Chapter 7
Statistical Treatment

This chapter contains a detailed description of the statistical procedures implemented in the analysis to extract the final results. The likelihood function and the statistical procedures used to test the fit results are described. In addition this chapter illustrates the tools used to scrutinize and validate the fit results.

7.1 General Statistical Treatment

7.1.1 The Likelihood Function

To measure the signal yield, a binned maximum likelihood fit is performed. The observable used in the fit as final discriminant is the large-R jet mass m_J. The fit is done simultaneously to all the analysis regions and the three lepton channels extracting the VH and VZ contributions. The likelihood is defined as the product over all bins of the Poisson probability to observe n_i events when m_i events are expected in a certain bin i:

$$\begin{aligned}\mathcal{L}_{Pois}(\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\tau}) &= \prod_{i\in\text{bins}} \text{Pois}(n_i|m_i(\boldsymbol{\mu},\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\tau})) \\ &= \prod_{i\in\text{bins}} \text{Pois}(n_i|\mu_{VH}s_i(\boldsymbol{\alpha}) + \mu_{VZ}b_i^{VZ}(\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\tau}) + b_i^{oth}(\boldsymbol{\alpha},\boldsymbol{\gamma},\boldsymbol{\tau})) \\ &= \prod_{i\in\text{bins}} \frac{(\mu_{VH}s_i + \mu_{VZ}b_i^{VZ} + b_i^{oth})^{n_i}}{n_i!} e^{-(\mu_{VH}s_i+\mu_{VZ}b_i^{VZ}+b_i^{oth})}\end{aligned} \tag{7.1}$$

where $\boldsymbol{\mu} = (\mu_{VH}, \mu_{VZ})$ and the number of expected events m_i in bin i is obtained summing the expected signal s_i and background events b_i. The contribution of the

G. Di Gregorio, *A Beauty-ful Boson*, Springer Theses,
https://doi.org/10.1007/978-3-031-20013-7_7

expected background is split in contribution of the VZ background b_i^{VZ} and contribution of the remaining backgrounds b_i^{oth} ($b_i = \mu_{VZ} b_i^{VZ} + b_i^{oth}$). The expected signal s_i and VZ background b_i^{VZ} events are multiplied by the parameters μ_{VH} and μ_{VZ}, respectively. The parameter μ_{VH} (μ_{VZ}) is referred to as *signal strength* and it is defined as the ratio of the measured cross-section time branching ratios $\sigma \times$BR for the VH (VZ) process divided by its SM expectation. The signal strength parameters are also called Parameter of Interests (PoIs). In the analysis presented in this thesis, a multi-PoIs fit is performed since the values of μ_{VH} and μ_{VZ} signal strengths are extracted simultaneously. In particular this convention is used: the (x+y) PoIs fit indicates a simultaneous fit in which x is the number of VH PoIs and y the number of VZ PoIs. This means that the multi-PoIs fit described in the following is a (1+1) PoIs fit.

In addition to the parameters μ_{VH} and μ_{VZ}, the likelihood depends on other parameters $\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\tau}$ called Nuisance Parameters (NPs). The NPs encode the dependence of the prediction on the systematic uncertainties into continuous parameter in the likelihood. The NPs can be categorized into three classes: $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, ..)$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, ..)$ and $\boldsymbol{\tau} = (\tau_1, \tau_2, ..)$. The $\boldsymbol{\tau}$ NPs are unconstrained parameters controlling the normalisation of the backgrounds and they are called *free-floating* because they are free to float in the fit. The $\boldsymbol{\gamma}$ NPs represent the statistical uncertainties caused by the limited size of simulated background samples. The signal process is usually chosen to not be affected by the $\boldsymbol{\gamma}$ NPs as the statistical uncertainties on the predicted signal simulation are small with respect to the backgrounds. A γ NP is applied in each bin of the analysis on the sum of all the backgrounds. The modelling and the experimental uncertainties enter in the fit through the $\boldsymbol{\alpha}$ NPs and they affect both signal and background events.

