Analysis Note Review for Background Estimation Strategy Comparison.

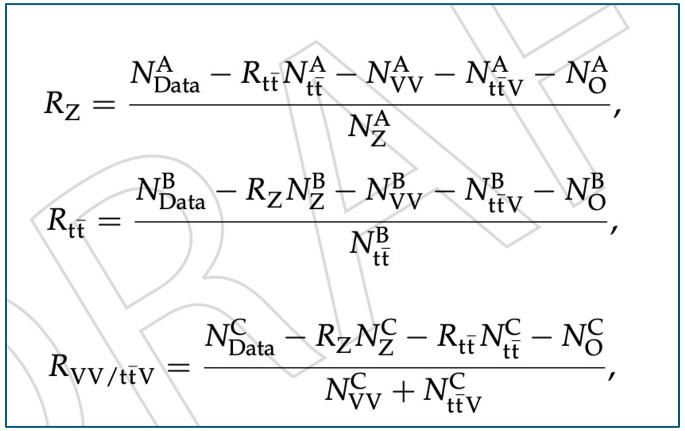
Each of these are at least in the pre-approval stage.

Exotica: EXO-20-011, EXO-21-019

SUSY: SUS-23-004, SUS-23-005, SUS-23-013, SUS-23-015, SUS-23-018

EXO-21-019 (AN-2019/233)

- Leptoquark search in final states with $2\mu + 1\mathrm{b} + 1\mathrm{j}$.
- Published June 2024.
- They use a method similar to ours (:

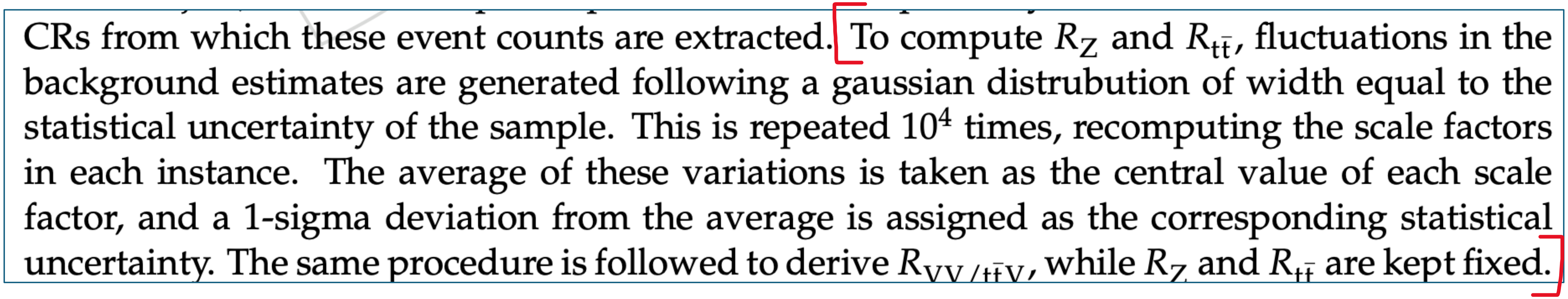


$$R_{\mathrm{Z}} = \frac{N^{\mathrm{A}}_{\mathrm{Data}} - R_{\mathrm{t\bar{t}}} N^{\mathrm{A}}_{\mathrm{t\bar{t}}} - N^{\mathrm{A}}_{\mathrm{VV}} - N^{\mathrm{A}}_{\mathrm{t\bar{t}V}} - N^{\mathrm{A}}_{\mathrm{O}}}{N^{\mathrm{A}}_{\mathrm{Z}}},$$

$$R_{\mathrm{t\bar{t}}} = \frac{N^{\mathrm{B}}_{\mathrm{Data}} - R_{\mathrm{Z}} N^{\mathrm{B}}_{\mathrm{Z}} - N^{\mathrm{B}}_{\mathrm{VV}} - N^{\mathrm{B}}_{\mathrm{t\bar{t}V}} - N^{\mathrm{B}}_{\mathrm{O}}}{N^{\mathrm{B}}_{\mathrm{t\bar{t}}}},$$

