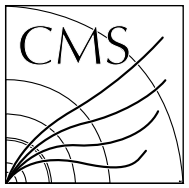
Available on CMS information server

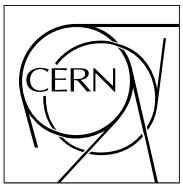
CMS NOTE -2017/001



The Compact Muon Solenoid Experiment

CMS Note

Mailing address: CMS CERN, CH-1211 GENEVA 23, Switzerland



19 January 2017 (v4, 20 January 2017)

Simplified likelihood for the re-interpretation of public CMS results

The CMS Collaboration

Abstract

In this note, a procedure for the construction of simplified likelihoods for the re-interpretation of the results of CMS searches for new physics is presented. The procedure relies on the use of a reduced set of information on the background models used in these searches which can readily be provided by the CMS collaboration. A toy example is used to demonstrate the procedure and its accuracy in reproducing the full likelihood for setting limits in models for physics beyond the standard model. Finally, two representative searches from the CMS collaboration are used to demonstrate the validity of the simplified likelihood approach under realistic conditions.

Contents

1 Introduction 2
2 Simplified likelihood 4
2.1 Defining the simplified likelihood 4
2.2 Example of constructing the simplified likelihood 6
3 Aggregated search regions 15
3.1 Example of the use of aggregated search regions 15
4 Validation of the simplified likelihood using CMS searches for BSM physics . . . 17
4.1 Monojet search 17
4.2 α_T search 20
5 Summary 22

1 Introduction

Searches for physics beyond the standard model (BSM) by the CMS collaboration [1] are performed using a wide variety of strategies and across different final states with varied kinematic properties. Often, the results of these searches are presented in terms of "model-independent" limits on the production cross-section for some hypothesised BSM particle. Common examples are searches for resonances whose decay products can be experimentally reconstructed with a high resolution resulting in narrow invariant mass peaks that can be readily distinguished from a smoothly varying background (as in [2, 3]. Many searches, however, involve the use of final states with low mass resolution or quantities such as the missing transverse momentum (as in [4, 5]) or the angles between objects in the final state (as in [6]). These searches are typically performed using the distributions of quantities for which the separation between standard model (SM) processes and BSM signals is limited. Furthermore, sensitivity to a wide range of BSM physics can often be improved using a categorisation of events based on the number of a particular object in the event, such as charged leptons or jets, or based on a multivariate analysis (MVA) of the kinematics and/or reconstruction and identification quality of the final state particles in the event. For such searches, limits can only be expressed in terms of the parameter space of some specific complete or simplified BSM scenario.

In the original CMS publications, the results of searches for BSM physics are often interpreted using a small subset of new physics models serving as benchmarks for the sensitivity of the search. Usually, the searches are re-interpreted to provide constraints on other models of new physics, not included in the publication. Re-interpretations can also be provided using complete models of new physics and the constraints from the searches are combined with measurements and searches from other experiments [7]. For these re-interpretations, the signal contribution is typically determined using an event generator such as PYTHIA [8] followed by a simulation of the detector response and resolution using tools such as DELPHES [9] or by using published efficiencies for generator level quantities (both methods are used in [10, 11]). The background contributions to the search regions and the associated systematic uncertainties often rely on simplifying assumptions, in particular where the search is performed using multiple event categories or the distributions of one or more discriminating variables, which can lead to inaccuracies in the re-interpretation.

In previous CMS publications, the total background predictions and systematic uncertainty have been provided by searches in each region for which the contribution from potential BSM signals is expected to be significant. These searches may also define "super" regions which cover larger regions of the relevant discriminating variables than those used in the analysis (as in [12, 13]). This allows re-interpretation of the results by selecting the most sensitive super region for a given BSM interpretation. This procedure avoids the need to provide correlations between the distributions of the discriminating variables. While the procedure is robust, the loss of information included in the regions neglected for each BSM scenario can result in a significant loss of sensitivity.

In this note, an alternative procedure for re-interpreting BSM physics searches by approximating the full background model and the associated systematic uncertainties is presented. The procedure uses a reduced set of information to describe the background model and the correlations between different regions used in the searches, minimising the loss of sensitivity. Using two representative searches, the procedure is validated to demonstrate that the assumptions of the simplified likelihood hold for realistic applications.

The following key points are discussed in this note:

- Many BSM CMS searches are sensitive to BSM models not discussed in the corresponding CMS publications.
- Often these searches are based on the distribution of one or more discriminating variables and commonly performed using event counts in many disjoint search regions. Although the event counts, background expectations, and background uncertainties are provided for each search region, re-interpretation of the results in different contexts is not possible without knowledge of the full background model.
- To facilitate re-interpretation CMS can make available an approximate covariance matrix for the background estimates in the various search regions, or a reduced set thereof.
- This covariance matrix is used to build a simplified likelihood for any signal model, as described in Section 2.1.
- Based on this likelihood, approximate limits on BSM models not considered in the original CMS publication can be extracted under a number of different statistical treatments, the choice of which is left to the user.
- Alternatively, for simplicity, the user may decide to define their own single search region tailored to the BSM model of interest. Provided this single search is defined as the union of a number of analysis search regions, the covariance matrix can be used to extract the total uncertainty on the background expectation in the single search region.
- Any re-interpretation using the simplified procedures outlined in this note can only approximate the result of a full CMS analysis, due to the imperfections in estimating the BSM acceptance and detector resolutions as well as their systematic uncertainties from simplified detector models and the underlying assumptions of the procedure itself.

A key ingredient for re-interpretations is the ability to predict the number of events expected in each search region for a given BSM model. The details of the production of signal model predictions and uncertainties will vary for different BSM searches and are not discussed within this note.

2 Simplified likelihood

While several methods for extracting limits are used at CMS in searches for BSM physics, a common approach follows the procedures described in Refs. [14] and [15]. The observed data are interpreted using a likelihood formalism in which the likelihood function is defined as

$$\mathcal{L}(\mu, \boldsymbol{\theta}) = \mathcal{P}(\text{data}|\mu \cdot s(\boldsymbol{\theta}) + b(\boldsymbol{\theta})) \cdot p(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta}), \tag{1}$$

where $\mathcal{P}(\text{data}|\mu \cdot s(\boldsymbol{\theta}) + b(\boldsymbol{\theta}))$ is the probability to observe the data, whether consisting of individual events or event counts in bins in one or more discriminating variables or event categories. The parameters $\boldsymbol{\theta} = (\theta_1, \theta_2 ...)$ are nuisance parameters, which are used to model the variation of the signal $s(\boldsymbol{\theta})$ and background $b(\boldsymbol{\theta})$ contributions due to systematic uncertainties. Often, these nuisance parameters are constrained by external measurements, $\tilde{\boldsymbol{\theta}}$, which are encoded in the probability density function, $p(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$. The parameter μ, typically referred to as signal strength, is a common scale factor for the expected signal contribution. A particular BSM model predicting $s(\boldsymbol{\theta})$ is said to be excluded at some confidence level when every value of $\mu \geq 1$ is excluded at least at that confidence level.

In order to achieve a good sensitivity to a wide range of BSM models, searches are often performed by categorising events with different final states or according to some discriminating variable. The precise modelling of the backgrounds can involve many nuisance parameters making the full likelihood too detailed to describe in a publication. The following describes a procedure for using a reduced set of information to re-interpret the results of searches for new physics through the use of a simplified likelihood.

2.1 Defining the simplified likelihood

In practice, the likelihood defined in Equation 1 can include regions in which negligible signal is expected under a wide range of BSM models. Typically these are referred to as "control" regions in that they allow the constraint of nuisance parameters that cause large variations in the background model. In this note, the distinction is made between these regions and "search" regions, which instead are expected to include significant contributions from the signal, under some particular type of BSM models. Search regions are defined by a set of criteria used to select events. These criteria can include categorisations based on the ranges of a number of a certain type of object in the event such as jets or charged leptons, and intervals in some discriminating variable such as the missing transverse momentum in the event. The data in each search region, i, is characterised by a single number, n_i, which is the observed number of events. The likelihood is therefore constructed from a product of counting experiments, where each counting experiment represents a search region. For a given search region, i, the probability to observe n_i events is

$$P(n_i|\mu \cdot s_i + b_i) = \frac{(\mu \cdot s_i + b_i)^{n_i} e^{-(\mu \cdot s_i + b_i)}}{n_i!}, \tag{2}$$

where s_i and b_i are the total expected signal and background contributions.[1]

In most cases, the background contribution in each search region will not be known with perfect accuracy and is therefore subject to systematic uncertainties. These uncertainties are mod-

[1] In the case that the search region R_i is a bin or interval in a distribution of some observable x, for which the signal and background models $s(x)$ and $b(x)$ are continuous functions of x the values of s_i and b_i are taken as $s_i = \int_{R_i} s(x)dx$ and $b_i = \int_{R_i} b(x)dx$.

elled by modifying the background contributions as $b_i \rightarrow b_i + \theta_i$. The probability to simultaneously observe each of the n_i events in N search regions is the product of probabilities across the search regions such that

$$\mathcal{P}(\text{data}|\mu \cdot s(\boldsymbol{\theta}) + b(\boldsymbol{\theta})) = \prod_{i=1}^{N} P(n_i|\mu \cdot s_i + b_i + \theta_i). \tag{3}$$

Some simplifying assumptions must be made in order to reduce the complexity of the full probability density function $p(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$.