The $\boldsymbol{\alpha}$ NPs are estimated from data or auxiliary measurements which provide both central values and uncertainties. For each $\boldsymbol{\alpha}$ NP, the likelihood function is multiplied by an *auxiliary term* that constrains the value of the systematic uncertainty around its estimate, within the uncertainty on such estimate. The auxiliary terms are Gaussian functions with mean equal to zero and variance equal to one:

$$\mathcal{L}_{aux}(\boldsymbol{\alpha}) = \prod_{\alpha \in \boldsymbol{\alpha}} \text{Gauss}(\boldsymbol{\alpha}|0, 1) = \prod_{\alpha \in \boldsymbol{\alpha}} \frac{1}{\sqrt{2\pi}} e^{-\alpha^2/2} \tag{7.2}$$

where the product is extended over all the systematic uncertainties considered in the analysis. The NPs are defined such that for $\alpha_j = 0$ the nominal predictions are obtained and for $\alpha_j = \pm 1$ two modified templates, called *up/down template*, for the $\pm 1\sigma$ variation are obtained. The NPs are expressed in unit of their uncertainties σ_α.[1]

The uncertainties on the background predictions due to the limited number of simulated events are also accounted in the likelihood function considering Poisson terms:

[1] This means, for example, that moving the parameter α of the jet energy scale by 1 corresponds in shifting the energy calibration of the jet by 1σ.

$$\mathcal{L}_{Stat}(\boldsymbol{\gamma}) = \prod_{i \in \text{bins}} \frac{(\gamma_i b_i)^{b_i}\, e^{-(\gamma_i b_i)}}{\Gamma(b_i + i)} \tag{7.3}$$

where

$$\Gamma(x) = \int dt\, t^{x-1} e^{-t} \tag{7.4}$$

For each bin i of the histograms a γ_i NP is considered and it represents the uncertainty on the sum of all the background processes in that bin. Dedicated studies have shown that, while the statistical uncertainty of the VH sample is small, the one of VZ is much more in-line with the other backgrounds. Consequently, in the multi-PoIs fit the VZ is treated like all the other backgrounds for the γ parameters.

The full likelihood used in the final fit can be schematically written as:

$$\mathcal{L}(\mu, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\tau}) = \mathcal{L}_{Pois}(\mu, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\tau}) \cdot \mathcal{L}_{aux}(\boldsymbol{\alpha}) \cdot \mathcal{L}_{Stat}(\boldsymbol{\gamma}) \tag{7.5}$$

A binned likelihood fit is performed to determine the PoIs and their uncertainties. The measured signal strengths and the NPs are obtained as the values of the parameters that maximize the likelihood function $\mathcal{L}(\mu, \boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\tau}) = \mathcal{L}(\mu, \boldsymbol{\theta})$ or, equivalently, minimize $-\ln \mathcal{L}(\mu, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents the set of NPs introduced previously, $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\tau})$. The likelihood maximization without fixing the values of signal strengths is called *unconditional fit*. Instead, the *conditional fit* is performed maximising the likelihood for particular values of the signal strengths.

The uncertainties on the signal strengths is determined evaluating $\mu_{+/-} = \hat{\mu}^{+\sigma_\mu^+}_{-\sigma_\mu^-}$ as the value that satisfies this equation:

$$-2 \ln \frac{\mathcal{L}(\mu_{+/-}, \hat{\hat{\boldsymbol{\theta}}})}{\mathcal{L}(\hat{\mu}, \hat{\boldsymbol{\theta}})} = 1 \tag{7.6}$$

where $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$ are the parameters that maximise the overall likelihood and $\hat{\hat{\boldsymbol{\theta}}}$ are the NP values that maximise the likelihood for a particular value of μ.

7.1.2 Profile Likelihood Ratio and Test Statistic

The Profile Likelihood Ratio (PLR) $\lambda(\mu)$ is defined as the ratio of two Likelihood functions:

$$\lambda(\mu) = \frac{\mathcal{L}(\mu, \hat{\hat{\boldsymbol{\theta}}}(\mu))}{\mathcal{L}(\hat{\mu}, \hat{\boldsymbol{\theta}})} \tag{7.7}$$

where $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$ are the parameters that maximise the overall likelihood and $\hat{\hat{\boldsymbol{\theta}}}(\mu)$ are the NP values that maximise the likelihood for a particular value of μ as men-

tioned previously. The PLR is used to construct the *test statistic* $q_\mu = -2\ln\lambda(\mu)$, which differentiates the background-only hypothesis with $\mu = 0$ from the alternative hypothesis with $\mu > 0$. The PLR takes values $0 \leq \lambda(\mu) \leq 1$, where large values close to unity imply good agreement between the hypothesised signal strengths μ and the observed data.