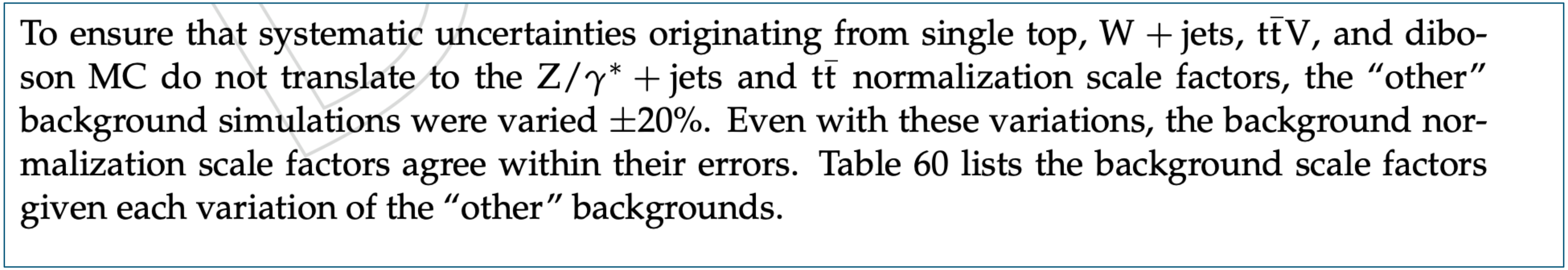
$$R_{\mathrm{VV/t\bar{t}V}} = \frac{N^{\mathrm{C}}_{\mathrm{Data}} - R_{\mathrm{Z}} N^{\mathrm{C}}_{\mathrm{Z}} - R_{\mathrm{t\bar{t}}} N^{\mathrm{C}}_{\mathrm{t\bar{t}}} - N^{\mathrm{C}}_{\mathrm{O}}}{N^{\mathrm{C}}_{\mathrm{VV}} + N^{\mathrm{C}}_{\mathrm{t\bar{t}V}}},$$

- However, they take extra steps to account for variations of non-targeted backgrounds…



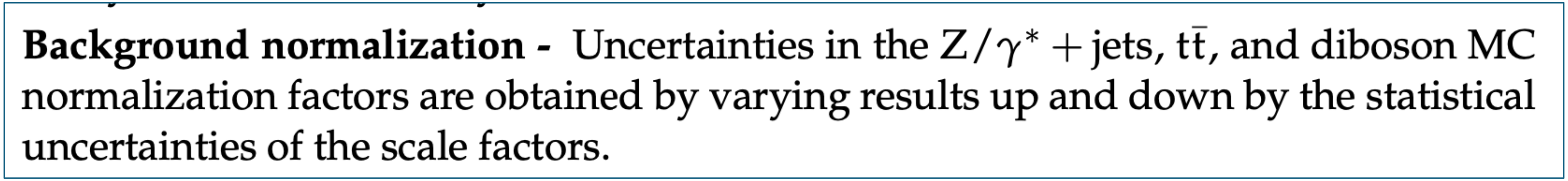
CRs from which these event counts are extracted. To compute R_Z and $R_{t\bar{t}}$, fluctuations in the background estimates are generated following a gaussian distrubution of width equal to the statistical uncertainty of the sample. This is repeated 10^4 times, recomputing the scale factors in each instance. The average of these variations is taken as the central value of each scale factor, and a 1-sigma deviation from the average is assigned as the corresponding statistical uncertainty. The same procedure is followed to derive $R_{VV/t\bar{t}V}$, while R_Z and $R_{t\bar{t}}$ are kept fixed.

- Their SFs "R" range from 0.95 to 2.8 depending on jet multiplicity.
- They also take extra steps to vary non-targeted backgrounds by up to up to 20% and confirm their SFs are not significantly changed (Appendix B).