- The constraints on the background contributions are Gaussian such that the distribution of the number of background events is symmetric about the expectation, b_i, and its variance is independent of $\boldsymbol{\theta}$. Often, the background contributions are estimated from control regions in data with large sample sizes, which makes this assumption valid.
- The covariance, and therefore only the linear correlation, between the background contribution in each region is sufficient to approximate $p(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$ at least for values of $\boldsymbol{\theta}$ which are close to $\tilde{\boldsymbol{\theta}}$.
- The numbers of events, n_i, are statistically independent from one another. This is true when there are no events which are included in more than one search region and the estimates of the background contributions, b_i, and covariance matrix $\mathbf{V}$ have not been obtained from data which are statistically dependent on the data from any search region.
- The systematic uncertainties in the signal model can be neglected. The validity of this assumption will strongly depend on the specific BSM physics model being considered. Systematic uncertainties on the signal can be accounted for by adding appropriate nuisance parameters with Gaussian constraints as for the background contributions.

Under these assumptions, $p(\tilde{\boldsymbol{\theta}}|\boldsymbol{\theta})$ can be modelled as a multivariate Gaussian distribution with the mean vector identified with external measurements of $\tilde{\boldsymbol{\theta}} = 0$. The simplified likelihood can now be expressed as

$$\mathcal{L}_S(\mu, \boldsymbol{\theta}) = \prod_{i=1}^{N} \frac{(\mu \cdot s_i + b_i + \theta_i)^{n_i} e^{-(\mu \cdot s_i + b_i + \theta_i)}}{n_i!} \cdot \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T \mathbf{V}^{-1} \boldsymbol{\theta}\right), \tag{4}$$

where $\mathbf{V}$ represents the covariance matrix for the expected background contributions across the search regions and is defined as

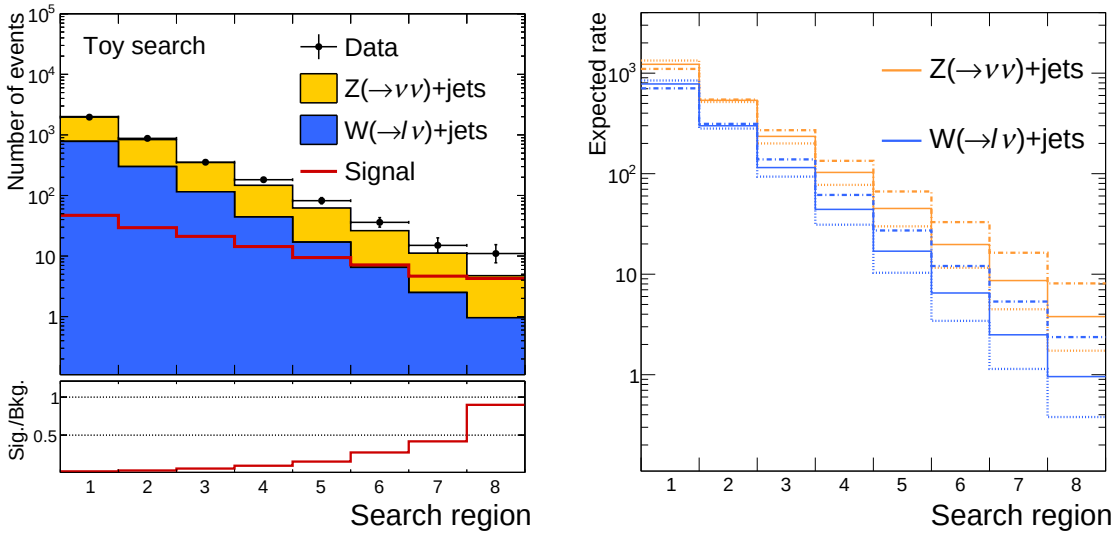
$$V_{ij} = E[\theta_i \times \theta_j], \tag{5}$$

where $E[x]$ is the expectation value of x.

For the results shown for the following example, the simplified likelihood is constructed using the statistics package ROOFIT [16] but it should be noted that the procedure does not prohibit the use of any other tool for its construction.

2.2 Example of constructing the simplified likelihood

A toy example of a search for BSM physics is represented in Figure 1. The search is performed using 8 mutually exclusive search regions which vary in the expected contributions from backgrounds and a potential BSM physics signal. Here we assume background contributions from two sources, in the following labelled $Z(\to \nu\nu) + \text{jets}$ and $W(\to l\nu) + \text{jets}$. Figure 1 (a) indicates the number of observed events in each of the search regions as well as the expected contribution from the two different sources of background, stacked on top of one another. The background estimates include systematic uncertainties related to theoretical uncertainties, when the backgrounds are derived using simulation, and experimental uncertainties related to the resolutions and response of the detector. The uncertainties can also include a statistical component due to limited sample sizes of events in control regions used to estimate the backgrounds. In this example, two systematic uncertainties are included, one affecting the $Z(\to \nu\nu) + \text{jets}$ background only and one affecting the $W(\to l\nu) + \text{jets}$ background only. These two systematic uncertainties are modelled as two nuisance parameters in the full likelihood of Equation 1. Figure 1 (b) shows the variation of the expected contribution due to the systematic uncertainties. The dotted and dot–dashed histograms show the expected contributions when each of the two nuisance parameters is varied positively and negatively by one sigma, respectively. Finally, the expected contribution from a BSM physics signal is shown. The ratio of the expected signal to the total background estimate in each region is shown in the lower panel of Figure 1 (a). In this example, the signal is provided, however, the signal can be exchanged for some other BSM physics process using a number of publicly available event generators and tools that simulate the response and resolution of the CMS detector. The details of producing signal estimates is beyond the scope of this note.



(a) (b)

Figure 1: (a) Number of data, expected background and expected signal in the 8 search regions. The lower panel shows the ratio of the expected signal to the total expected background. (b) Uncertainties in the two background contributions. The dotted and dot–dashed histograms show the expected contributions when each of the two nuisance parameters is varied positively and negatively by one sigma, respectively.

The full likelihood in this toy example is given by

$$\mathcal{L}(\mu, \phi_{\mathrm{W}}, \phi_{\mathrm{Z}}) = \left[\prod_{i=1}^{N} \mathrm{P}\left(n_i | \mu \cdot s_i + B_i^{\mathrm{Z}}(\phi_{\mathrm{Z}}) + B_i^{\mathrm{W}}(\phi_{\mathrm{W}})\right)\right] \cdot e^{-\frac{1}{2}\phi_{\mathrm{Z}}^2} \cdot e^{-\frac{1}{2}\phi_{\mathrm{W}}^2}, \tag{6}$$

where $B_i^{\mathrm{Z}}(\phi_{\mathrm{Z}})$ and $B_i^{\mathrm{W}}(\phi_{\mathrm{W}})$ are the expected contributions from the $\mathrm{Z}(\to \nu\nu) + \text{jets}$ and $\mathrm{W}(\to l\nu) + \text{jets}$ backgrounds in the i–th signal region, which are functions of nuisance parameters ϕ_{Z} and ϕ_{W}, respectively. These functions are such that the nominal background expectations in Figure 1 (b) is attained when $\phi_{\mathrm{Z}} = \phi_{\mathrm{W}} = 0$ and the expectations represented by the dotted and dot-dashed lines are attained when $\phi_{\mathrm{Z}} = \phi_{\mathrm{W}} = -1$ and $\phi_{\mathrm{Z}} = \phi_{\mathrm{W}} = +1$, respectively. The functions are shown in Figure 2. The functions used here are designed to be differentiable for the full range in ϕ_{Z} and ϕ_{W} except at the boundary, which is included to avoid negative expected rates.