A test statistic used in the $VH(b\bar{b})$ analysis is the one for the discovery of a positive signal in which the background-only hypothesis with $\mu = 0$ is tested. The compatibility of the data with the background-only hypothesis is evaluated from the test statistic $q_0 = -2\ln\lambda(\mu = 0) = -2\ln\frac{\mathcal{L}(\mu=0,\hat{\hat{\boldsymbol{\theta}}}(\mu=0))}{\mathcal{L}(\hat{\mu},\hat{\boldsymbol{\theta}})}$. If the data are compatible with the background-only hypothesis, the nominator and the denominator of the test statistic are similar and q_0 is close to 0. Differently, if the event yield is larger than the expectation, the test statistics q_0 assumes larger values indicating higher incompatibility between the data and the tested hypothesis. The incompatibility can be expressed with the p-value, in this case named p_0, defined as:

$$p_0 = \int_{q_{o,obs}}^{\infty} f(q_0|0)dq_0 \tag{7.8}$$

where $q_{o,obs}$ is the value of the test statistic measured from the observed data and $f(q_0|0)$ is the probability density function of the test statistic q_0 under the $\mu = 0$ assumption. The hypothesis of a test can be considered excluded if its p-value is observed below a specific threshold. The p-value can be expressed in terms of the significance Z which is defined such that a Gaussian distributed variable found Z standard deviations above its mean has an upper-tail probability equal to p_0 (Fig. 7.1). In a more formal way, the significance is the quantile (inverse of the cumulative distribution) of the standard Gaussian, computed for $(1 - p_0)$:

$$Z = \Phi^{-1}(1 - p_0) \tag{7.9}$$

The rejection of the background hypothesis with a significance of at least Z = 5 (which correspond to $p_0 = 2.87 \times 10^{-7}$) is considered as an appropriate level to quote a discovery.

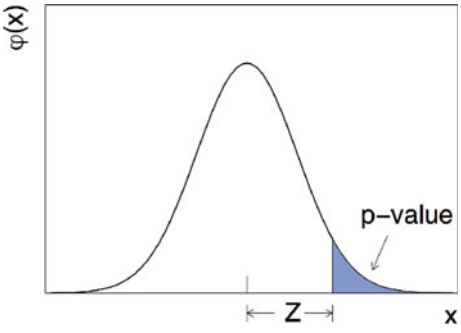

Fig. 7.1 Relationship between a p-value and a significance of Z sigma [3]

To compute the p-value, the distribution of $f(q_0|0)$ is required. This can be achieved by sampling the distribution exploiting the Monte Carlo method [1]. However, the procedure is computationally expensive and approximate solutions are adopted. Assuming the null hypothesis to be true, the Wilk's theorem [2] ensures that q_0 is asymptotically distributed as a χ^2 with one degree of freedom. This means that the value of q_0 can be easily compared to the χ^2 value. With few steps it can be shown that the significance can be $Z = \sqrt{q_0}$. In the following, all the statistical tests are done using the asymptotic approximation ensured by the Wilk's theorem.

7.2 Fit Input

The signal and control regions used in the fit have been summarised in Table 5.5. The following processes are considered in the fit, either as signal or backgrounds:

- signal VH, $H \to b\bar{b}$ (summed over all the production modes);
- Z+jets and W+jets. The V+jets backgrounds are split into three different components depending the flavour composition of the two jets used to reconstruct the Higgs boson decay, V+HF, $V + cl$, $V + ll$;
- $t\bar{t}$;
- single-top: s-, t- and Wt-channels. The s- and t-channels are treated as one component, while the Wt-channel is treated independently;
- diboson: WW, ZZ, WZ. The WZ and ZZ processes are treated as one component and they have an associated PoI in the fit;
- multi-jet in 1-lepton channel. The multi-jet contribution in 0- and 2-lepton channel is negligible.

Signal and background m_J templates are determined from the MC simulation in all the cases except fot the multi-jet background in the 1-lepton channel, whose contribution is extracted from the data.

The likelihood is built from the m_J histograms for each process listed above. The choice of using different binnings and ranges for the m_J distribution has been made to maximise the resolution taking into account the following aspects:

- avoid empty bins in the templates;
- minimise empty bins in the data distributions;
- have a statistical uncertainty in each bins lower than 20% to avoid potential biases on μ.

