On the presentation of correlated systematic uncertainties in Higgs boson rate measurements

P. Bechtle, T. Stefaniak

November 12, 2013

1 Motivation

For an application like `HiggsSignals` [1], or in future combinations of Higgs data from different experiments, one important question is how to treat the correlated systematics. This question arises specifically in two different aspects:

1. The contributions of the different Higgs production and decay modes to the relative rate measurements $\hat{\mu}$ in the search (sub)channels, in the following called observables, may vary when using models differently from the SM. In this case, the uncertainty matrix $\mathbf{C}$ of all observables needs to be re-evaluated.

2. For the interpretation of the Higgs rate measurements in different models the (correlated) contribution of the model-dependent theory uncertainty needs to be disentangled from other uncertainty sources, since they need to be treated differently from the SM.

If $\mathbf{C}$ is not publicly available (as is currently the case), or if the public information about $\mathbf{C}$ is not detailed enough, the above tasks can not be fulfilled with full precision. Therefore, we like to make a proposal for the presentation of correlated systematic uncertainties in Higgs boson rate measurements, inspired by Ref. [2] and the procedure employed in the LEP Higgs combination [3, 4]. Including this information in global analyses of the Higgs signal strength measurements is expected to be of great importance, in particular in the future, as statistical uncertainties are expected to be reduced faster than theoretical and presumably also experimental systematic uncertainties.

2 Covariance matrix decomposition

Under the assumption, that experimental systematics and theory uncertainties can be treated as Gaussian distributed uncertainties with only linear correlations[1], the covariance matrix of N signal rate measurements, $\hat{\mu}_i$ $(i = 1, \ldots, N)$, is given by

$$\mathbf{C} = C_{ii'} = \rho_{ii'}\sigma_i\sigma_{i'}, \tag{1}$$

[1] For a large number of uncertainty sources, the central limit theorem suggests that this assumption approximately holds, no matter what the *unknown* probability density function of a given theory uncertainty truly is.

where $\rho_{ii'}$ is the correlation matrix for the different observables i and i', that are in this case the measured rates in different (sub)channels of Higgs searches. This and all following covariance matrices preferably contain only *relative* uncertainties with respect to the measured signal strengths, such that the absolute uncertainty is given by $\sigma_i = \Delta\sigma_i \cdot \hat{\mu}_i$, where $\Delta\sigma_i$ is the relative uncertainty. If an absolute systematic signal strength uncertainty is present in the measurement, which is independent of $\hat{\mu}_i$, relative and absolute uncertainties should be decomposed.

However, Eq. (1), as simple as it seems, is neither directly useable for expressing uncertainties of a variable number of measurements, nor for testing models that predict signal strength contributions of the various signal topologies (i.e. in most cases the various production modes included in an analysis) different than in the SM. In addition, it is not possible to decompose $\mathbf{C}_{ii'}$ or $\rho_{ii'}$ into different uncertainty sources, such as experimental systematics (independent of the model) and theoretical uncertainties (model-dependent). Therefore, we propose to employ an extended set of covariance matrices, describing the uncertainties and their correlations in a complete way and for each analysis individually. This procedure relies only on the assumptions of Gaussianity of the probability density functions and linearity of the correlations, but not on implicit model assumptions.

In addition to i and i' denoting the observable in the (sub)channels, let $j = 1, ..., J$ enumerate the different channels (or signal topologies) considered within an analysis.[2] Let furthermore $k = 1, ..., K$ denote the different completely uncorrelated uncertainty sources (e.g. statistics), $l = 1, ..., L$ the different $+100\,\%$ correlated uncertainty sources (like typically a lepton energy scale (LES) uncertainty) and $m = 1, ..., M$ the different $-100\,\%$ correlated uncertainty sources (like typically a tagging efficiency of a given category, where an increased number of events in one category corresponds to a reduced number of events in an orthogonal category). Not all observables need necessarily to be affected by all uncertainty sources, and some observables might have opposite correlations for the same uncertainty source. Note also that, while several observables might approximately be fully correlated with respect to a given uncertainty source, they do not need to have the same relative uncertainty. This makes the ansatz proposed here so general: The $\rho_{ii'}$ in Eq. (1) are generally not only taking the values $0, -1$ or $+1$ but anything in between since they are the result of a convolution of many different error sources, which contribute with different strength to each measurement. However, the correlation factor for each individual error source k, l or m on any single measurement μ_i^j can usually accurately be described as $0, -1$ or $+1$. The aim of the following discussion is to provide a framework to reliably provide the necessary information to calculate $\mathbf{C}$ for every possible signal composition (of the signal topologies j) in each observable i and the resulting, potentially different theory uncertainty of the signal rates. In other words, this proposed procedure enables the derivation of the covariance matrix for every testable model different to the SM.

