The Compact Muon Solenoid Experiment
Conference Report

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

12 August 2009 (v2, 18 August 2009)

MUSiC - An Automated Scan for Deviations between Data and Monte Carlo Simulation

Arnd Meyer for the CMS Collaboration

Abstract

A model independent analysis approach is presented, systematically scanning the data for deviations from the standard model Monte Carlo expectation. Such an analysis can contribute to the understanding of the CMS detector and the tuning of event generators. The approach is sensitive to a variety of models of new physics, including those not yet thought of.

Presented at SUSY09: 17th International Conference On Supersymmetry And The Unification Of Fundamental Interactions

MUSiC – An Automated Scan for Deviations between Data and Monte Carlo Simulation

Arnd Meyer¹, meyera@physik.rwth-aachen.de

III. Physikalisches Institut A, RWTH Aachen, 52056 Aachen, Germany

Abstract. A model independent analysis approach is presented, systematically scanning the data for deviations from the standard model Monte Carlo expectation. Such an analysis can contribute to the understanding of the CMS detector and the tuning of event generators. The approach is sensitive to a variety of models of new physics, including those not yet thought of.

Keywords: New Phenomena, Hadron Collider, LHC, CMS

PACS: 12.60.-i, 13.85.Rm, 02.70.Uu, 29.85.Fj

The startup of the LHC opens up unknown territory in particle physics. However, the initial understanding of the detector and the validation of the Monte Carlo (MC) generator predictions are important tasks which precede most searches for new physics. In addition, it is not clear which effects of new physics will appear in the first data, and which theory beyond the standard model (SM) will describe them. Therefore it is planned to systematically analyze the data with as little bias as possible. A special algorithm called “MUSiC” (Model Unspecific Search in CMS) has been developed. Details of the strategy, the algorithm and some representative results are presented in [1] and summarized here, assuming an integrated luminosity of 1 fb^{-1} at $\sqrt{s} = 14 \text{ TeV}$. Similar strategies have already been applied successfully at other experiments [2, 3].

The present study is restricted to events containing at least one lepton (e or μ). Events are classified into event classes which group events according to their final state topology. Each class is defined by the number of physics objects in the event, e.g. $1\mu \text{ 3jet}$. Exclusive and inclusive event classes are considered: in the inclusive classes a minimal number of objects (e.g. $1\mu \text{ 3jet} + X$, at least one muon and 3 jets) is required. The following physics objects are included: muons (μ), electrons (e), photons (γ), hadronic jets, and missing transverse energy (E_T). This leads to approximately 300 – 400 event classes. Inclusion of additional objects such as τ leptons or b -jets is planned for the future. It is assumed that new physics will appear in events with high- p_T objects.

The CMS software framework including a full detector simulation is used in order to process the simulated samples and to reconstruct the physics objects. The SM backgrounds W +jets, Z +jets and $t\bar{t}$ +jets are simulated with the ALPGEN [4] generator. Diboson samples, lepton enriched QCD, bottomonia and charmonia events as well as QCD multi-jets, γ +jets backgrounds, and minimum bias events are simulated using PYTHIA [5]. Some benchmark supersymmetry (SUSY) samples are generated using the SoftSusy [6] program to calculate the mass spectrum, and PYTHIA for the event genera-

¹ On behalf of the CMS collaboration

tion. For the dominant backgrounds of many new physics signals ($t\bar{t}$ +jets and W/Z +jets) a constant k -factor has been applied consistently for all subsamples in order to reweight the leading order cross section to the next-to-leading order (NLO) prediction obtained from MCFM [7]. For SUSY the Prospino 2 [8] NLO order cross sections are used.

Event selection. Single and dilepton triggers (e, μ) are used. Muons which are measured by the muon system and the inner tracker are selected with $p_T(\mu) > 30$ GeV (well above the trigger threshold) and $|\eta(\mu)| < 2.1$. Isolation is required, mainly to reject nonisolated muons from heavy flavor decays. Additional offline criteria are applied to reject misreconstructed muon candidates. Electrons are identified using tracker and calorimeter information. Isolated well-identified electrons with $p_T(e) > 30$ GeV and $|\eta(e)| < 2.5$ are selected. Jets reconstructed with the “iterative cone” algorithm with a radius of $R = \sqrt{\Delta\phi^2 + \Delta\eta^2} = 0.5$ are selected. Jet energy scale (JES) corrections are applied, and jets with $p_T(\text{jet}) > 60$ GeV and $|\eta(\text{jet})| < 2.5$ are selected. The hadronic fraction of jets is required to be $E_{\text{had}}/E_{\text{tot}} > 0.05$. Finally, the missing transverse energy is considered when exceeding $\cancel{E}_T > 100$ GeV, after accounting for JES corrections and subtracting muon momenta. Further cuts are applied to reduce ambiguities, misidentifications and duplications of the various physics objects.