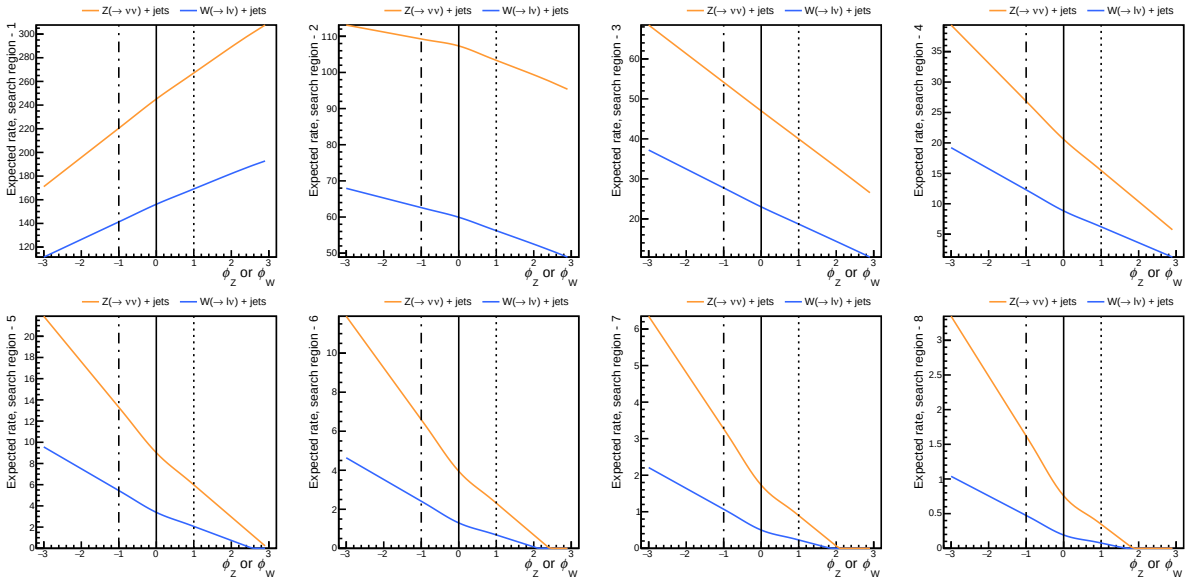


Figure 2: The functions $B_i^{\mathrm{Z}}(\phi_{\mathrm{Z}})$ and $B_i^{\mathrm{W}}(\phi_{\mathrm{W}})$, representing the expected rates of the $\mathrm{Z}(\to \nu\nu) + \text{jets}$ and $\mathrm{W}(\to l\nu) + \text{jets}$ backgrounds as a function of the nuisance parameters ϕ^{Z} or ϕ_{W} in each of the search regions. The solid, dotted, and dot–dashed vertical lines show the values of the nuisance parameters for which the expected background rate in each process attains the nominal, positive shifted and negative shifted rates in Figure 1 (b).

While the likelihood for this example is rather straight-forward, often the systematic uncertainties are derived through the use of one or more control regions which would be included as additional terms in the full likelihood. Moreover, BSM searches often consider many more sources of background with many different systematic uncertainties that yield more complicated variations for those backgrounds than those considered here. This motivates the use of the simplified likelihood.

The relevant information to construct the simplified likelihood for the toy search is summarised in Table 1. Furthermore, Figure 3 gives the covariance, $\mathbf{V}$, between the total rate of background contributions expected in each of the search regions. The non-zero off-diagonal values are a result of the fact that the estimate of the backgrounds are not independent in each region since the systematic effects cause correlated variations of the expected contributions of the two backgrounds across the regions.

Table 1: Observed number of events and total expected background and signal rates for each of the 8 search regions.

Search region (i)	Data (n_i)	Total background (b_i)	Signal (s_i)
1	1964	2006.4	47.0
2	877	836.4	29.4
3	354	350.0	21.1
4	182	147.1	14.3
5	82	62.0	9.4
6	36	26.2	7.1
7	15	11.1	4.7
8	11	4.7	4.3

The information provided in Table 1 and Figure 3 is all that is necessary to reconstruct the likelihood in Equation 4 and can be summarised as follows:

- The number of observed events in each search region, n_i.
- The background and signal expectations in each search region, b_i and s_i.
- The covariance between the number of background events in each search region, $\mathbf{V} = \left\{V_{ij}\right\}_{i,j=1}^{8}$.

The third item in the list is that which provides the simplified description of the true probability density function for the nuisance parameters. In the toy example, the true distribution is given by

$$\rho_T(\theta_1, \theta_2, ...) = \rho_T\left(B_1^Z(\phi_Z) + B_1^W(\phi_W) - b_1, B_2^Z(\phi_Z) + B_2^W(\phi_W) - b_2, ...\right), \tag{7}$$

where the identification $\theta_i = B_i^Z(\phi_Z) + B_i^W(\phi_W) - b_i$ is made. The Gaussian approximation for the distribution takes the same form for the constraint term in Equation 4, namely,

$$\rho_G(\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\boldsymbol{\theta}^T\mathbf{V}^{-1}\boldsymbol{\theta}\right). \tag{8}$$

Figure 4 shows a comparison between the projections of ρ_T, compared to ρ_G. In this toy example, the distribution ρ_T is evaluated in pseudo-experiments, by randomly varying the values of ϕ_Z and ϕ_W, assuming each to be distributed as a unit Gaussian. An alternative approach to determining ρ_T has been suggested using an analytic derivation [17], however, the use of such an approach is beyond the scope of this note. In this toy example, there is a good agreement between the true probability density function and the Gaussian approximation.

The simplified likelihood can be used to perform a number of statistical tests using both Bayesian or frequentist techniques. In the rest of this section, the statistical procedures used will follow the frequentist paradigm, employing the use of the profiled likelihood ratio defined as

$$q(\mu) = -2\ln\frac{\mathcal{L}_S(\mu, \hat{\hat{\boldsymbol{\theta}}}_\mu)}{\mathcal{L}_S(\hat{\mu}, \hat{\hat{\boldsymbol{\theta}}})}, \tag{9}$$

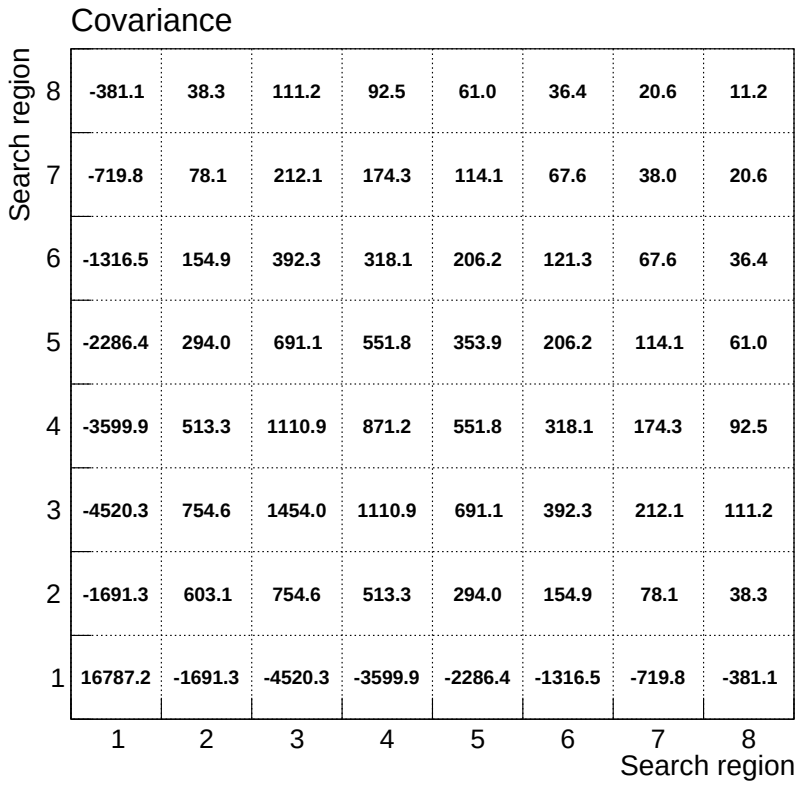


Figure 3: Covariance between the total rate of background contributions expected in each of the search regions.

where $\hat{\mu}$ and $\hat{\hat{\boldsymbol{\theta}}}$ are the values of the parameters μ and $\boldsymbol{\theta}$ respectively, which maximise the likelihood. The values $\hat{\boldsymbol{\theta}}_\mu$ are the values of $\boldsymbol{\theta}$ which maximise the likelihood for a fixed value of μ. The value of $\hat{\mu}$ is usually referred to as the "best-fit" value.

A common estimate of the uncertainty in μ is to determine the interval in μ such that $q(\mu) \leq 1$ and define the uncertainty as the difference between the end points of that interval and $\hat{\mu}$. Furthermore, the profiled likelihood ratio is a common test statistic for setting limits and quantifying excesses in BSM searches performed by the CMS collaboration, a full description of which can be found in Ref. [14]. The uncertainty in μ often provides a good indication of the sensitivity of a search to a given BSM signal.

Figure 5 shows the value of $q(\mu)$ as a function of μ. The values when $q(\mu)$ is defined using the likelihood of Equation 4 are shown and compared to the same definition but assuming no correlations between the background yields by setting $V_{ij} = 0$ for $i \neq j$. In this case, the systematic uncertainty in each region is assumed to be independent of the systematic uncertainty in any other region. The results substituting $\mathcal{L}_S \to \mathcal{L}$ in Equation 9, namely using the full likelihood of Equation 1, are also shown. The simplified likelihood shows good agreement with the full likelihood. For this example, ignoring the correlations results in a discrepancy in the estimate of $\hat{\mu}$. In addition, the width of the curve ignoring the correlation is larger than the other curves which will lead to an overestimation of the uncertainty on μ. In general, this agreement will depend on the relative importance of the off-diagonal terms in the covariance matrix and for smaller correlations, the agreement can be expected to improve.