To ensure that systematic uncertainties originating from single top, W + jets, $t\bar{t}V$, and diboson MC do not translate to the Z/γ^* + jets and $t\bar{t}$ normalization scale factors, the "other" background simulations were varied ±20%. Even with these variations, the background normalization scale factors agree within their errors. Table 60 lists the background scale factors given each variation of the "other" backgrounds.

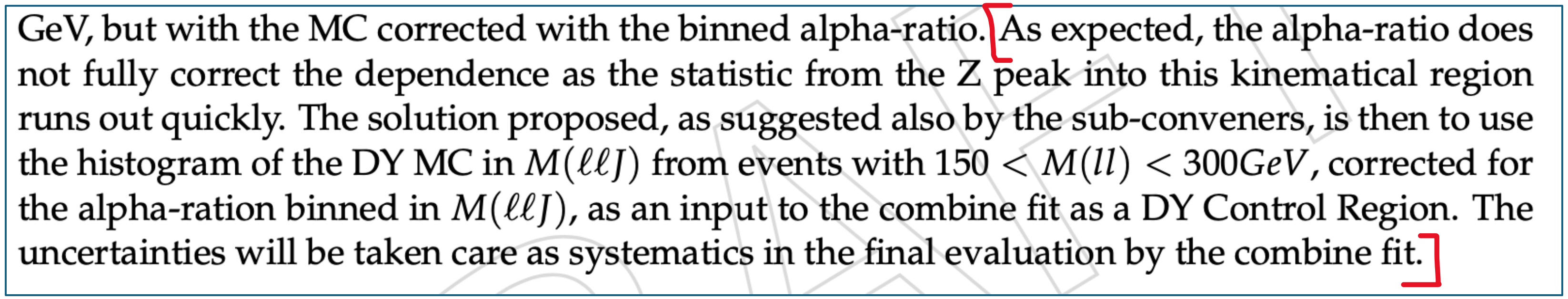
- They account for uncertainties related to this SF method in a similar fashion to our analysis:



Background normalization - Uncertainties in the $Z/\gamma^* +$ jets, $t\bar{t}$, and diboson MC normalization factors are obtained by varying results up and down by the statistical uncertainties of the scale factors.

EXO-20-011 (AN-2018-111)

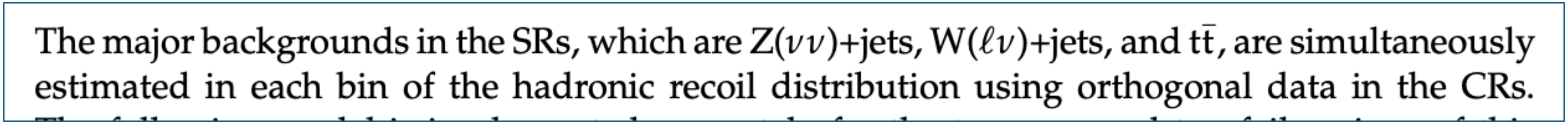
- Search for heavy composite Majorana neutrino in $eeqq$ and $\mu\mu qq$ final states.
- They initially use a similar BG estimation strategy similar to ours for their dominant BGs (sec. 5.1.2), but then use a simultaneous fit per convener recommendation (sec.5.1.3).



GeV, but with the MC corrected with the binned alpha-ratio. As expected, the alpha-ratio does not fully correct the dependence as the statistic from the Z peak into this kinematical region runs out quickly. The solution proposed, as suggested also by the sub-conveners, is then to use the histogram of the DY MC in $M(\ell\ell J)$ from events with $150 < M(ll) < 300 GeV$, corrected for the alpha-ration binned in $M(\ell\ell J)$, as an input to the combine fit as a DY Control Region. The uncertainties will be taken care as systematics in the final evaluation by the combine fit.