We first introduce some basic definitions for the total signal strength of an observable and how it can be decomposed in channel signal strengths and according weights. The weighted signal strengths for the channels j of the signal strength measurement i is defined by

$$\alpha_i^j = \omega_i^j \mu^j, \tag{2}$$

[2] For an analysis targeting a single Higgs decay mode, these are typically the five LHC Higgs production modes, $\{\mathrm{ggH}, \mathrm{VBF}, WH, ZH, t\bar{t}H\}$. That could be further generalized, though.

where we using the definitions (following Ref. [1])

$$\mu^j = \frac{[\sigma \times \mathrm{BR}]^j}{[\sigma_{\mathrm{SM}} \times \mathrm{BR}_{\mathrm{SM}}]^j}, \tag{3}$$

and

$$\omega_i^j = \frac{\epsilon_i^j \, [\sigma_{\mathrm{SM}} \times \mathrm{BR}_{\mathrm{SM}}]^j}{\sum_{j'} \epsilon_i^{j'} \, [\sigma_{\mathrm{SM}} \times \mathrm{BR}_{\mathrm{SM}}]^{j'}} \tag{4}$$

for the individual signal strength and SM weights for the channel j. The SM weights carry also an index i since they depend on the signal efficiencies ϵ_i^j of the analysis i in the channel j. Using these definitions, the measured signal strength can be decomposed as

$$\mu_i = \frac{\sum_j \epsilon_i^j \, [\sigma \times \mathrm{BR}]^j}{\sum_j \epsilon_i^j \, [\sigma_{\mathrm{SM}} \times \mathrm{BR}_{\mathrm{SM}}]^j} = \sum_j \omega_i^j \mu^j = \sum_j \alpha_i^j, \tag{5}$$

assuming the efficiencies are identical for the investigated model and the SM.[3]

In order for the proposed procedure to work, note that the uncertainty sources k, l and m need to be labeled unambiguously. Then it is possible to report for each observable i individually, to what extent the channel j reacts to the uncertainty sources k, l or m, namely with the uncertainty $\sigma_i^{j,\{k,l\,\mathrm{or}\,m\}} = \mu^j \Delta\sigma_i^{j,\{k,l\,\mathrm{or}\,m\}}$. This information is independent of all other analyses and can be combined with other observables in an unambiguous way for any given signal composition.

For each analysis, the required values $\Delta\sigma_i^{j,\{k,l\,\mathrm{or}\,m\}}$ could be deduced for every analysis channel i by letting each signal contribution j vary in the profile likelihood limit setting individually. Then, the variation of the best fit $\hat{\alpha}_i^j = \omega_i^j \hat{\mu}^j$, (as opposed to just measuring $\hat{\mu}_i$ in the analysis (sub)channel i in the case of the full profile likelihood analysis) can be observed for upward variations and downward variations of each nuisance parameter $\theta^{k,l\,\mathrm{or}\,m}$. For ranges of values of $\theta^{k,l\,\mathrm{or}\,m}$ which represent about a $+1\sigma$ variation of that nuisance parameter, the median value of $\langle \alpha_i^j \rangle_{\mathrm{up}}$ is observed. Likewise, for a range of values of $\theta^{k,l\,\mathrm{or}\,m}$ which represent a -1σ variation, $\langle \alpha_i^j \rangle_{\mathrm{down}}$ is observed. Then, the uncertainty on the systematics associated with $\theta^{k,l\,\mathrm{or}\,m}$ can be approximated as $\sigma_i^{j;k,l\,\mathrm{or}\,m} = (\langle \alpha_i^j \rangle_{\mathrm{up}} - \langle \alpha_i^j \rangle_{\mathrm{down}})/2$. The correlations $+1$ or -1 can be deduced by the upward or downward variation of $\hat{\alpha}_i^j$ for the same $\theta^{k,l\,\mathrm{or}\,m}$. This procedure might be expensive on the computing side due to the required high sampling density for several different values of $\theta^{k,l\,\mathrm{or}\,m}$, but is in principle straightforward.

As a first approximation the output can be constrained to the dominant sources of systematic uncertainties while neglecting all sources with minor impact on the given measurement. A computationally cheaper, albeit potentially less accurate method of extracting the $\Delta\sigma_i^{j,\{k,l\,\mathrm{or}\,m\}}$ values would be to fix each individual $\theta^{k,l\,\mathrm{or}\,m}$ consecutively, once to its $+1$ or -1σ values, as observed in a conventional profile likelihood fit. Then, the best fit values $\hat{\alpha}_i^j|_{\mathrm{up}}$ and $\hat{\alpha}_i^j|_{\mathrm{down}}$ could be observed directly for each individual profile likelihood fit and $\Delta\sigma_i^{j,\{k,l\,\mathrm{or}\,m\}}$ could be calculated from them as above.

[3] The formalism can easily be generalized for the case where this assumption is not fulfilled.

