PHYSICAL REVIEW D **78,** 035012 (2008)

Background modeling in new physics searches using forward events at CERN LHC

Victor Pavlunin and David Stuart

Department of Physics, University of California, Santa Barbara, California, 93106-9530, USA

(Received 13 June 2008; published 28 August 2008)

We present a method to measure dominant standard model backgrounds using data containing high rapidity objects in pp collisions at the Large Hadron Collider (LHC). The method is developed for analyses of early LHC data when robustness against imperfections of background modeling and detector simulation can be a key to the discovery of new physics at LHC.

DOI: 10.1103/PhysRevD.78.035012 PACS numbers: 12.60.−i, 13.85.Qk, 14.70.Fm, 14.70.Hp

I. INTRODUCTION

The Large Hadron Collider (LHC) will soon start operating in an unexplored energy regime at $\sqrt{s} \sim 14$ TeV, about 7 times higher than that achieved at the Tevatron. At that center-of-mass energy, a large number of new particles could be produced even in a data sample of modest integrated luminosity. The challenge is to distinguish events with new particles from those, many orders of magnitude more copious, attributed to the standard model (SM), and to do so using tools and methods appropriate for early data. The challenge is magnified by the fact that signatures of the physics beyond the SM realized in nature are not known.

Heavy new particles are produced, approximately at threshold, via interactions of energetic partons. Their decay products tend to be distributed uniformly over solid angle, which corresponds to a narrow central rapidity region [1]. SM particles are light on the mass scale of the LHC and tend to be produced in interactions of soft, often very asymmetric in energy, partons. They receive a significant boost along the beam line, which makes them distributed over a wide rapidity range.

In this paper, we present a new method to measure dominant SM backgrounds in searches for heavy new particles. It uses data containing high rapidity objects to predict SM yields at small rapidity. We apply this to the SM processes: Z + jets, W + jets, γ + jets, QCD jets, and $t\bar{t}$, that are the largest background sources in many new physics searches. We also discuss the usage of a ratio constructed from event yields in central and forward rapidity regions as a generic search variable.

The method is presented in the context of a new physics search involving leptons, photons, jets and missing transverse energy. In the absence of a single most compelling model of new physics, the search is developed in a model independent way. The only assumption we make is that new particles are heavy and they decay to SM particles via a multistage cascade producing a large number of jets, so that the number of jets is a main search variable. A key feature of our method is that systematic uncertainties associated with incomplete knowledge of the SM production rates and detector artifacts cancel to first order. The emphasis throughout is on robustness against imperfections of background modeling required for new physics searches in early LHC data.

II. METHOD OVERVIEW

We consider final states involving many jets, 4 or more. The SM V + jets production rates, where for brevity V stands for a Z, W, γ, or a jet [2], fall steeply as the number of jets grows, but they are difficult to predict from first principles. Monte Carlo (MC) techniques are unreliable in predicting backgrounds with a large number of jets. Theory calculations [3] do not exist at sufficiently high order. The structure functions have significant uncertainties for partons carrying a small fraction, x, of the proton momentum that is relevant for LHC [4]. Large uncertainties in the calibration of the experimental apparatus are expected in early data taking. For these reasons, instead of relying on MC simulation of the detector response to SM processes, we use control regions in data to determine dominant SM backgrounds. We identify control samples in kinematic regimes where the SM dominates and extrapolate backgrounds measured there into the signal region where new physics may contribute. In V + jets, the SM dominates when the transverse momentum, $|\vec{p}_T|$, of V or the number of jets, N_J, is small. These control regions have been used previously for data-based background determination [5]. We use, in addition, control samples with high rapidity objects that are background dominated even when $|\vec{p}_T|$ or N_J is large. Jet rapidity has been successfully used previously in di-jet resonance searches at the Tevatron [6].

Figure 1 shows the (pseudo-)rapidity distributions for Z + jets (a), W + jets (b), γ + jets (c), and multijets (d). In the Z + jets channel, we use the rapidity of the Z boson, y_Z, as a key discriminating rapidity variable. The W boson rapidity cannot be unambiguously determined due to the undetected neutrino. We instead use the lepton pseudorapidity [1], η_{lepton}, for W + jets. The pseudorapidities of the photon, η_γ, and the highest $|\vec{p}_T|$ jet, η^*_{jet}, are used for γ + jets and multijets, respectively. As seen in Fig. 1, the (pseudo-)rapidity distributions for decays of new massive particles are central, while that for the SM processes are approximately uniform in a wide rapidity range.