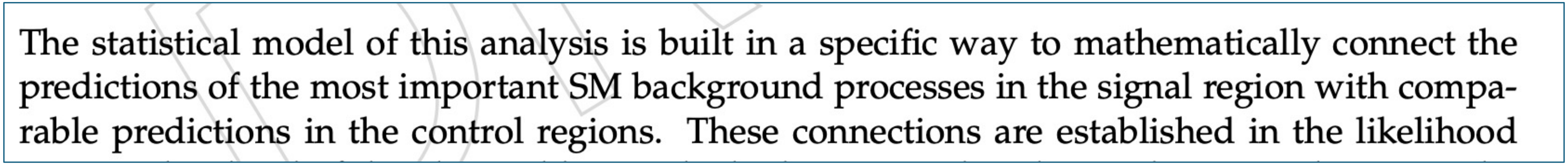
SUS-23-004 (AN-2020/061)

- SUSY search for mono-top jet in final state.
- It seems they use a SF method related to correcting efficiencies only related to top tagging because of their us of AK15 jets (See sections 6.6.1 and 6.6.3. for a description of their process)
- However, they ultimately use a simultaneous fit to estimate BG in SR (in PAS):



The major backgrounds in the SRs, which are $Z(\nu\nu)$+jets, $W(\ell\nu)$+jets, and $t\bar{t}$, are simultaneously estimated in each bin of the hadronic recoil distribution using orthogonal data in the CRs.

- And in the AN section 10.3 on statistical methods:



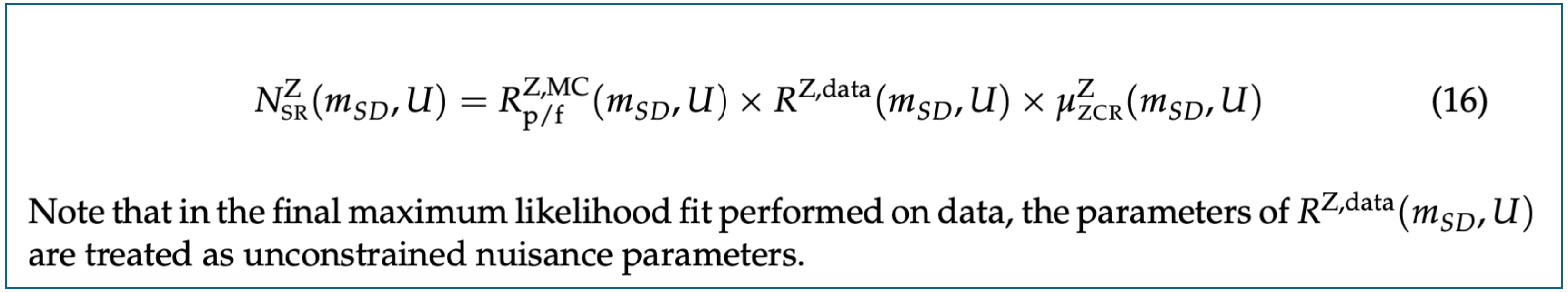
The statistical model of this analysis is built in a specific way to mathematically connect the predictions of the most important SM background processes in the signal region with comparable predictions in the control regions. These connections are established in the likelihood

SUS-23-005 (AN-2020/018)

- Search for light pseudoscalar Higgs in $\mu\mu\tau\tau$ final states.
- They use a SF method to estimate their BGs in their $\tau_h ID$ SF derivation because they are at τ_h with $p_T = [10,20] GeV$ (Sec. 5).
- They use the exact prescription used by the Tau POG.
- They ultimately use purely data-driven methods for their actual signal region BG estimation. (Sec. 8)

SUS-23-013 (AN-2018/026)

- Search DM in association with a b-quark pair.
- They’re BG estimation strategy involves deriving transfer factors in MC and data from CRs (Sec.7).
- A complex treatment of CR data and Bernstein polynomials are used to extract data TFs, which are treated as unconstrained nuisance parameters in the final SR fit.