7.3 Nuisance Parameters

The impact of all the experimental and modelling uncertainties affecting the m_J templates is quantified using histograms that correspond to $\pm 1\sigma$ variation of each specific NP. The up $(+1\sigma)$ and down (-1σ) variations are calculated relative to the

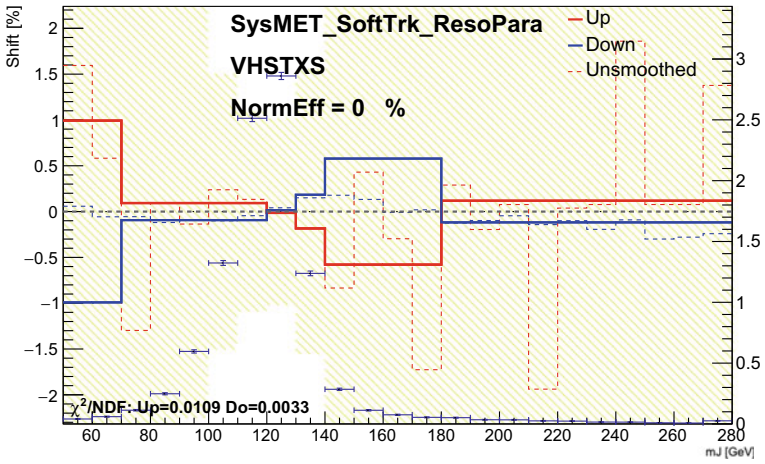

Fig. 7.2 Effect of the smoothing procedure on the m_J distribution for the signal sample in LP SR, $p_{\mathrm{T}}^{V} \geq 400$ GeV in 1-lepton channel

nominal template. There are few cases where the systematic variation lead only to $+(-)1\sigma$ effect. The jet resolution uncertainty, for example, is a one-sided uncertainty and consequently its effect is symmetrised with respect to the nominal histogram in order to have also the variation in the other direction. In all these cases the effect of the systematic is symmetrized with respect to the nominal histogram to obtained the variation also in the other direction.

Smoothing and Pruning of the Systematic Uncertainties

Certain systematic uncertainties such as the large-R jet energy scale uncertainties can cause bin migrations of events in the m_J distribution. This migration of the events causes large statistical fluctuations which are not physical. To prevent these effects, the *smoothing* procedure is adopted to all the systematic variation across all regions. In the first step, bins are merged until there is one maximum in the varied distribution relative to the nominal distribution. In the second step, bins are furthermore grouped until the statistical uncertainty in each bin is below 5%. Figure 7.2 shows the effect of the smoothing on a $E_{\mathrm{T}}^{\mathrm{miss}}$ systematic for the signal VH sample. The red and blue lines show the shifts of the variation of the systematic with respect to the nominal (referred to the left y-label). The dotted lines represents the variations of the systematic before the smoothing while the continuous lines the variations of the systematic after the smoothing procedure. The points with the error bars show the m_J distribution of the signal sample, referring to the right y-label.

From the total list of NPs, only some of them have a sizeable impact on the fit templates. To reduce the number of NPs in order to obtain a more solid fit model, a *pruning* procedure is applied. The procedure removes systematics uncertainties that have a negligible impact on the final result. Normalisation and shape uncertainties are dropped if the variation of the corresponding template is below 0.5% in all the bins.

Additional pruning criteria are applied in all the analysis regions where the signal contribution is less than 2% of the total background and the systematic variation impact of the total background is less than 0.5%.

7.4 Tool for the Validation of the Fit Results

All the tools described in the following are used to understand the statistical analysis and to identify potential problems and errors in the fit.

7.4.1 Pull Plots

The information of the fit results can be visualised using plots. The *pull plots* shows the pull of the nuisance parameters which is the comparison of the central value and uncertainty of the nuisance parameters before and after the fit [3]. The pull of the NP θ, with an expectation value θ_0 and standard deviation σ_θ, is defined as:

$$pull(\theta) = \frac{\hat{\theta} - \theta_0}{\sigma_\theta} \tag{7.10}$$

where $\hat{\theta}$ is the NP value obtained from the maximum likelihood fit. The pull quantifies how far from its expected value the NP is "pulled" by the fit in number of σ_θ. A healthy situation is when the pull is zero, if this is not the case, further investigation is required. If the pull is not zero, the NP value extracted from the fit is different from the expected NP value.

In the pull plots the parameters corresponding to the floating normalisations are also shown but following a different convention. The value shown in the plot is not the pull since the floating normalisations do not have any prior, but it is the absolute value of the normalisation with its uncertainty.