The uncorrelated part of the (weighted) uncertainty matrix on the absolute errors on the measured signal contributions of signal j to each observable i can be written as:

$$(\mathbf{C}^{jj',k}_{0\%})_{ii'} = {\alpha_i^j}^2 {\Delta\sigma_i^{j,k}}^2 \; \delta_{ii'}\delta_{jj'}. \tag{6}$$

Due to the α_i^j introduced here, any signal composition predicted by the model can be accommodated. For the +100% correlated uncertainties, we can write

$$(\mathbf{C}^{jj',l}_{+100\%})_{ii'} = \alpha_i^j \Delta\sigma_i^{j,l} \alpha_{i'}^{j'} \Delta\sigma_{i'}^{j,l}. \tag{7}$$

Note that the product $\alpha_i^j \alpha_{i'}^j$ should not be mistaken as a correlation factor. The correlation coefficients $\rho_{ii'}^{jj',l}$ are all +1, hence the diagonal entries of the matrices $\mathbf{C}^{jj',l}$ are weighted by the same factors of α as the off-diagonal entries. Likewise, we write for the −100% correlated part,

$$(\mathbf{C}^{jj',m}_{-100\%})_{ii'} = \pm\; \alpha_i^j \Delta\sigma_i^{j,m} \alpha_{i'}^{j'} \Delta\sigma_{i'}^{j,m}, \tag{8}$$

where the + [-] sign is used for $i = i'$ [$i \neq i'$].

The covariance matrices given in Eq. (6)-(8) are given in the most general form. They still hold in the case, where the relative uncertainties $\Delta\sigma_i^{j,\{k,l\,\mathrm{or}\,m\}}$ are approximately the same for all channels j considered within the measurement, i.e. $\Delta\sigma_i^{j,\{k,l\,\mathrm{or}\,m\}} \approx \Delta\sigma_i^{k,l\,\mathrm{or}\,m}$. If this approximation holds[4], the amount of information necessary from the experiments reduces significantly, but the formalism does not change.

In addition to the re-weighting by α (accounting for changed signal compositions), the individual uncertainties $\Delta\sigma_i^{j,\{k,l\,\mathrm{or}\,m\}}$ can also be varied, which is necessary if, for instance, the uncertainties of the model predictions are different than in the SM. Using the above decomposition scheme, the covariance matrices $\mathbf{C}'^{jj'}$ for the signal topologies j and j' can be reconstructed at each point in the parameter space of the model by

$$\mathbf{C}'^{jj'} = \sum_k \mathbf{C}^{jj',k}_{0\%} + \sum_l \mathbf{C}^{jj',l}_{+100\%} + \sum_m \mathbf{C}^{jj',m}_{-100\%} \equiv \rho'^{jj'}_{ii'} \sigma'^{jj'}_i \sigma'^{jj'}_{i'}\,. \tag{9}$$

From this, the overall covariance matrix for the signal rate observables, defined as $\mathbf{C}' = {\rho'}_{ii'}\sigma'_i\sigma'_{i'}$, is simply given for any given model point by

$$\mathbf{C}' = \sum_{j,j'} \mathbf{C}'^{jj'}. \tag{10}$$

Thus, an individual, re-weighted uncertainty matrix for the observables can be reconstructed unambiguously at any model point, taking all individual correlations into account.

Note, that in general neither $\rho'_{ii'}$ nor σ'_i are identical to the SM quantities in the combined approach in Eq. (1). To summarize, the specific advantages of the proposed procedure are:

1. All uncertainty sources can be individually re-weighted, e.g. theoretical uncertainties, for each production mode.

[4] Note, that the approximation does obviously not hold in the presence of strong correlations *among* the channels j of a single measurement, as e.g. typically introduced by a tagging efficiency, see the discussion above.

2. All uncertainties can be re-weighted for each channel/signal topology (i.e. in most cases for each Higgs production mode) in each observable, according to a different signal composition as predicted by the model.

All that is needed for this procedure to work is, that the information $\Delta\sigma_i^{j,\{k,l \text{ and } m\}}$ for the uncorrelated, fully correlated, and fully anti-correlated relative uncertainties, respectively, of the channel/signal topologies j considered in the measurement i, is made publicly available.

References

[1] P. Bechtle, S. Heinemeyer, O. Stål, T. Stefaniak and G. Weiglein, "HiggsSignals: Confronting arbitrary Higgs sectors with measurements at the Tevatron and the LHC," arXiv:1305.1933 [hep-ph].

[2] A. Read, "Higgs statistics menu", Talk given at *Higgs Days at Santander 2013*.

[3] R. Barate *et al.* [LEP Working Group for Higgs boson searches and ALEPH and DELPHI and L3 and OPAL Collaborations], "Search for the standard model Higgs boson at LEP," Phys. Lett. B **565** (2003) 61 [hep-ex/0306033].

[4] S. Schael *et al.* [ALEPH and DELPHI and L3 and OPAL and LEP Working Group for Higgs Boson Searches Collaborations], "Search for neutral MSSM Higgs bosons at LEP," Eur. Phys. J. C **47** (2006) 547 [hep-ex/0602042].