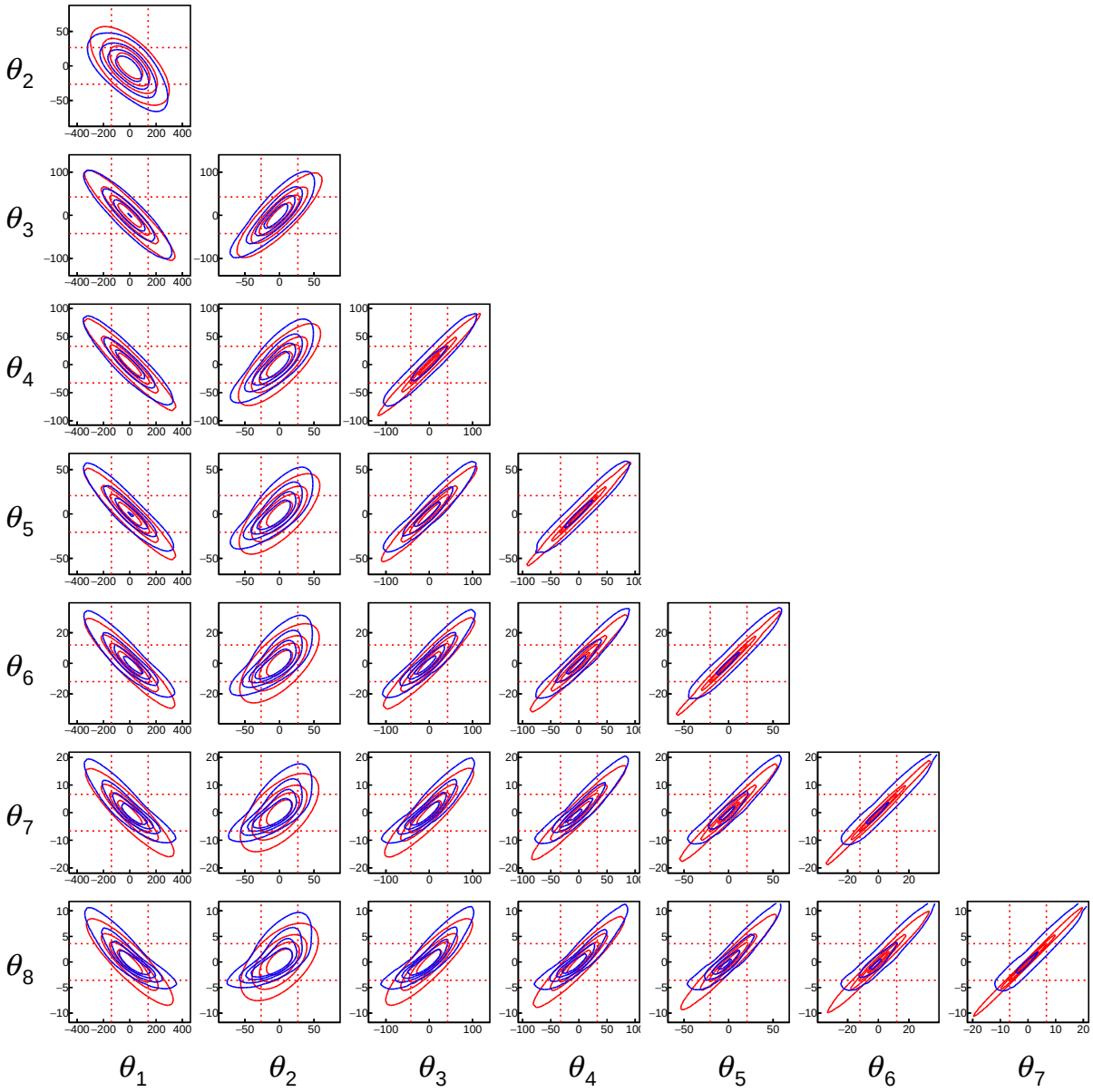


Figure 4: Contours of constant probability density for the true probability density function and the Gaussian approximation for the nuisance parameters in the toy search. The red dotted horizontal and vertical lines indicate the regions for which $|\theta_i| < \sqrt{V_{ii}}$, where θ_i is the nuisance parameter along the vertical and horizontal axes, respectively.

While typically the approximations necessary to construct the simplified likelihood are valid in a wide range of searches, it is important to note that there are scenarios in which they are expected to be invalid. In the toy example given, the expected background and data in the search regions with a low signal to background ratio, are in good agreement. The fitted values of the nuisance parameters, therefore will be small compared to square root of the variance of the background expectation, $|\hat{\theta}_{i,\mu}| < \sqrt{V_{ii}}$. This means the nuisance parameters do not need to be pulled too far from 0 in order to fit the data. The variation of these values as a function of μ are shown in Figure 6 (a).

The multivariate Gaussian is not expected to provide a good approximation of the true proba-

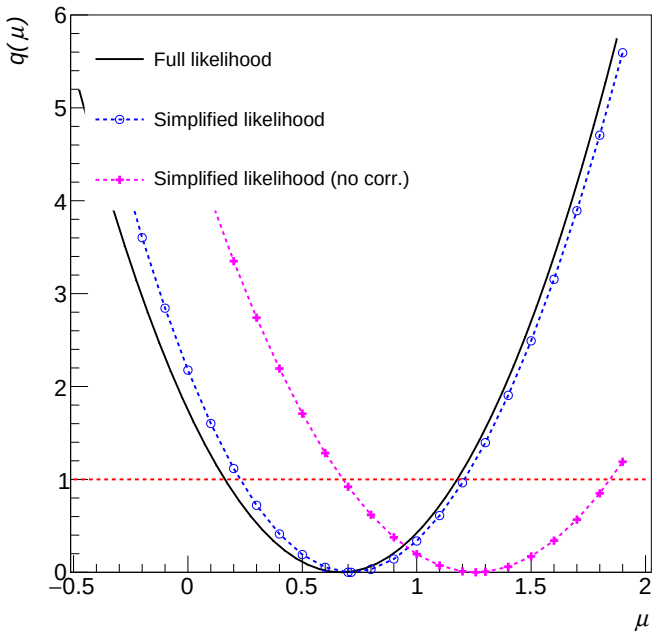


Figure 5: The value of $q(\mu)$ defined using the simplified likelihood using the full covariance matrix, assuming no correlations between the background yields and defined using the full likelihood.

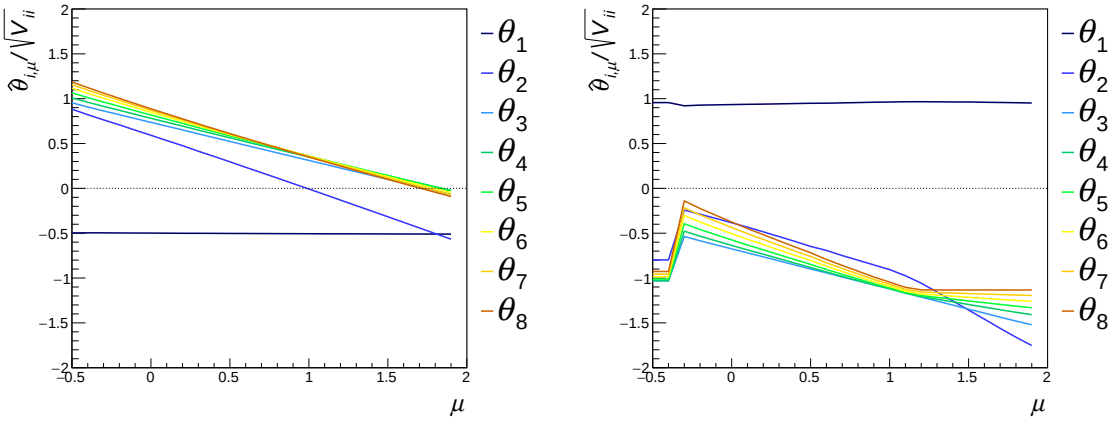
bility density function of the nuisance parameters for large values of the nuisance parameters. To illustrate this, a new set of observed event counts for each of the 8 search regions, n'_i, is produced for which the values of $\hat{\boldsymbol{\theta}}_\mu$ will be further from 0. Figure 7 shows a comparison of the event counts n_i and n'_i and the total expected backgrounds b_i. Unlike in the original data, the difference between b_1 and n'_1 is significant. Since the expected contribution of the signal is relatively small, the fit must account for this difference by moving the nuisance parameters away from 0. Figure 6 (b) shows the fitted values of the nuisance parameters, relative to the square root of the background variance, as a function of μ.

Figure 8 (a) shows the resulting values of $q(\mu)$ as a function of μ when replacing $n_i \to n'_i$ in Equation 4. The agreement between the full and simplified likelihoods is significantly poorer compared to Figure 5. This highlights the limitations of the multivariate Gaussian approximation as a model of the probability density function for the nuisance parameters when the values of the nuisance parameters are large.

The multivariate Gaussian approximation can also fail in the presence of systematic uncertainties which yield highly asymmetric variations in the background expectations. To illustrate this case, an additional uncertainty is included in the toy example which allows the total rate of the two background contributions to vary in a fully correlated way. This uncertainty is introduced via including an additional nuisance parameter η for which

$$B_i^{\mathrm{Z}}(\phi_{\mathrm{Z}}) \to (1.5)^{\eta} \cdot B_i^{\mathrm{Z}}(\phi_{\mathrm{Z}}), \qquad B_i^{\mathrm{W}}(\phi_{\mathrm{W}}) \to (1.5)^{\eta} \cdot B_i^{\mathrm{W}}(\phi_{\mathrm{W}}) \tag{10}$$

and multiplying the likelihood by the constraint term $e^{-\frac{1}{2}\eta^2}$.