A possible estimate of the error of the NP can be performed studying the PLR as function of the parameter θ around $\hat{\theta}$. The estimate is done applying the same method used to evaluate the μ uncertainty. For the NPs with a Gaussian constraint in the likelihood, the expected interval of the pull is $[-1, +1]$. If the interval is smaller than the expected one, the performed measurement is more accurate with respect to the auxiliary measurement. In this case, the systematic uncertainty is "constrained" by the data and it needs to be understood.

Once all the systematics are considered inside the fit, the first fit is performed using the *Asimov dataset*. The *Asimov dataset* [2] has as data the expected yields predicted from the simulation and they are used in replacement of the real data to test the fit performance and to quantify the expected sensitivity. By definition, the value of the pull from the fit to the Asimov dataset will not change, it can be only constrained.

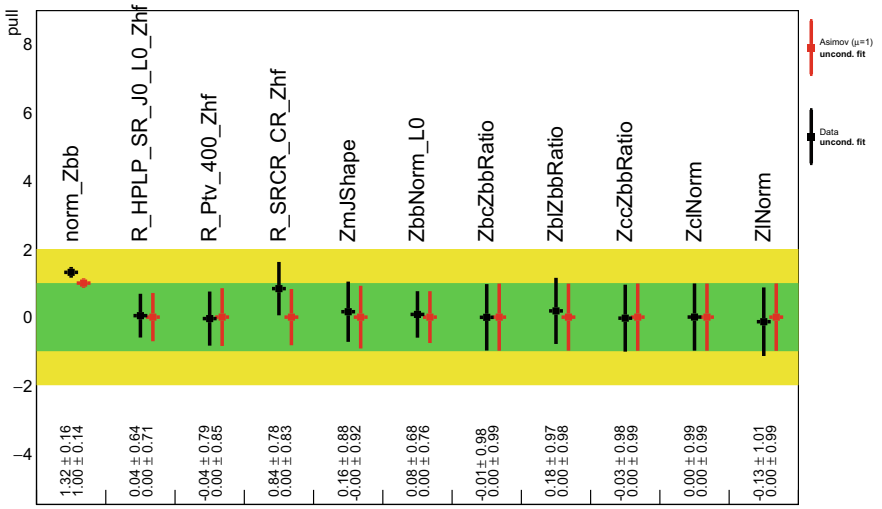

Fig. 7.3 Example of pull plot obtained from an unconditional fit to Asimov (*red dots*) and real dataset (*black dots*)

The pulls of NPs with a Gaussian constraints are set to zero, while the values of the floating normalisations are equal to 1. The use of the Asimov dataset is important to spot suspicious behaviours and to predict the expected precision of the floating normalisation factors. Figure 7.3 shows an example of pull plot in which there is a comparison of the pulls obtained from the unconditional fit applied to Asimov (red dots) and to real dataset (black dots).

Another way to study the stability of the fit without using the information of the PoIs is to perform a conditional fit with $\mu = 1$. With this fit the value of the PoIs is fixed but it is possible to extract information on the pulled NPs.

7.4.2 Correlation Matrix

Another tool used for the validation of the fit model is the correlation matrix. The correlation between θ_i and θ_j NPs or PoIs is obtained from the covariance matrix of the estimator of all the parameters, $V_{ij} = \text{cov}[\hat{\theta}_i, \hat{\theta}_j]$. In the large sample limit, the covariance matrix is defined as the inverse of the second derivative of the log-likelihood function evaluated at $\hat{\mu}$ and $\hat{\theta}$ [2]:

$$\text{cov}[\theta_i, \theta_j] = \left[-\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \theta_i \partial \theta_j} \right]^{-1} \tag{7.11}$$