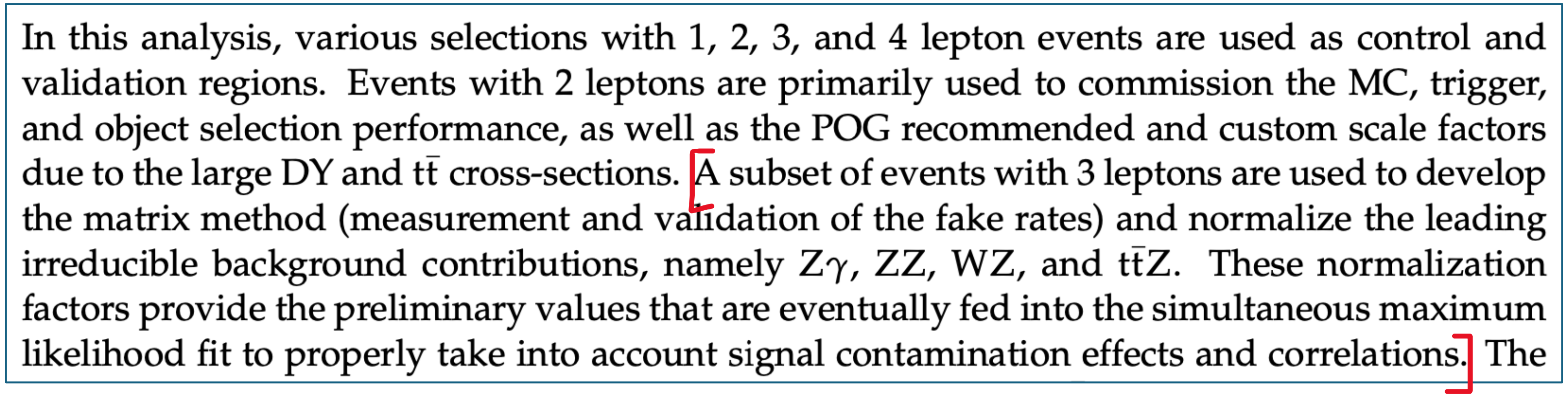


$$N^{\mathrm{Z}}_{\mathrm{SR}}(m_{SD}, U) = R^{\mathrm{Z,MC}}_{\mathrm{p/f}}(m_{SD}, U) \times R^{\mathrm{Z,data}}(m_{SD}, U) \times \mu^{\mathrm{Z}}_{\mathrm{ZCR}}(m_{SD}, U) \tag{16}$$

Note that in the final maximum likelihood fit performed on data, the parameters of $R^{\mathrm{Z,data}}(m_{SD}, U)$ are treated as unconstrained nuisance parameters.

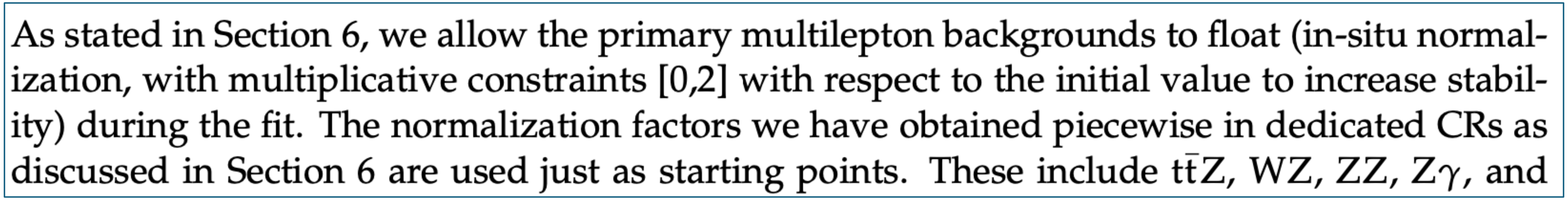
SUS-23-015 (AN-2022/173)

- SUSY search in trilepton + jets final states.
- They use 3-lepton CRs to estimate BGs. They use a data-driven approach to get “normalization factors,” and these are eventually used in a simulataneous fit (Sec. 6).