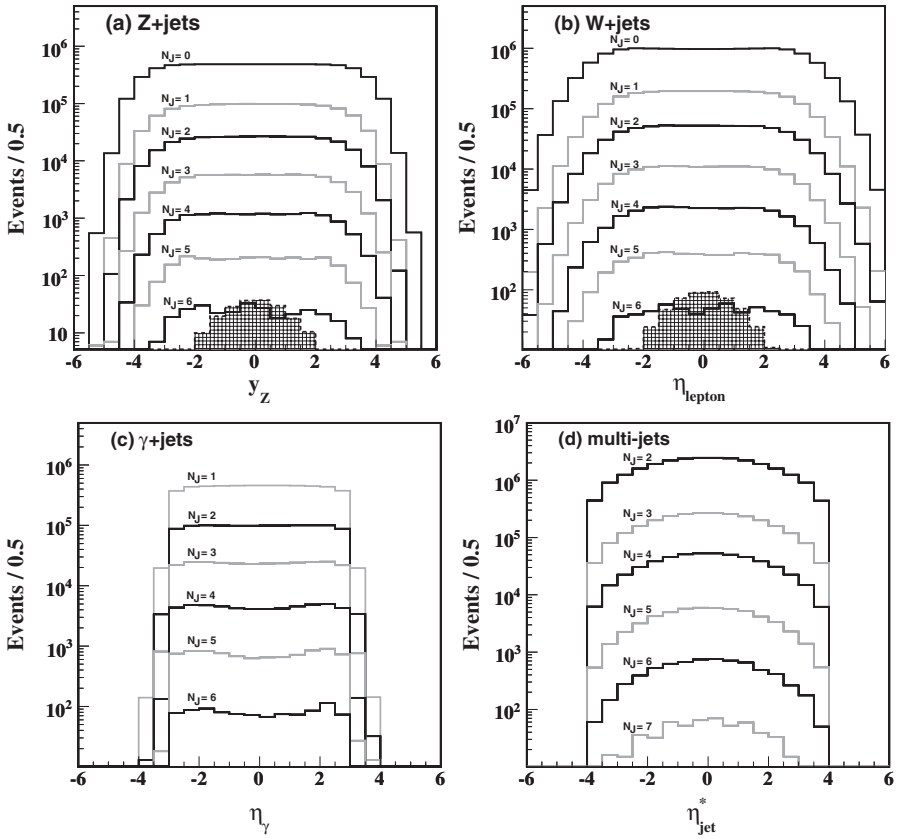

FIG. 1. Rapidity of Z bosons from SM Z + jets (a), pseudorapidity of charged leptons from SM W + jets (b), pseudorapidity of γ for SM γ + jets (c), and pseudorapidity of the highest $|\vec{p}_T|$ jet in SM QCD multijet events (d). Generator level requirements of $|\eta_\gamma| < 3.0$ and $|\eta_{\text{jet}}| < 4.0$ are imposed in plots (c) and (d). Shapes of rapidity distributions from LM4 and LM6 mSUGRA benchmark points [14] are shown by black hatched histograms in the Z + jets and W + jets cases, respectively.

Furthermore, the rapidity distributions vary slowly as the number of jets increases.

The object providing the discriminating rapidity variable is called a tag [7]. We use events with forward tags to determine backgrounds for events with central tags, using an algorithm described in Sec. IV.

In this paper, for brevity, we discuss searches at high N_J, since N_J is a particularly simple and robust variable. Other distributions considered in our search include: the highest jet $|\vec{p}_T|$ ($|\vec{p}_T^{\text{lead}}|$) and the $J_T \equiv \sum |\vec{p}_T^{\text{jet}}|$ spectra in each N_J bin; and N_J^* distributions, which are closely related to N_J but obtained as a sum of weights of either $|\vec{p}_T^{\text{lead}}|$ or J_T in each N_J bin. The N_J^* distributions have higher discriminating power compared to the N_J distributions since new particles are expected to be heavy. However, reliance on the $|\vec{p}_T^{\text{lead}}|$ or J_T spectra is more susceptible to uncertainties in the jet energy scale.

