sum_{s \in \text{ samples}} \underbrace{\left(\prod_{\kappa \in \boldsymbol{\kappa}} \kappa_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right)\right)}_{\text{multiplicative modifiers}} \left(\nu_{scb}^0\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right) + \underbrace{\sum_{\Delta \in \boldsymbol{\Delta}} \Delta_{scb}\left(\boldsymbol{\eta}, \boldsymbol{\chi}\right)}_{\text{additive modifiers}}\right). \tag{0.1.15}$$

How this modification affects the likelihood depends then also on our modeling of the error. This is governed by the auxiliary measurments c_χ and will be explained later.

For intuition let us consider just one systematic uncertainty to a nominal bin height estimate ν_{scb}^0. If modifying ν_{scb}^0 by μ while the constraint term $c_\chi(\mu)$ (note the dependence on μ) stays small, it can be beneficial for our goal of maximizing the likelihood. This means we find the most likely/compatible value to the observed data within our modeling of the errors.

---

[1]People call this also the rate as in the definition of the Poisson distribution.

The Modifiers

In HistFactory there are by convention four types $\{\lambda, \mu, \gamma, \alpha\}$ of such multiplicative rate modifiers that will be explained throughout the rest of 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

$$\nu_{scb}(\mu) = \mu \nu_{scb}^0. \tag{0.1.16}$$

These are bin-independent and shape(of the histogram)-preserving normalization factors and could be for example the signal strength from above.

Further there are **bin-wise modifiers** γ_b (uncorrelated shape)

$$\nu_{scb}(\gamma_b) = \gamma_b \nu_{scb}^0. \tag{0.1.17}$$

These are useful for example to include the Monte Karlo uncertainties or 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

$$\nu_{scb}(\alpha) = \eta(\alpha) \nu_{scb}^0, \tag{0.1.18}$$

and additive versions

$$\nu_{scb}(\alpha) = \nu_{scb}^0 + \eta(\alpha). \tag{0.1.19}$$

This is useful to include systematic uncertainties. In a typical ATLAS analysis the effect of some standard deviation to a final to be calculated quantity is usually inferred at some step by plugging in the upper ('1up') or lower ('1down') standard deviation values (pretending they were the nominal values entering the calculation) and observe the impact on a final quantity. The interpolation function transports this knowledge into the likelihood model by scaling just like before the nominal

value ν_{scb}^0 and at the same time applying a penalization c_α according to our modeling of the errors.

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

$$\eta_0 = \eta(\alpha = 0) = \begin{cases} 1, & \text{multiplicative modifier, } (\kappa) \\ 0, & \text{additive modifier, } (\lambda). \end{cases} \tag{0.1.20}$$

One example of these interpolation function that scales the bin height linear over the known deviations $\eta_{-1} = \nu_{scb}^{1\text{down}}$ and $\eta_1 = \nu_{scb}^{1\text{up}}$ is

$$\eta_{\text{linear}}(\alpha) = \begin{cases} \alpha(\eta_0 - \eta_1), & \alpha > 0 \\ \alpha(\eta_0 - \eta_{-1}), & \alpha < 0 \end{cases} \tag{0.1.21}$$

This is illustrated in fig. 3(a). For the other ones the reader is referred to e.g. [9]. Remember α is the nuisance parameter and not the function $\eta(\alpha)$ and there is an associated constraint term c_α to each α that uses Gaussian and Poissonian modeling. The Guassian implementation is straightforward as the nuisance parameter is scaled to the standard deviation values as described before Gauss= $(\alpha \mid a, \sigma = 1)$.

For a poissonian constraint to a multiplicative factor γ_b, with a nominal (most probable) value $\gamma_0 = 1$, we must scale the Poisson distribution with a factor f so it reflects the original bin-height uncertainty σ. We basically need to multiply all parameters of the Poisson distribution with a factor f and then solve for the one with the desired uncertainty. Since the variance of a Poissonian like eq. 0.1.3 is just the rate parameter λ it follows

$$\text{Var}\left[\text{Pois}(k = f\gamma_0, \lambda = f\gamma)\right] = \lambda \overset{\gamma = \gamma_0}{=} f\gamma_0 = (f\sigma)^2 \quad \rightarrow \quad f = (1/\sigma^2). \tag{0.1.22}$$