In this analysis, various selections with 1, 2, 3, and 4 lepton events are used as control and validation regions. Events with 2 leptons are primarily used to commission the MC, trigger, and object selection performance, as well as the POG recommended and custom scale factors due to the large DY and $t\bar{t}$ cross-sections. [A subset of events with 3 leptons are used to develop the matrix method (measurement and validation of the fake rates) and normalize the leading irreducible background contributions, namely $Z\gamma$, ZZ, WZ, and $t\bar{t}Z$. These normalization factors provide the preliminary values that are eventually fed into the simultaneous maximum likelihood fit to properly take into account signal contamination effects and correlations.] The

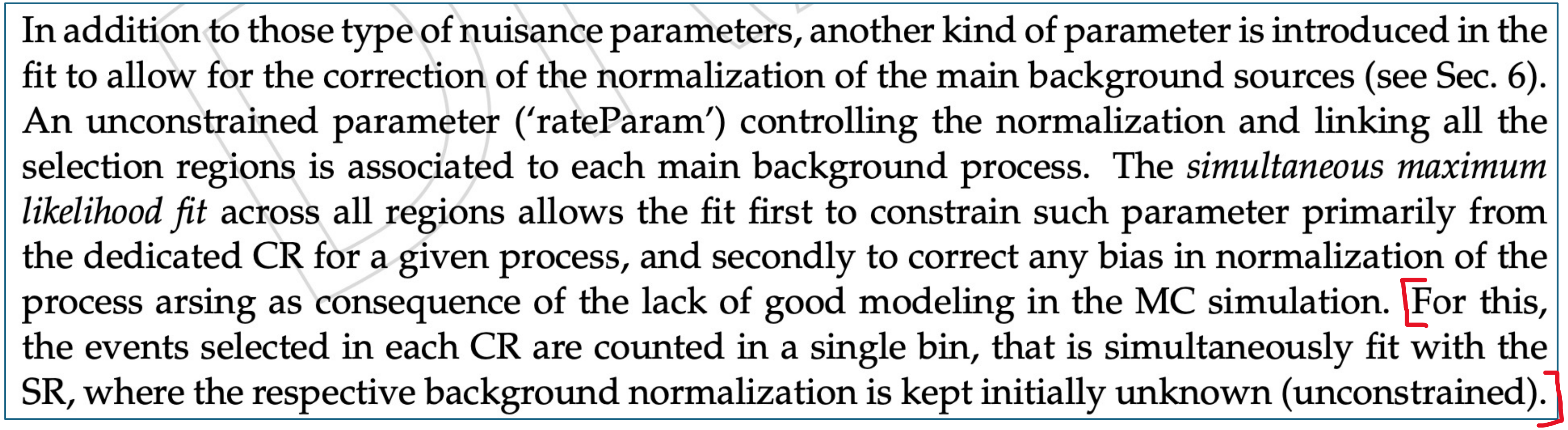
- And in Sec. 9 regarding their fitting techniques:



As stated in Section 6, we allow the primary multilepton backgrounds to float (in-situ normalization, with multiplicative constraints [0,2] with respect to the initial value to increase stability) during the fit. The normalization factors we have obtained piecewise in dedicated CRs as discussed in Section 6 are used just as starting points. These include $\mathrm{t\bar{t}Z}$, WZ, ZZ, $\mathrm{Z}\gamma$, and

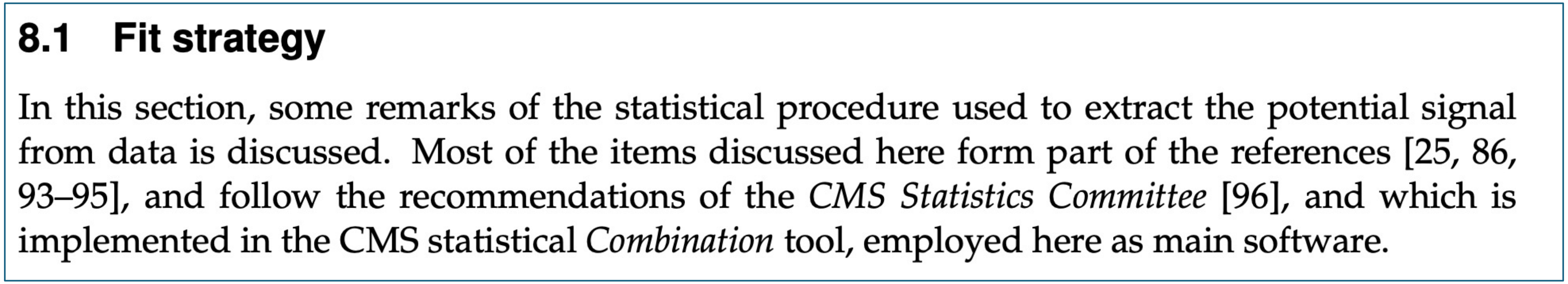
SUS-23-018   (AN-2023/042)