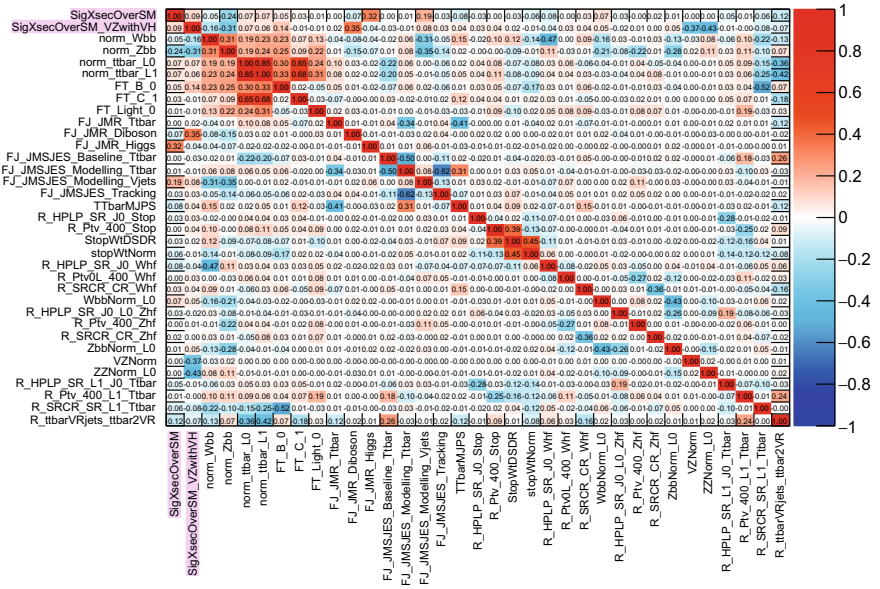

Fig. 7.4 Correlation matrix obtained from a fit to the Asimov dataset

The value of the correlation coefficients can vary from -1 to $+1$. If two variables are not related, their correlation is zero. Figure 7.4 shows an example of correlation matrix obtained from a fit to the Asimov dataset. The correlation matrix contains also the correlation coefficients between the PoIs (highlighted in magenta) and the NPs. Since some NPs are correlated, the correlation matrix helps to understand why some NPs are constrained or pulled. To simplify a bit the plot, only NPs that have the absolute value of the correlation of magnitude 0.25 or higher with an other NP are shown. In general the NPs have a small correlation with the exception of few cases as the correlation among the normalisation factors and the correlation among the large-R jet systematics (see Fig. 7.4).

7.4.3 Ranking Plot

An important information of the fit is how much the PoI value varies when changing the value of an NP. The impact of a NP θ on the fitted PoI is defined as [3]:

$$\text{impact} = \Delta\mu^{\pm} = \hat{\hat{\mu}}_{\theta_0 \pm \sigma_\theta} - \hat{\mu} \tag{7.12}$$

where $\hat{\mu}$ is the value maximising the likelihood and $\hat{\hat{\mu}}_{\theta_0 \pm \sigma_\theta}$ is the value of the PoI extracted from a fit where all the NPs are allowed to vary except for θ which is

fixed to the values at the edge of the intervals of the pulls, $\theta_0 \pm \sigma_\theta$. Each NP has its impact and not all the NPs are equally important. The *ranking plot* is used to sort the NPs with the largest impact. Figure 7.5 shows an example of ranking plot obtained from a fit to the Asimov dataset, in which all the uncertainties are listed in a decreasing order of their impact. The plot shows only 15 NPs with the highest