III. EXPERIMENTAL ASPECTS

The ATLAS and CMS experiments use multipurpose detectors that are in the final stages of construction at the European Organization for Nuclear Research (CERN). Detailed descriptions of the detectors can be found in Ref. [8]. Of primary importance for our studies are the detectors' rapidity coverages and kinematic thresholds. The detectors are capable of efficiently reconstructing electrons and muons with low fake rates for lepton $|\vec{p}_T| > 20$ GeV within $|\eta| < 2.5$. Photons and jets are reconstructed in the $|\eta| < 2.5$ and $|\eta| < 3.0$ range, respectively. Missing transverse energy, E_T^{miss}, is calculated using E_T measurements of all reconstructed objects in each event. Mismeasured or misreconstructed objects, calorimeter noise, malfunctioning detector subsystems and channels, and background unrelated to pp collisions constitute sources of unphysical E_T^{miss} that may complicate the usage of E_T^{miss} in early searches. Accordingly, we perform studies with and without a requirement on E_T^{miss} in the event selection.

To study the effectiveness of the method, we have produced mock data samples for the following SM processes: Z + jets (5.0 fb^{-1}, up to 5 partons, $Z \to l^+l^-$), W + jets (1.0 fb^{-1}, up to 5 partons, $W \to l\nu_l$), $t\bar{t}$ (1.0 fb^{-1}, up to 4 partons, $t\bar{t} \to l\nu_l bbjj$ and $t\bar{t} \to l\nu_l \tau\nu_\tau bb$), γ + jets (400.0 pb^{-1}, up to 5 partons), and QCD jets (1.0 pb^{-1}, up to 5 partons), where l is μ or e. The integrated luminosity indicated in parentheses for each channel specifies the sample size used in our studies, except where specified

otherwise. These samples were generated with ALPGEN [9] using CTEQ5L parton distribution functions (PDFs) [10], and PYTHIA [11] was used for parton showering, hadronization, simulation of the underlying event, and jet reconstruction. To model features of a new physics signal in search distributions, we produced mock signal data samples for minimal supergravity (mSUGRA) benchmark points LM4 and LM6 [12–14] using PYTHIA.

Kinematic selection criteria are applied as follows. Electrons and muons are required to have $|\vec{p}_T|$ of at least 20 GeV in the $|\eta| < 2.5$ range. Photons are reconstructed above the $|\vec{p}_T|$ threshold of 30 GeV in the $|\eta| < 2.5$ range. Jets are reconstructed using the PYCELL algorithm [11] and required to be within $|\eta| < 3.0$ for $|\vec{p}_T|$ thresholds varying between 30 and 100 GeV. Low thresholds are used for background studies, while higher thresholds are used to study signal dominated regions.

Detector response is not directly simulated, although an assumed reconstruction efficiency of 50% is applied in each channel. The $E_T^{\rm miss}$ vector is approximated by a vector opposite to the sum of $\vec{p}_T$ measurements of charged leptons, photons, and jets. Using the γ + jets sample, we find that the jet energy resolution function in our mock data samples is approximately Gaussian with σ varying from about 15% at 30 GeV to about 8% at 100 GeV. To simulate effects of $E_T^{\rm miss}$ mismodeling due to jet energy fluctuations with non-Gaussian tails and incomplete hermeticity of the detectors, we perform robustness tests where jet energies are varied according to the hypothetical probability density function shown in Fig. 2, and jets are removed in selected regions, as described in Sec. VI.

These selection criteria and sample sizes are chosen generally and are not optimized to any new physics model. The new physics reference models listed above are used only for illustration. Our goal in this paper is to demonstrate the scope of the method and its performance rather than to attain high sensitivity to a specific model for a specific final state or quantify that sensitivity.

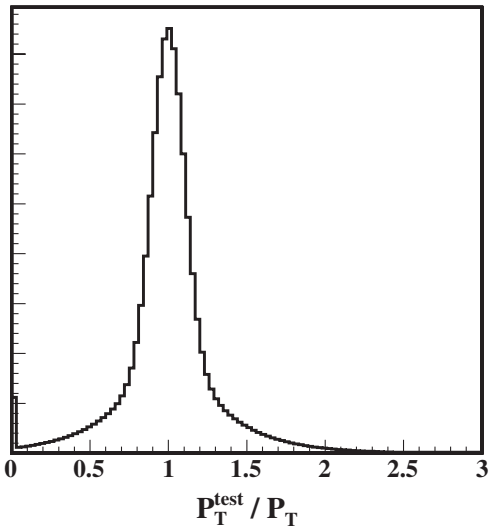

FIG. 2. A hypothetical probability density function used for jet energies in modeling the effect of artificial $E_T^{\rm miss}$.

IV. ALGORITHM

To describe and illustrate the algorithm and tests of its robustness, in the next several sections we center the discussion on the Z + jets channel. The discussion applies to all four V + jets channels, however, and differences among these channels are pointed out where significant.