The search algorithm. The composition of the selected events, i.e. the number of muons, jets etc. determines to which event class it is assigned. At the present time three distributions are investigated for each event class: the scalar sum of the transverse momentum $\sum p_T$ of all physics objects; the invariant mass M_{inv} of all physics objects (transverse mass M_T for classes with $\cancel{E}_T$); and for classes with missing transverse energy: $\cancel{E}_T$. The $\sum p_T$ distribution is the most general observable. The invariant mass has an obvious advantage for new particles produced as resonances. Models beyond the SM which aim to provide a dark matter candidate might be spotted with the $\cancel{E}_T$ distribution. The implementation of additional variables is straightforward.

All distributions are input to the MUSiC algorithm (similar to [3]) which scans them systematically for deviations, comparing the MC prediction with the measured data. Each connected bin region is considered within the distributions, including individual bins as well as broad regions. For each connected region, a counting experiment is performed, adding up the various Monte Carlo contributions (N_{SM}) and comparing this sum to the data (N_{data}). In addition to these two numbers, the systematic uncertainty of the prediction δN_{SM} is used. A Poisson probability is computed, determining how likely the prediction fluctuates to the number seen in the data. The systematic uncertainties, taking correlations into account, are included using a convolution with a Gaussian function. From all possible combinations of connected bins, the region with the smallest p -value ($p_{\text{min}}^{\text{data}}$) is chosen. This is called the *region of interest*. This approach is sensitive to an excess of data as well as a deficit. A statistical penalty factor derived from toy MC experiments is included to account for the number of regions investigated, thus obtaining the *event class significance* $\tilde{P}$. The value of $\tilde{P}$ can be translated into standard deviations and is comparable to the widely used CL_b [9].

The following systematic uncertainties are included in the present study. The algorithm accounts for known correlations within one error when computing p -values and

when generating pseudo-data. A 10% cross section uncertainty on all SM processes reflects the current theoretical knowledge which might improve in the future. The uncertainty on the integrated luminosity is assumed to be 5%. The JES uncertainty is expected to be 5%, and is also propagated to the value of $\cancel{E}_T$. The reconstruction and identification efficiencies for electrons and muons are assumed to be known with a precision of 2% (1% for jets), while the uncertainty of the misidentification rate for electrons and muons is 100%. Finally, the statistical uncertainty of the MC prediction is included. The contributions of other background sources can be regarded as another systematic uncertainty.

Results. The sensitivity of MUSiC is tested with typical models. In addition to producing pseudo-data for the background-only hypothesis, one can also assume *signal + background*, i.e. add a signal to the SM distributions. Several pseudo-experiments are repeated, and the expected significance of a possible signal is determined. MUSiC has been tested for various use cases and benchmark signals. Three main scenarios which fit to the concept of a model independent search are detector effects and generator tuning, a prominent single deviation, and complex multiple deviations. In this paper, only the latter is discussed, while examples for the former two are found in [1]. Examples for detector effects are an underestimation of the JES uncertainty, and examples for a single prominent deviation include new gauge bosons, which would only show up in one or few event classes. In MUSiC, as a reasonable threshold for a significant deviation, event classes with $\tilde{P} < 1 \cdot 10^{-3}$ ($> 3.3\sigma$) are considered to be “interesting”.

There are several reasons why a generic analysis strategy is a good supplement to SUSY searches. The free parameters in most SUSY models lead to a large parameter space from which nature can pick the scenario realized. Through their decay chains, SUSY particles often lead to spectacular cascades with high multiplicities of leptons and jets and a large amount of $\cancel{E}_T$ due to the LSP (lightest SUSY particle). SUSY often does not predominantly favor a single topology, but contributes to a multitude of event classes. The generic search can give an overall picture of the SUSY signatures. The results shown here use a typical mSUGRA point (“LM4”) with $m_0 = 210$ GeV, $m_{1/2} = 285$ GeV, $\tan\beta = 10$, $\text{sgn}(\mu) = +$, $A_0 = 0$, and $\sigma(\text{NLO}) = 27.7$ pb. For LM4, the decay of the $\tilde{\chi}_2^0$ into on-shell Z ’s is characteristic. The production of $\tilde{q}\tilde{g}$ is dominant, contributing about half of the total cross section. The squark and gluino masses are below ~ 700 GeV, leading to relatively large total cross sections.