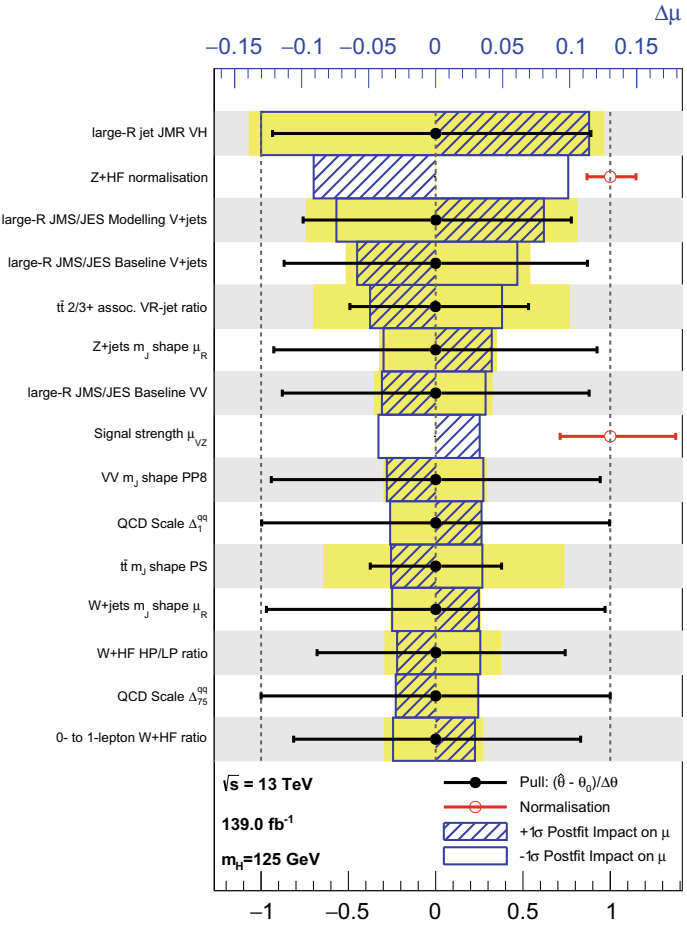

Fig. 7.5 Impact of the uncertainties on the VH signal strength μ_{VH} parameter obtained from a fit to the Asimov dataset. The uncertainties are sorted in a decreasing order. The *boxes* show the variation of $\hat{\mu}_{VH}$, referring to the top x-axis, when fixing the corresponding NP. The impact of the up- and down-variations can be distinguished via the dashed and plane box fillings. The *yellow boxes* show the pre-fit impact, referring to the top x-axis, by varying each NP by $\pm 1\sigma$. The filled black points with the corresponding error bars show the pull of each NP, referring to the bottom x-axis. The *open red circles* with the error bars show the fitted values and the uncertainties of the normalisation factors which are freely floating in the fit. The *dotted vertical* are placed at ± 1 and are referred to the bottom x-axis

impact. The boxes show the impact of the NP on $\hat{\mu}_{VH}$, referring to the top x-axis. The hatched and plane boxes represent the up- and down-variation, respectively. The yellow band shows the pre-fit impact, referring to the top x-axis, by varying each nuisance parameter by $\pm 1\sigma$. The filled black circles with the error bars show the pulls of the NPs and their uncertainties, referring to the bottom x-axis. The open red circles with the error bars show the fitted values and uncertainties of the floating normalisations. By definition of fit to Asimov dataset, all the black points are set to zero and all the red points are set to one. The dotted vertical lines are referred to the bottom x-axis and placed at ± 1. In the ranking plot shown in Fig. 7.5 the highest ranked parameters are the parameters that shows the largest correlation to the signal strength.

7.4.4 Breakdown of the Uncertainties

The uncertainties with similar origin can be grouped together to study the uncertainty impact of the group on the fitted signal strength. In this way it is possible to find which systematics have a big impact on the measurement precision. The uncertainty impact of a group of uncertainties is the result of the comparison of the uncertainties on the signal strengths:

$$\text{uncertainty impact} = \sqrt{\sigma^2_{\hat{\mu}} - \sigma^2_{\hat{\mu}'}} \tag{7.13}$$

where $\sigma^2_{\hat{\mu}}$ is the uncertainty on the signal strength obtained from the nominal fit[2] and $\sigma^2_{\hat{\mu}'}$ is the uncertainty on the signal strength running a fit with all the NPs belonging to a group fixed to their best fit values. When testing the impact of the systematic on one PoI, the understudy PoI is fixed to the value extracted from the nominal fit while the other PoI is left floating in the fit. The "total statical" impact is evaluated comparing the result of the nominal fit with the result of the fit with all the NPs fixed to their best fit values except for the floating normalisations. The "data stat only" impact is defined as the comparison between the nominal fit and the fit with all the NPs fixed to their nominal expectation values. The "floating normalisation" contribution is the quadratic difference between the total error and the error from the fit with only the normalisation factors fixed to the best fit values. The "total systematic" impact is the quadratic difference between the total error and the "total statistical" error. The sum in quadrature of the individual contributions of the systematic uncertainties differs from the total systematic contribution due to correlations between the NPs.

[2] The nominal fit indicates the maximum likelihood fit in which all the NPs and PoIs are left floating.

References

1. Metropolis N, Ulam S (1949), The Monte Carlo method. J Am Stat Assoc 44: 335. https://doi.org/10.1080/01621459.1949.10483310
2. Cowan G, Cranmer K, Gross E, Vitells O (2011) Asymptotic formulae for likelihood-based tests of new physics. Eur Phys J C 71:1554. https://doi.org/10.1140/epjc/s10052-011-1554-0 arXiv:1007.1727 [physics.data-an], Erratum: Eur Phys J C 73: 2501 (2013)
3. Gross E (2018) Practical statistics for High Energy Physics. In: CERN Yellow Reports. School Proceedings, vol 3, ed. by Mulders M, Zanderighi G 199, https://doi.org/10.23730/CYRSP-2018-003.199