(a) (b)

Figure 6: Values of $\frac{\hat{\theta}_{i,\mu}}{\sqrt{V_{ii}}}$ as a function of μ when using the original data n_i and the replacement data n'_i.

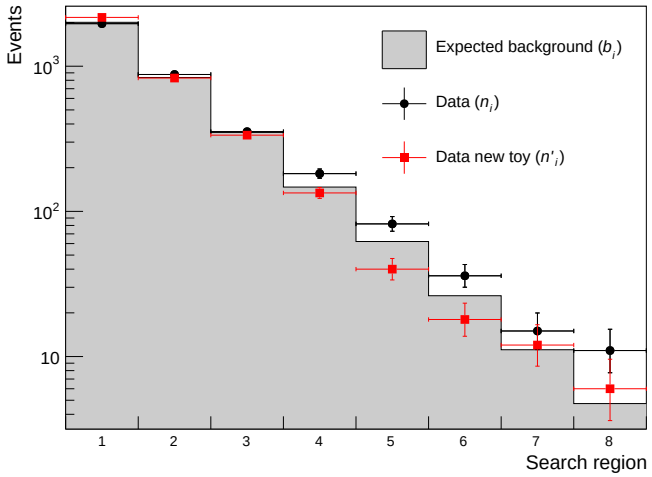
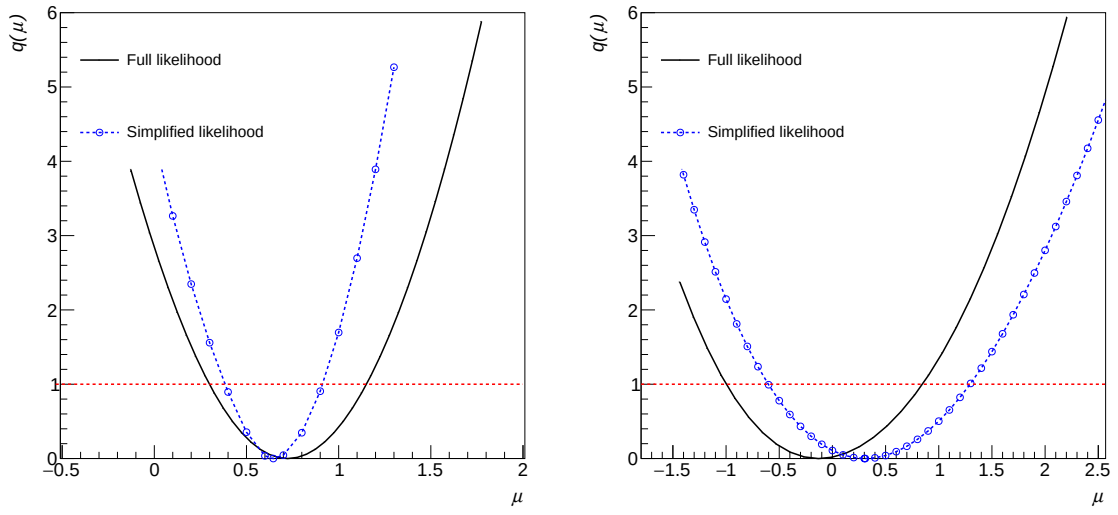


Figure 7: Comparison of the background expectations, b_i, and the original and replacement data, n_i and n'_i, in each of the 8 search regions.

Figure 9 shows comparisons between the true probability density function for $\boldsymbol{\theta}$, derived from pseudo-experiments in which the nuisance parameters of the full likelihood are varied, and the Gaussian approximation. The approximation is not appropriate in this case.

Figure 8 (b) shows the resulting value of $q(\mu)$ as a function of μ for the full and simplified likelihoods, after the introduction of the asymmetric systematic uncertainty. The use of the simplified likelihood in this case yields a significant difference in the estimate of $\hat{\mu}$, compared to the full likelihood. In general, the level of disagreement between the full and simplified likelihoods will depend on the validity of the Gaussian approximation.



(a) (b)

Figure 8: (a) The value of $q(\mu)$ defined using the simplified likelihood replacing $n_i \to n_i'$ and defined using the full likelihood. (b) The value of $q(\mu)$, when including a highly asymmetric systematic uncertainty on the total background expectations, defined using the simplified likelihood and defined using the full likelihood.

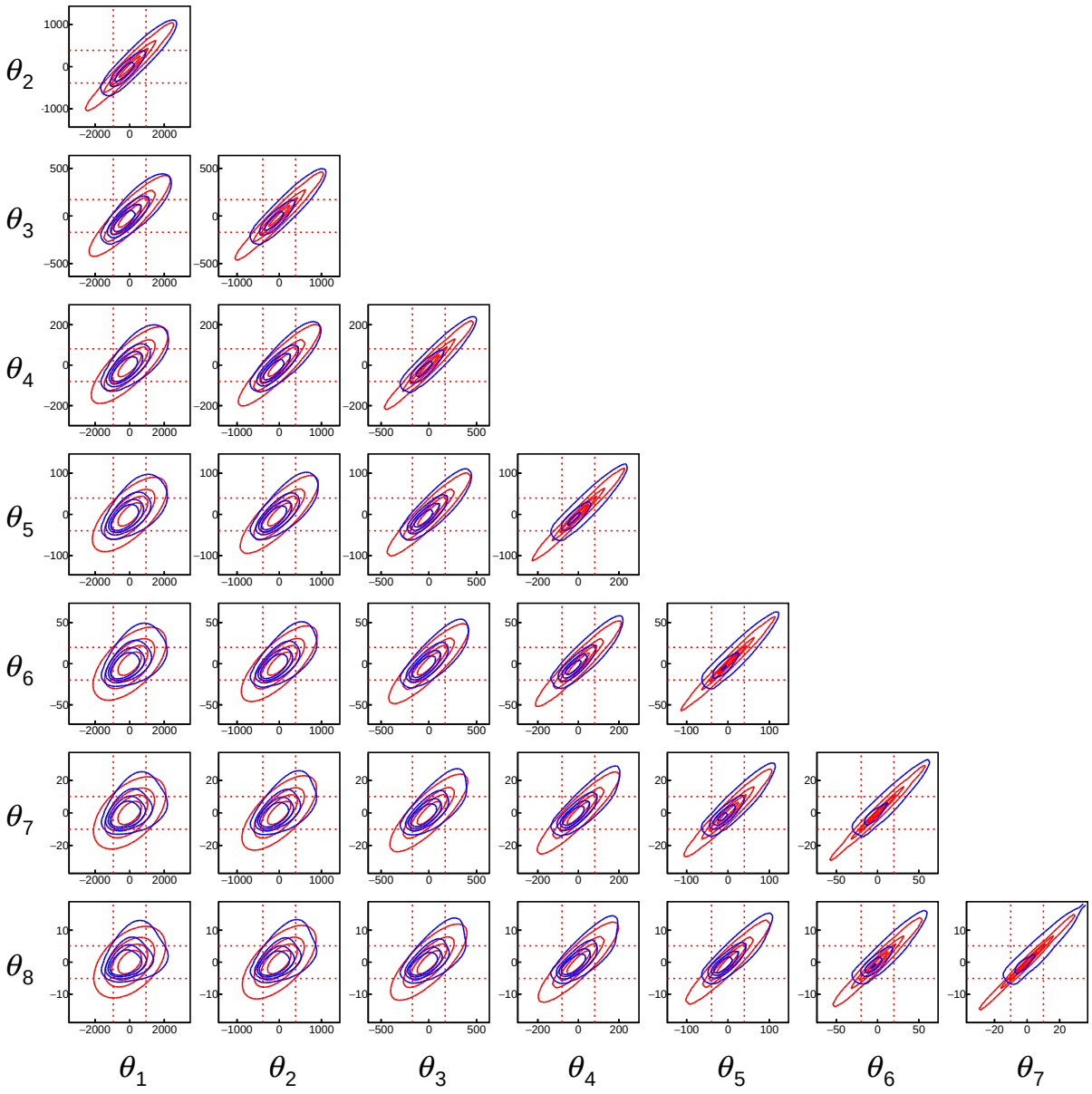


Figure 9: Contours of constant probability density for the true probability density function and the Gaussian approximation for the nuisance parameters in the toy search where an asymmetric background systematic is included. The red dotted horizontal and vertical lines indicate the regions for which $|\theta_i| < \sqrt{V_{ii}}$, where θ_i is the nuisance parameter along the vertical and horizontal axes, respectively.

3 Aggregated search regions

Inclusive searches for new physics typically use a fine categorisation to allow sensitivity to a wide range of models. This approach can make the use of the search for re-interpretation impractical. This section describes how the categorisation may be simplified without neglecting the correlation between the search regions by defining aggregated search regions.