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 1.
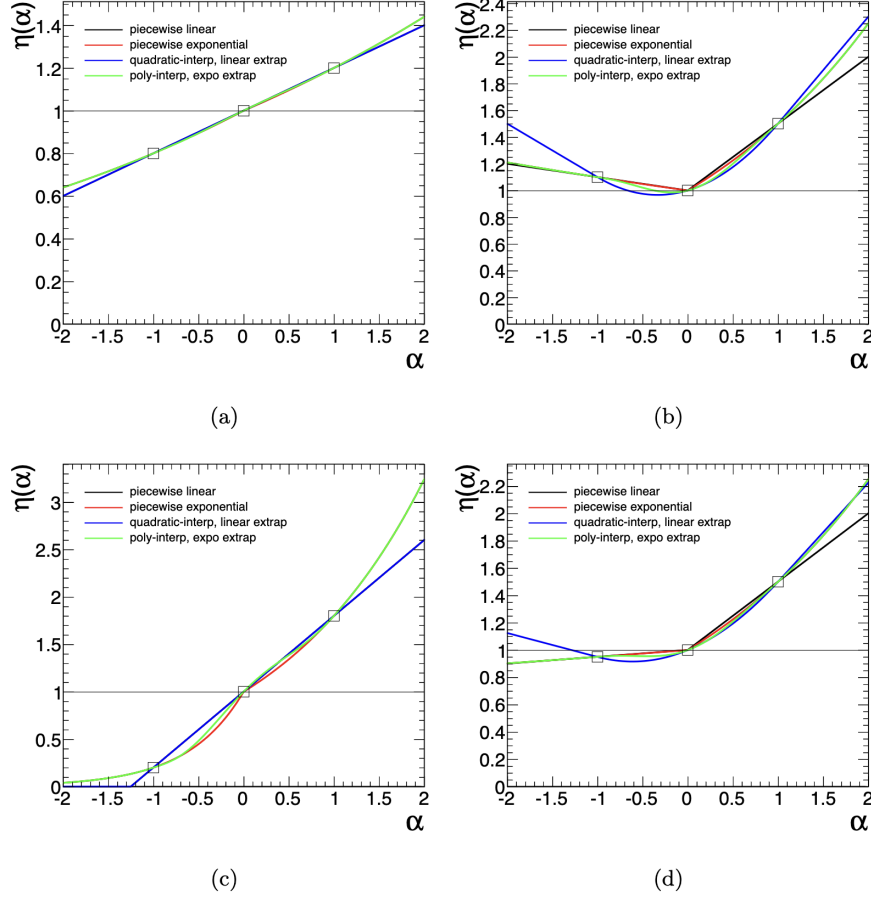
Figure 3: The four Interpolation functions $\eta(\alpha)$ for different up and down standard deviation values. For example in (a) the bin height will be scaled with a factor of 0.8 for an $\alpha = -1$ (1.2 for an $\alpha = 1$). From [7].

Table 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. From [3]

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	$\Delta_{scb}(\alpha) = f_p\left(\alpha\| \Delta_{scb,\alpha=-1}, \Delta_{scb,\alpha=1}\right)$	$\text{Gaus}\left(a = 0\| \alpha, \sigma = 1\right)$	$\Delta_{scb,\alpha=\pm 1}$
Normalisation Unc.	$\kappa_{scb}(\alpha) = g_p\left(\alpha\| \kappa_{scb,\alpha=-1}, \kappa_{scb,\alpha=1}\right)$	$\text{Gaus}\left(a = 0\| \alpha, \sigma = 1\right)$	$\kappa_{scb,\alpha=\pm 1}$
MC Stat. Uncertainty	$\kappa_{scb}(\gamma_b) = \gamma_b$	$\prod_b \text{Gaus}\left(a_{\gamma_b} = 1\| \gamma_b, \delta_b\right)$	$\delta_b^2 = \sum_s \delta_{sb}^2$
Luminosity	$\kappa_{scb}(\lambda) = \lambda$	$\text{Gaus}\left(l = \lambda_0\| \lambda, \sigma_\lambda\right)$	$\lambda_0, \sigma_\lambda$
Normalisation	$\kappa_{scb}(\mu_b) = \mu_b$		
Data-driven Shape	$\kappa_{scb}(\gamma_b) = \gamma_b$		

Bibliography

[1] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, The European Physical Journal C **71**, 1 (2011).

[2] O. Behnke, K. Kröninger, G. Schott, and T. Schörner-Sadenius, *Data analysis in high energy physics: a practical guide to statistical methods* (John Wiley & Sons, 2013).

[3] L. Heinrich, "Introduction to model building within pyhf," `https://pyhf.readthedocs.io/en/v0.7.2/intro.html` (2021), [Online; accessed 27-June-2023].

[4] A. Wald, Transactions of the American Mathematical society **54**, 426 (1943).

[5] S. S. Wilks, The annals of mathematical statistics **9**, 60 (1938).

[6] A. L. Read, Journal of Physics G: Nuclear and Particle Physics **28**, 2693 (2002).

[7] K. Cranmer, G. Lewis, L. Moneta, A. Shibata, and W. Verkerke (ROOT), "HistFactory: A tool for creating statistical models for use with RooFit and RooStats," Tech. Rep. (New York U., New York, 2012).

[8] L. Heinrich, M. Feickert, G. Stark, and K. Cranmer, Journal of Open Source Software **6**, 2823 (2021).

[9] L. Heinrich, *Searches for Supersymmetry, RECAST, and Contributions to Computational High Energy Physics*, Ph.D. 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, 07.07.2023

_____

Frederic Renner