A global scan of all event classes is performed. In total 375 inclusive classes and 315 exclusive classes are populated (either signal and/or background MC). Deviations are found in a multitude of places: LM4 contributes to 160 (260) exclusive (inclusive) classes, 94 (170) classes with $\cancel{E}_T$:

15% (36%) show significant deviations with $\tilde{P}(\text{expected}) < 1 \cdot 10^{-3}$ in $\sum p_T$;

38% (59%) show significant deviations with $\tilde{P}(\text{expected}) < 1 \cdot 10^{-3}$ in $\cancel{E}_T$.

In the case of inclusive classes, the deviations are partially overlapping since 1μ 5jet events contribute to 1μ 2jet + X , 1μ 3jet + X and so on. When comparing similar final state topologies, the inclusive classes tend to have smaller expected event class significances than the exclusive ones. The two kinematic distributions examined, $\sum p_T$ and $M_{\text{inv}}(M_T)$, lead to similar results. However, one can observe a clear gain when using the $\cancel{E}_T$ distribution which is a prominent signature of the LSP.

In Fig. 1(a) the $\cancel{E}_T$ distribution for a single pseudo-experiment is shown. A deviation well above 4σ is found, despite the 20% – 30% systematic uncertainties on the $t\bar{t}$ +jets and W +jets backgrounds, in particular due to the JES uncertainties. Besides the single lepton plus jets event classes, another type of class with significant deviations contains multileptons plus jets ($+\cancel{E}_T$). Many combinations of 2 or 3 leptons show small values of $\tilde{P}$. One example is the $1e\ 1\mu\ 3\text{jet}\ \cancel{E}_T + X$ class where a deviation $> 4.4\sigma$ is found in Σp_T (region of interest 1000 – 2650 GeV with $N_{\text{data}} = 188$ and $N_{\text{MC}} = 61 \pm 18$).

FIGURE 1. (a) Distribution of $\cancel{E}_T$ in an event class with single lepton+jets+ $\cancel{E}_T$. Region of interest: $N_{\text{data}} = 424$ and $N_{\text{MC}} = 142 \pm 40$. (b) Frequency distribution of the $\tilde{P}$ values using all exclusive event classes which have pseudo-data entries, using the Σp_T distribution. Histogram: averaged CMS experiment (SM only), points: single experiment with LM4.

In order to quantify the global compatibility of data and SM expectation, one can plot the frequency distribution of the $\tilde{P}$ values using all event classes. If there is a signal leading to deviations in several event classes, one would expect the tails of this distribution to differ from the SM-only scenario. More entries than expected with small $\tilde{P}$ should be observed. Fig. 1(b) is an example for such a distribution, using the Σp_T distribution in the exclusive case. Here the $\tilde{P}$ values ($-\log_{10} \tilde{P}$, thus $3 \hat{=} 3.3\sigma$) of all event classes with pseudo-data entries are shown. One can clearly see that SUSY leads to significant deviations in numerous classes. Note that classes where only an upper limit can be set ($\tilde{P} < X$, indicated by the red arrow) all contribute to the rightmost bin.

REFERENCES

1. CMS Collaboration, PAS EXO-08-005, 2008.
2. B. Abbott *et al.* (D0 Collaboration), Phys. Rev. D **62** (2000) 092004; V.M. Abazov *et al.* (D0 Collaboration), Phys. Rev. D **64** (2001) 012004; D0 Collaboration, Conference Note 5777-CONF (2009); T. Aaltonen *et al.* (CDF Collaboration), Phys. Rev. D **78** (2008) 012002, Phys. Rev. D **79** (2009) 011101; T. Hebbeker, L3 note 2305 (1998).
3. A. Aktas *et al.* (H1 Collaboration), Phys. Lett. B **602** (2004) 14.
4. M.L. Mangano *et al.*, ALPGEN, JHEP 0307:001 (2003), hep-ph/0206293.
5. T. Sjostrand, S. Mrenna and P. Skands, PYTHIA, JHEP 0605:026 (2006), arXiv:hep-ph/0603175.
6. B.C. Allanach, Comput. Phys. Commun. **143** (2002) 305, arXiv:hep-ph/0104145.
7. J.M. Campbell and R.K. Ellis, Phys. Rev. D **60**, 113006 (1999), <http://mcfm.fnal.gov/>.
8. W. Beenakker, R. Hopker and M. Spira, arXiv:hep-ph/9611232.
9. A.L. Read, J. Phys. G **28** (2002) 2693.