Considering the simplified likelihood defined in Equation 4, the likelihood for aggregated search regions may be written by substituting the background predictions by

$$b_i + \theta_i \rightarrow b_I + \theta_I \equiv \sum_i b_i + \theta_I, \tag{11}$$

where the sum runs over all search regions being aggregated into search region I. Note that the number of aggregated search regions must be fewer than the total number of search regions and every search region must be contained in a single aggregated search region. Similarly, the signal expectations and observed number of events are substituted by

$$s_i \rightarrow s_I \equiv \sum_i s_i, \qquad n_i \rightarrow n_I \equiv \sum_i n_i. \tag{12}$$

The covariance matrix in the aggregated search regions is related to the full covariance matrix as

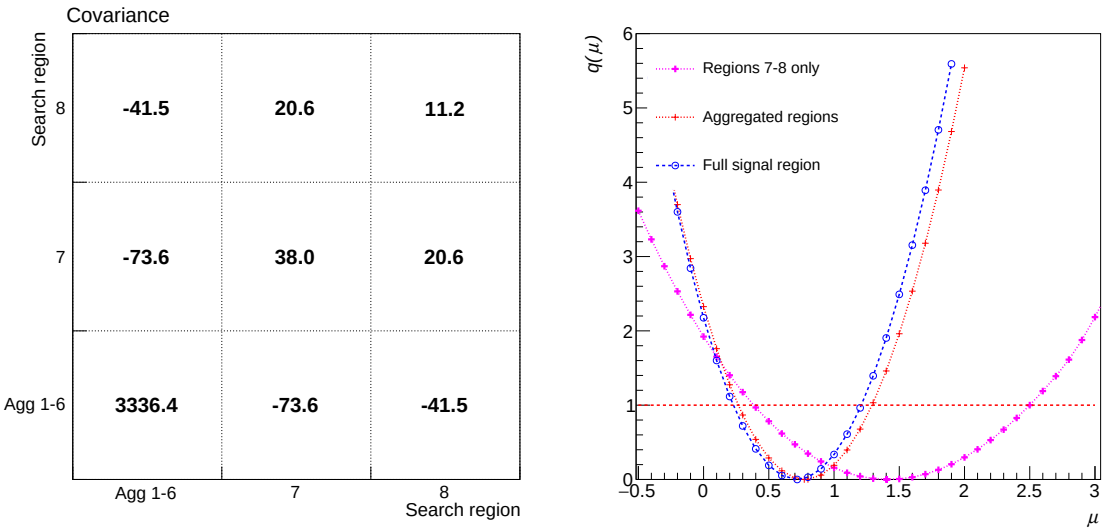
$$V_{IJ} = E[\sum_i \theta_i \times \sum_j \theta_j] = \sum_i \sum_j V_{ij}, \tag{13}$$

where the sum is over all search regions aggregated into search regions I, J and the linearity of the expectation value has been used. The use of the simplified likelihood for the aggregated search regions relies on the same conditions outlined in Section 2.

3.1 Example of the use of aggregated search regions

The same toy search described in Section 2.2 is considered for use with aggregated search regions. From Figure 1 it is clear that search regions 7 and 8 provide the dominant contribution to the sensitivity of the search as the ratio of the expected signal to the background is the largest in these two regions. One may therefore expect that search regions 1–6 can be neglected when deriving constraints on this particular signal model. However, as the background expectations in these search regions are correlated with those in search regions 7 and 8, neglecting search regions 1–6, while still valid for re-interpreting the search, will result in a loss of sensitivity. Alternatively, the search regions 1–6 may be aggregated using the covariance matrix shown in Figure 3 and Equation 13. The resulting covariance between the aggregated search regions is shown in Figure 10 (a).

Figure 10 (b) shows the value of $q(\mu)$ as a function of μ. The values when $q(\mu)$ is defined using the likelihood of Equation 4 using all 8 search regions, the aggregated search regions described above, and considering search regions 7 and 8 only are shown. There is good agreement between the curves using the aggregated search regions and using all 8 search regions.



(a) (b)

Figure 10: (a) Covariance between the total rate of background contributions expected in each of the aggregated search regions. (b) The value of $q(\mu)$ defined using the simplified likelihood considering all 8 search regions, the aggregated search regions and search regions 7 and 8 only.

Neglecting the search regions 1–6 causes a shift in the value of $\hat{\mu}$. In addition, the width of the likelihood curve when neglecting these search regions is considerably larger, implying a larger uncertainty estimate on μ and therefore a poorer sensitivity to the signal. This can be expected from the loss of constraint on the nuisance parameters when search regions 1–6 are neglected. Generically, the impact of removing search regions will depend on the size of the systematic uncertainties and the correlations between the search regions used and those which are neglected.

4 Validation of the simplified likelihood using CMS searches for BSM physics

This section details a validation of the use of simplified likelihoods in two example searches for BSM physics at CMS. Both of these searches use multiple criteria to categorise events and are designed to be sensitive to a range of BSM models. The two searches are typical of those for which the simplified likelihood approach is intended for re-interpretation of the results. These searches use the 12.9 fb^{-1} dataset of 13 TeV proton-proton collisions collected by the CMS detector at the LHC.

4.1 Monojet search

The simplified likelihood approach has been validated in a search for dark matter (DM) with a large missing transverse momentum and jets resulting from radiation of a gluon or a vector boson [18].

The search comprises two categories of events that are defined by the nature of the jets in the event. The first category, labelled mono-V, selects events which contain a "fat-jet" with a mass and substructure consistent with that of a hadronically decaying W or Z boson. The second category, the monojet category, includes events failing the selection of the first but containing at least one high transverse momentum jet. In both categories, the magnitude of the missing transverse momentum $E_{\text{T}}^{\text{miss}}$ is used to further discriminate the signal from the background. Each category is divided into several search regions defined as intervals in $E_{\text{T}}^{\text{miss}}$, 7 in the mono-V category and 22 in the monojet category making a total of 29 search regions. The results of the search are interpreted in terms of simplified models for DM production [19–21], mediated via vector, axial vector, scalar or pseudoscalar interactions.

4.1.1 Validation of the simplified likelihood

The dominant backgrounds in this search are derived by defining several control regions in the data and simultaneously fitting for background contribution in the control and search regions. The full procedure is described in detail in Section 4 of Ref. [18]. The smaller backgrounds are estimated from simulation, tuned using data.

The simplified likelihood in this example is constructed using the number of observed events and total expected background in each of the $E_{\text{T}}^{\text{miss}}$ bins in the mono-V and monojet search regions. The covariance between the expected background is estimated by propagating the information from the control regions to the search regions using pseudo-experiments. In some cases, the correlations between the mono-V and monojet categories are substantial and are therefore included in the constriction of the simplified likelihood. The signal expectations in the bins of $E_{\text{T}}^{\text{miss}}$ are taken directly from the yields reported in Ref [18].

Figure 11 shows a comparison of $q(\mu)$ as a function of μ for an example signal point in the simplified model parameter space. The results are compared to the ones obtained using the full likelihood and using the simplified likelihood, ignoring correlations between the background expectations. The results using an Asimov dataset [22] assuming a signal strength of $\mu = 1$ are also shown as these are used to calculate limits using the asymptotic approximations detailed in Ref [22].

Figure 12 shows a comparison of the 95% confidence level upper limit on the signal strength, $\mu_{\text{up}}^{95\%}$, as a function of the mass of a scalar or pseudoscalar mediator, m_{MED}, for a fixed value of the coupling parameters and a fixed dark matter mass, $m_{\text{DM}} = 10\,\text{GeV}$. The results are

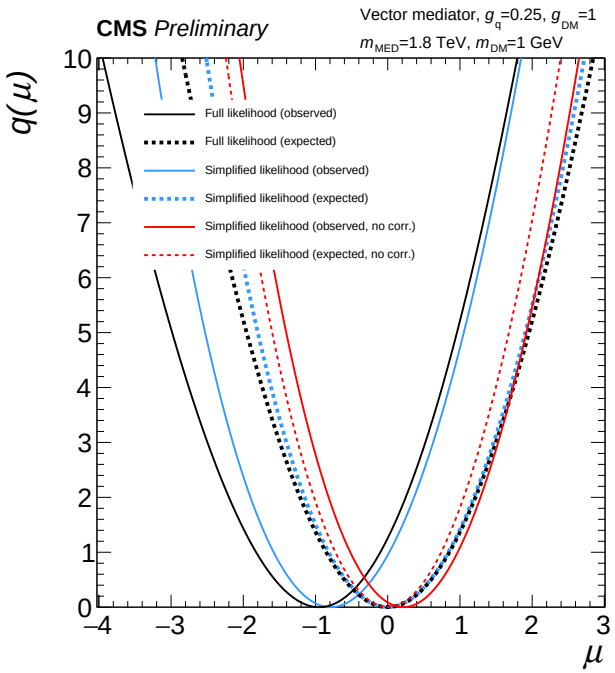
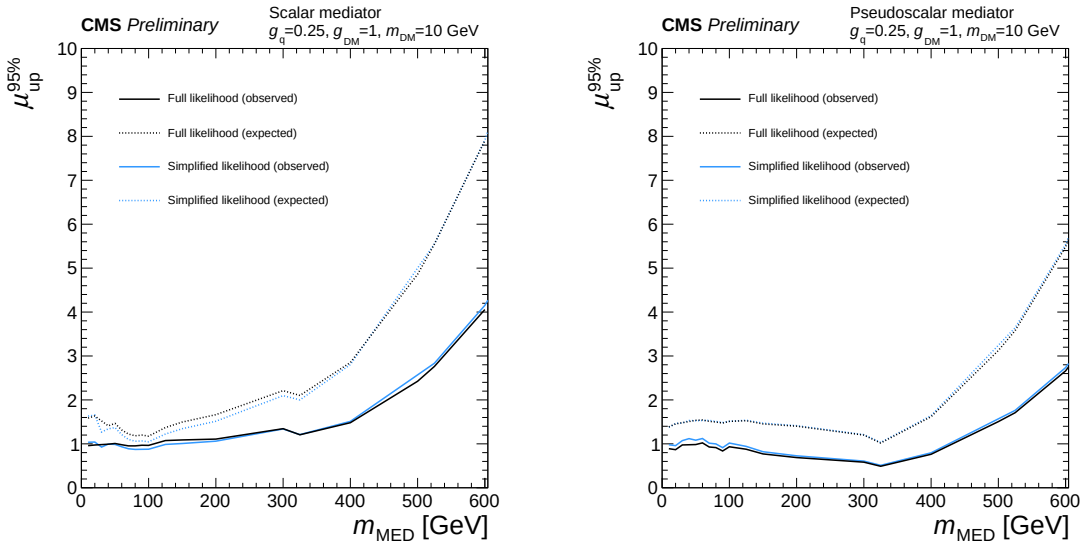