- Search for DM in association with b-quark and lepton pairs.
- They use a simultaneous fit of CRs and SR



In addition to those type of nuisance parameters, another kind of parameter is introduced in the fit to allow for the correction of the normalization of the main background sources (see Sec. 6). An unconstrained parameter ('rateParam') controlling the normalization and linking all the selection regions is associated to each main background process. The *simultaneous maximum likelihood fit* across all regions allows the fit first to constrain such parameter primarily from the dedicated CR for a given process, and secondly to correct any bias in normalization of the process arsing as consequence of the lack of good modeling in the MC simulation. For this, the events selected in each CR are counted in a single bin, that is simultaneously fit with the SR, where the respective background normalization is kept initially unknown (unconstrained).

- It is noteworthy that a SUSY third gen. convener is a primary author on this search and it is stated in this AN that this is the recommended treatment according to the statistics committee:

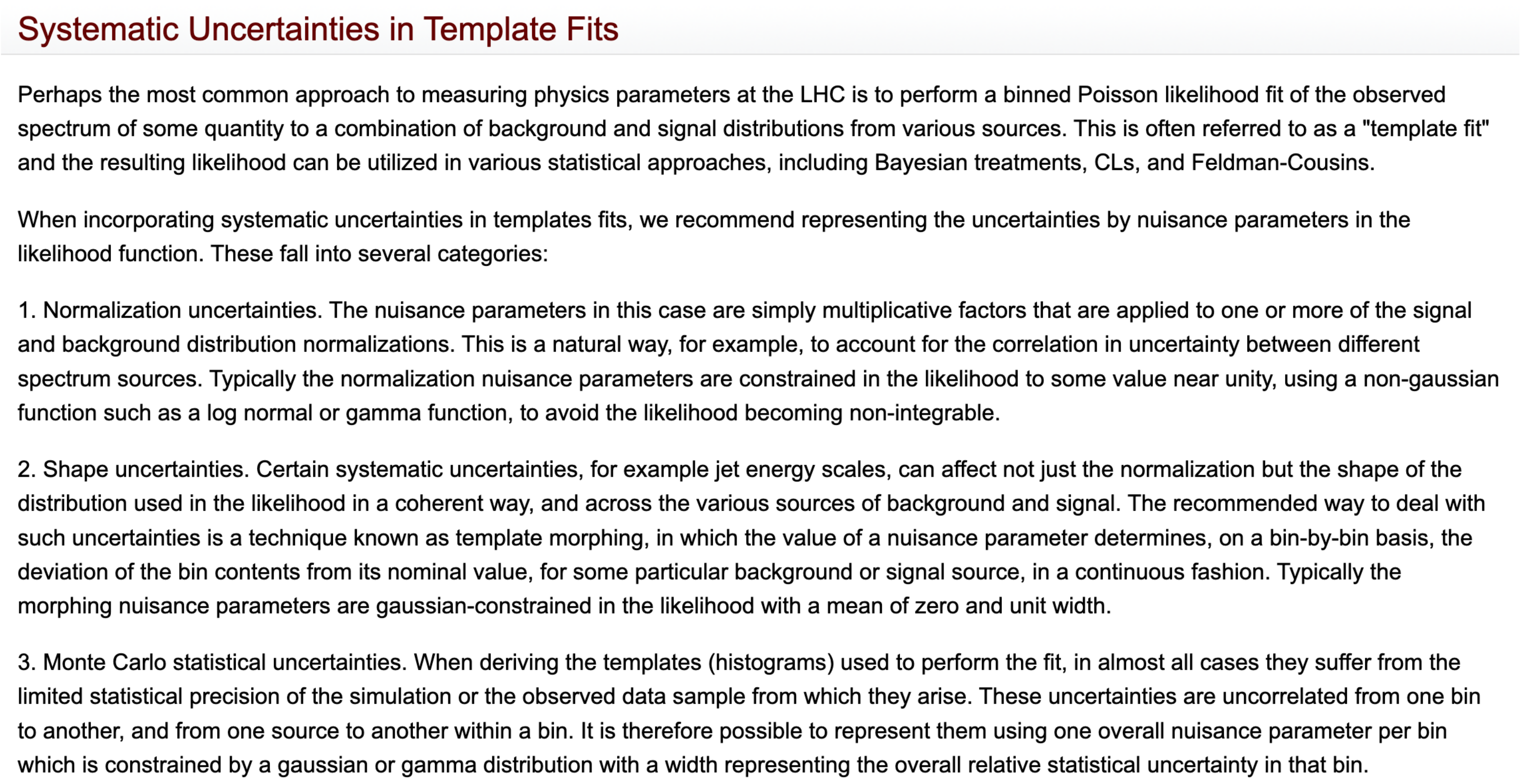


8.1 Fit strategy

In this section, some remarks of the statistical procedure used to extract the potential signal from data is discussed. Most of the items discussed here form part of the references [25, 86, 93–95], and follow the recommendations of the *CMS Statistics Committee* [96], and which is implemented in the CMS statistical *Combination* tool, employed here as main software.

- Statistics committee reccomnedations can be found here, though I couldn't immediately identify one regarding simultaneous fits of CRs with SR.

https://twiki.cern.ch/twiki/bin/viewauth/CMS/StatisticsCommittee#Recommendations_from_the_Committ



Systematic Uncertainties in Template Fits

Perhaps the most common approach to measuring physics parameters at the LHC is to perform a binned Poisson likelihood fit of the observed spectrum of some quantity to a combination of background and signal distributions from various sources. This is often referred to as a "template fit" and the resulting likelihood can be utilized in various statistical approaches, including Bayesian treatments, CLs, and Feldman-Cousins.