Figure 11: Scan of q_μ as a function of μ using the full and simplified likelihood and simplified likelihood ignoring background correlations for the Monojet search assuming mediator and DM masses of 1.8 TeV and 1 GeV respectively. Coupling values of $g_\mathrm{q} = 0.25$ and $g_\mathrm{DM} = 1$ are assumed for the simplified model for DM production. The results are shown for the observed data and using an Asimov dataset assuming $\mu = 1$.

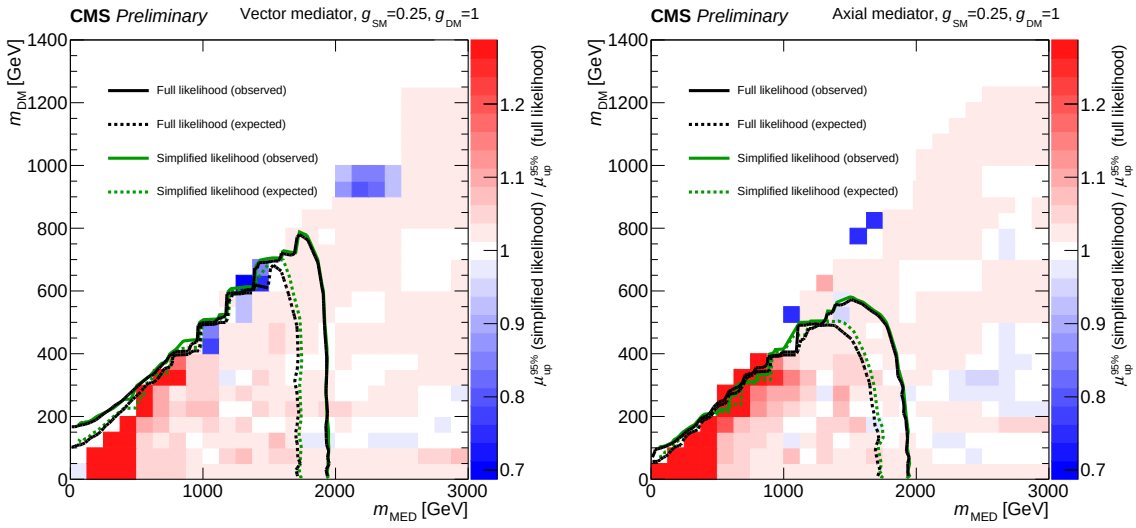
compared using the full likelihood and the simplified likelihood in the definition of the test statistic for both expected and observed upper limits.

Figure 13 shows a comparison of the limits set in the m_MED–m_DM plane for a vector or axial vector mediator with fixed coupling parameters using the full and simplified likelihoods. The ratios of the observed values of $\mu_{up}^{95\%}$ between the two likelihood constructions is shown in the colour scale. The contours bound the regions for which the observed and expected values of $\mu_{up}^{95\%}$ are less than 1 in the two cases. This comparison is performed on a grid of points which is less granular than that used for the results in Ref. [18] such that the contours provided therein do not precisely compare to those in Figure 13. Nevertheless, in general a good agreement is found between the full and simplified likelihoods for this set of signal points. The disagreement in the region where $2m_\mathrm{DM} \approx m_\mathrm{MED}$ results from the use of a sparse grid in a region over which the cross section, and hence the limit, in the simplified models varies rapidly. This rapid change in the limits exaggerates the real discrepancy between the simplified and full likelihoods.



(a) (b)

Figure 12: Expected and observed $\mu^{95\%}_{up}$ as a function of m_{MED}, assuming a DM mass of 10 GeV, for a (a) scalar or (b) pseudoscalar mediator for the Monojet search. Coupling values of $g_q = 0.25$ and $g_{DM} = 1$ are assumed for the simplified model for DM production. The results are compared between the limits calculated using the full and the simplified likelihoods.



(a) (b)

Figure 13: Expected and observed exclusion contours defined as the boundary of the region where $\mu^{95\%}_{up} < 1$ in the m_{MED}–m_{DM} plane for a (a) vector or (b) axial vector mediator for the Monojet search. Coupling values of $g_{SM} = 0.25$ and $g_{DM} = 1$ are assumed for the simplified model for DM production. The results are compared between the limits calculated using the full and the simplified likelihoods. The colour scale shows the ratio of $\mu^{95\%}_{up}$ calculated using the simplified likelihood to the value using the full likelihood.

4.2 α_T search

The α_T search is an inclusive search for supersymmetry (SUSY) which uses a fine categorisation in n_{jet} (number of jets), n_b (number of jets identified as b-quark jets), H_T (scalar sum of the transverse momentum of the jets), and H_T^{miss} (magnitude of the vector sum of the transverse momentum of the jets) to provide sensitivity to a wide range of new physics models. The results are described fully in Ref. [13]. In this section the use of the simplified likelihood method is illustrated and validated for an example SUSY simplified signal model, T2bb, for which a pair of bottom squarks $\tilde{b}$ are produced and decay to a bottom quark and lightest SUSY particle $\tilde{\chi}^0$ ($pp \rightarrow \tilde{b}\tilde{b}^* \rightarrow b\tilde{\chi}^0 \bar{b}\tilde{\chi}^0$) [23]. Similar validations have been carried out for the simplified models T2tt and T1bbbb (defined in Ref. [13]).

4.2.1 Categorisation

The simplified likelihood is defined from the aggregated regions used for the α_T search. The categorisation is summarised in Table 2. These aggregated regions are defined to be disjoint, contiguous and cover the full search region phase space.

Table 2: To define the aggregate regions for the α_T search the H_T dimension is merged to $\geq$ 200 GeV and n_b to two categories of $n_b < 2$ and ≥ 2. The merged n_{jet} categories are summarised in this table. Each category is further binned using eight H_T^{miss} bins with lower bounds from $100 - 800$ GeV.

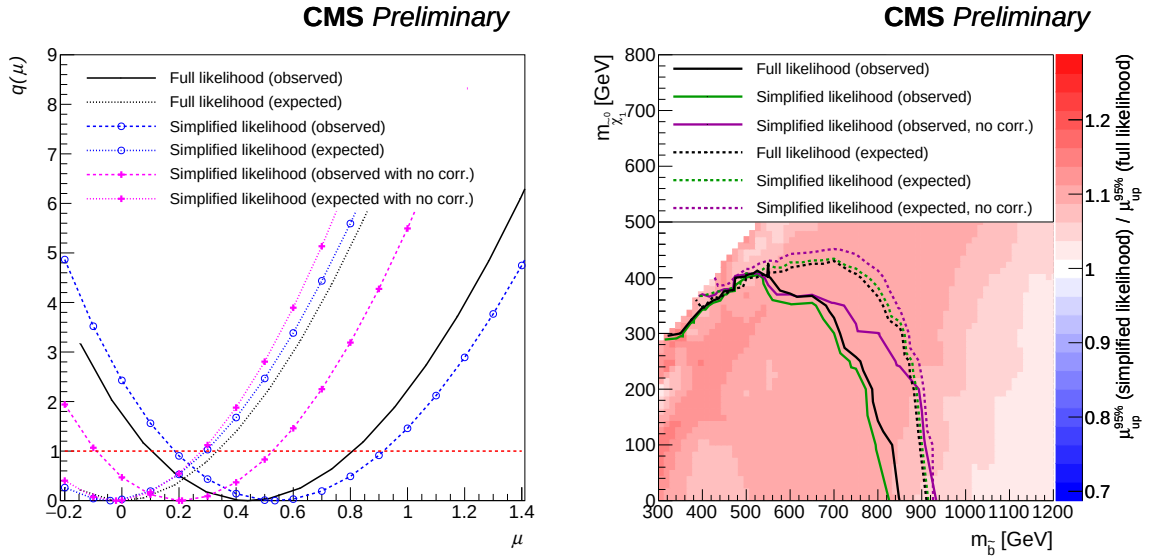
n_{jet} topology	Merged jet categories		
	Monojet	**Asymmetric**	**Symmetric**
Monojet-like	1	2	2
Asymmetric high n_{jet}	-	3, 4, ≥ 5	-
Mid n_{jet}	-	-	3, 4
High n_{jet}	-	-	≥ 5

4.2.2 Validation of the simplified likelihood

The predictions in the search regions are derived, as described in Section 5 of Ref. [13], using a simultaneous fit to several control regions and these predictions are compared with the observations. As for the Monojet search the covariances between the search regions are derived using pseudo-experiments.