When incorporating systematic uncertainties in templates fits, we recommend representing the uncertainties by nuisance parameters in the likelihood function. These fall into several categories:

1. Normalization uncertainties. The nuisance parameters in this case are simply multiplicative factors that are applied to one or more of the signal and background distribution normalizations. This is a natural way, for example, to account for the correlation in uncertainty between different spectrum sources. Typically the normalization nuisance parameters are constrained in the likelihood to some value near unity, using a non-gaussian function such as a log normal or gamma function, to avoid the likelihood becoming non-integrable.

2. Shape uncertainties. Certain systematic uncertainties, for example jet energy scales, can affect not just the normalization but the shape of the distribution used in the likelihood in a coherent way, and across the various sources of background and signal. The recommended way to deal with such uncertainties is a technique known as template morphing, in which the value of a nuisance parameter determines, on a bin-by-bin basis, the deviation of the bin contents from its nominal value, for some particular background or signal source, in a continuous fashion. Typically the morphing nuisance parameters are gaussian-constrained in the likelihood with a mean of zero and unit width.

3. Monte Carlo statistical uncertainties. When deriving the templates (histograms) used to perform the fit, in almost all cases they suffer from the limited statistical precision of the simulation or the observed data sample from which they arise. These uncertainties are uncorrelated from one bin to another, and from one source to another within a bin. It is therefore possible to represent them using one overall nuisance parameter per bin which is constrained by a gaussian or gamma distribution with a width representing the overall relative statistical uncertainty in that bin.