The covariance and predictions in the aggregated regions are used to define and validate the simplified likelihood. The signal expectations are taken directly from the yields reported in Ref [13]. Figure 14 (a) shows the value of $q(\mu)$ as a function of μ for the T2bb model with $m_{\tilde{b}} =$ 800 GeV and $m_{\tilde{\chi}^0} = 200$ GeV. The values of $q(\mu)$ are compared to those derived using the full likelihood and using the simplified likelihood, ignoring correlations between the background expectations. The results using an Asimov dataset, assuming a signal strength of $\mu = 1$, are also shown as these are used to calculate limits using the asymptotic approximations detailed in Ref. [22].

In Figure 14 (b) the ratio between the observed $\mu_{up}^{95\%}$ for the simplified and the full likelihood is shown for T2bb. The expected and observed contours of $\mu = 1$ excluded at 95% for the full likelihood, the simplified likelihood and the simplified likelihood where correlations are neglected are overlaid. The results using the full and simplified likelihoods are seen to agree well when the correlations between bins are considered.



(a) (b)

Figure 14: (a) The value of $q(\mu)$ for the search defined using the simplified likelihood using the full covariance matrix, assuming no correlations between the background yields and defined using the full likelihood. (b) Expected and observed exclusion contours defined as the boundary of the region where $\mu_{\text{up}}^{95\%} < 1$ in the T2bb plane for the α_T search. The results are compared between the limits calculated using the full and the simplified likelihoods and the simplified likelihood assuming no correlations between the background yields. The colour scale shows the ratio of $\mu_{\text{up}}^{95\%}$ calculated using the simplified likelihood to the value using the full likelihood.

5 Summary

Searches for new physics by the CMS collaboration are performed using a wide variety of strategies that exploit a number of different final states with wide ranging kinematic properties. Re-interpreting these searches requires approximating the background model and associated systematic uncertainties for the search. This can be achieved by defining a simplified likelihood combining one or more search regions in which the systematic uncertainties on the backgrounds are modelled as Gaussian constrained nuisance parameters. The definition of the simplified likelihood requires the observed data and the total expected background in each search region, and the covariance of the total background between each of the search regions. In addition, the number of search regions being considered can be reduced, where necessary, through the use of aggregated search regions. A toy example search has been used to demonstrate and validate the use of the simplified likelihood and aggregated search regions. Some examples of scenarios where the simplified likelihood does not provide a good approximation of the full likelihood have been highlighted. The use of the simplified likelihood approach has been validated in two CMS searches for BSM physics and a good agreement between the full and simplified likelihoods was found. Using only the reduced set of information required by the simplified likelihood approach allows for re-interpretation of CMS searches under different BSM physics models, where the conditions stipulated for the method are met.

References

[1] CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST* **3** (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[2] CMS Collaboration, "Search for high-mass diphoton resonances in proton-proton collisions at 13 TeV and combination with 8 TeV search", (2016). arXiv:1609.02507. Accepted by *Phys. Lett. B*.

[3] CMS Collaboration, "Search for narrow resonances in dilepton mass spectra in proton-proton collisions at $\sqrt{s} = 13$ TeV and combination with 8 TeV data", Technical Report arXiv:1609.05391. CERN-EP-2016-209. CMS-EXO-15-005, CERN, Geneva, Sep, 2016. Submitted to *Phys. Lett. B*.

[4] CMS Collaboration, "Search for top squark pair production in compressed-mass-spectrum scenarios in proton-proton collisions at $\sqrt{s} = 8$ TeV using the α_{T} variable", Technical Report arXiv:1605.08993. CMS-SUS-14-006. CERN-EP-2016-103, CERN, Geneva, May, 2016. Submitted to *Phys. Lett. B*.

[5] CMS Collaboration, "Search for dark matter in proton-proton collisions at 8 TeV with missing transverse momentum and vector boson tagged jets", *JHEP* **12** (2016) 083, doi:10.1007/JHEP12(2016)083, arXiv:1607.05764.

[6] CMS Collaboration, "Search for W' decaying to tau lepton and neutrino in proton-proton collisions at $\sqrt{(}s) = 8$ TeV", *Phys. Lett. B* **B755** (2016) 196, doi:10.1016/j.physletb.2016.02.002, arXiv:1508.04308.

[7] K. J. de Vries et al., "The pMSSM10 after LHC Run 1", *Eur. Phys. J.* **C75** (2015), no. 9, 422, doi:10.1140/epjc/s10052-015-3599-y, arXiv:1504.03260.

[8] T. Sjostrand, S. Mrenna, and P. Z. Skands, "A Brief Introduction to PYTHIA 8.1", *Comput. Phys. Commun.* **178** (2008) 852–867, doi:10.1016/j.cpc.2008.01.036, arXiv:0710.3820.

[9] DELPHES 3 Collaboration, "DELPHES 3, A modular framework for fast simulation of a generic collider experiment", *JHEP* **02** (2014) 057, doi:10.1007/JHEP02(2014)057, arXiv:1307.6346.

[10] M. Drees et al., "CheckMATE: Confronting your Favourite New Physics Model with LHC Data", *Comput. Phys. Commun.* **187** (2015) 227–265, doi:10.1016/j.cpc.2014.10.018, arXiv:1312.2591.

[11] M. Papucci, K. Sakurai, A. Weiler, and L. Zeune, "Fastlim: a fast LHC limit calculator", *Eur. Phys. J.* **C74** (2014), no. 11, 3163, doi:10.1140/epjc/s10052-014-3163-1, arXiv:1402.0492.

[12] CMS Collaboration, "Search for new physics in the all-hadronic final state with the MT2 variable", CMS Physics Analysis Summary CMS-PAS-SUS-16-015, 2016.

[13] CMS Collaboration, "An inclusive search for new phenomena in final states with one or more jets and missing transverse momentum at 13 TeV with the AlphaT variable", CMS Physics Analysis Summary CMS-PAS-SUS-16-016, 2016.

[14] CMS Collaboration, "Combined results of searches for the standard model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV", *Phys. Lett. B* **710** (2012) 26, doi:10.1016/j.physletb.2012.02.064, arXiv:1202.1488.

[15] The ATLAS , The CMS , The LHC Higgs Combination Group Collaboration, "Procedure for the LHC Higgs boson search combination in Summer 2011", Technical Report CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11, CERN, 2011.

[16] W. Verkerke and D. Kirkby, "The RooFit toolkit for data modeling", *ArXiv Physics e-prints* (June, 2003) arXiv:physics/0306116.

[17] S. Fichet, "Taming systematic uncertainties at the LHC with the central limit theorem", *Nucl. Phys. B* **911** (2016) 623, doi:10.1016/j.nuclphysb.2016.08.029, arXiv:1603.03061.

[18] CMS Collaboration, "Search for dark matter in final states with an energetic jet, or a hadronically decaying W or Z boson using 12.9 fb^{-1} of data at $\sqrt{s} = 13$ TeV", CMS Physics Analysis Summary CMS-PAS-EXO-16-037, 2016.

[19] G. Busoni, A. D. Simone, E. Morgante, and A. Riotto, "On the validity of the effective field theory for dark matter searches at the LHC", *Phys. Lett. B* **728C** (2014) 412., doi:10.1016/j.physletb.2013.11.069, arXiv:1307.2253.

[20] O. Buchmueller, M. J. Dolan, and C. McCabe, "Beyond effective field theory for dark matter searches at the LHC", *JHEP* **01** (2014) 025, doi:10.1007/JHEP01(2014)025, arXiv:1308.6799.

[21] O. Buchmueller, M. J. Dolan, S. A. Malik, and C. McCabe, "Characterising dark matter searches at colliders and direct detection experiments: Vector mediators", *JHEP* **1501** (2015) 037, doi:10.1007/JHEP01(2015)037, arXiv:1407.8257.

[22] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics", *Eur. Phys. J. C* **71** (2011) 1554, doi:10.1140/epjc/s10052-011-1554-0, arXiv:1007.1727. [Erratum: *Eur. Phys. J. C* **73** (2013) 2501].

[23] D. Alves et al., "Simplified models for LHC new physics searches", *J. Phys. G* **39** (2012) 105005, doi:10.1088/0954-3899/39/10/105005, arXiv:1105.2838.