The rapidity range for reconstructed Z bosons passing realistic event selection criteria is reduced (Fig. 3). We define forward events as those with a Z boson having $|y_Z| > 1.3$, and we call the detector region with $|\eta| > 1.3$ the forward region. Central events are defined as those with a Z boson at $|y_Z| < 1.0$, and the central region of the detector as that having $|\eta| < 1.0$. (This definition of central and forward categories is arbitrary and could be modified without significant effect.)

Small N_J bins are SM dominated for both central and forward events, and we use them to predict the SM contribution to the central, high N_J bins where signal would appear. This is done by measuring a ratio, denoted as R_{N_J}, of the central yield ($Y_{N_J}^{\rm Central}$) to the sum of forward ($Y_{N_J}^{\rm Forward}$) and central yields in each N_J bin: $R_{N_J} \equiv Y_{N_J}^{\rm Central}/(Y_{N_J}^{\rm Forward} + Y_{N_J}^{\rm Central})$. A linear fit to R_{N_J} is made in the low N_J bins and extrapolated into the high N_J region. The extrapolated ratios and the yields of forward events in

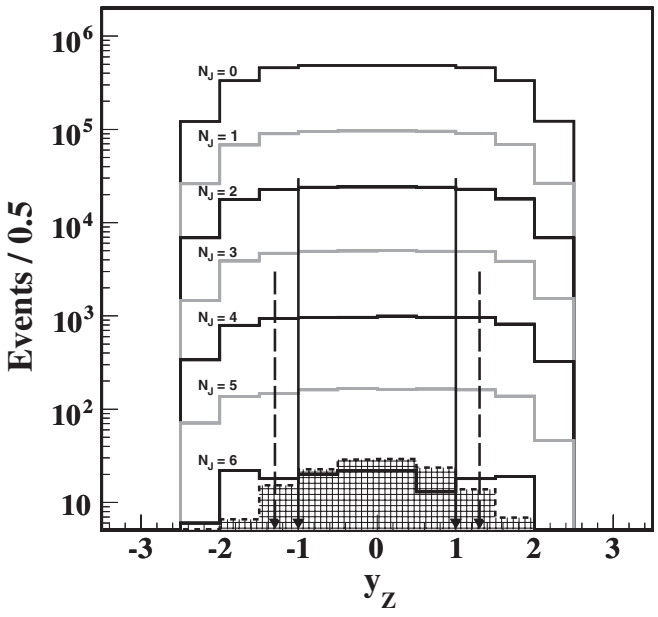

FIG. 3. Rapidity of Z bosons from the SM Z + jets production in the fiducial coverage of LHC detectors. The central signal region is indicated by solid arrows. The background dominated region is at $|y_Z|$ larger than that indicated by dashed arrows. The Z rapidity shape from the LM4 mSUGRA benchmark is shown by the black hatched histogram.

high N_J bins are combined to obtain a background prediction in the central, high N_J signal region.

The accuracy of this background prediction can be tested in mock data samples by comparing it to the yield in the central region at high N_J. This estimated-to-observed comparison is shown as a function of N_J in Fig. 4 for Z + jets, W + jets, γ + jets, and pure QCD jets. The prediction is made using fits in $1 \leq N_J \leq 3$ for Z + jets and W + jets. For γ + jets and multijets, $2 \leq N_J \leq 4$ is used. The observed central yield at high N_J is well matched to the prediction in all cases. Pull distributions, defined as $(N_{\text{Observed}} - N_{\text{Estimated}})/\sigma_{\text{Stat}}$, where N_{Observed} is the observed number of central events, $N_{\text{Estimated}}$ is the number of central events estimated using the algorithm, and σ_{Stat} is the total statistical uncertainty, are in the bottom plot of the same figure in black markers of the appropriate shape for each channel. Shaded markers in the bottom plot show how the pulls change with the addition of a 1% relative systematic uncertainty in each N_J bin. With at most a small systematic uncertainty, the algorithm estimates the background in the central region accurately.

The results in Fig. 4 are obtained with a jet threshold of 30 GeV. A higher threshold would likely improve signal sensitivity, but it could also affect the algorithm's performance. As the jet threshold changes, the R_{N_J} values may change, but the low N_J fit should properly account for any difference. We search for the presence of biases by varying the jet threshold between 30 and 100 GeV and repeating the tests in Fig. 4(e) for the Z + jets and W + jets channels. No evidence of a bias is found.

The performance of the algorithm when a signal is present is illustrated in Fig. 5, where we compare the central yields and the predictions with and without a signal contribution. A clear excess of a signal above the back-

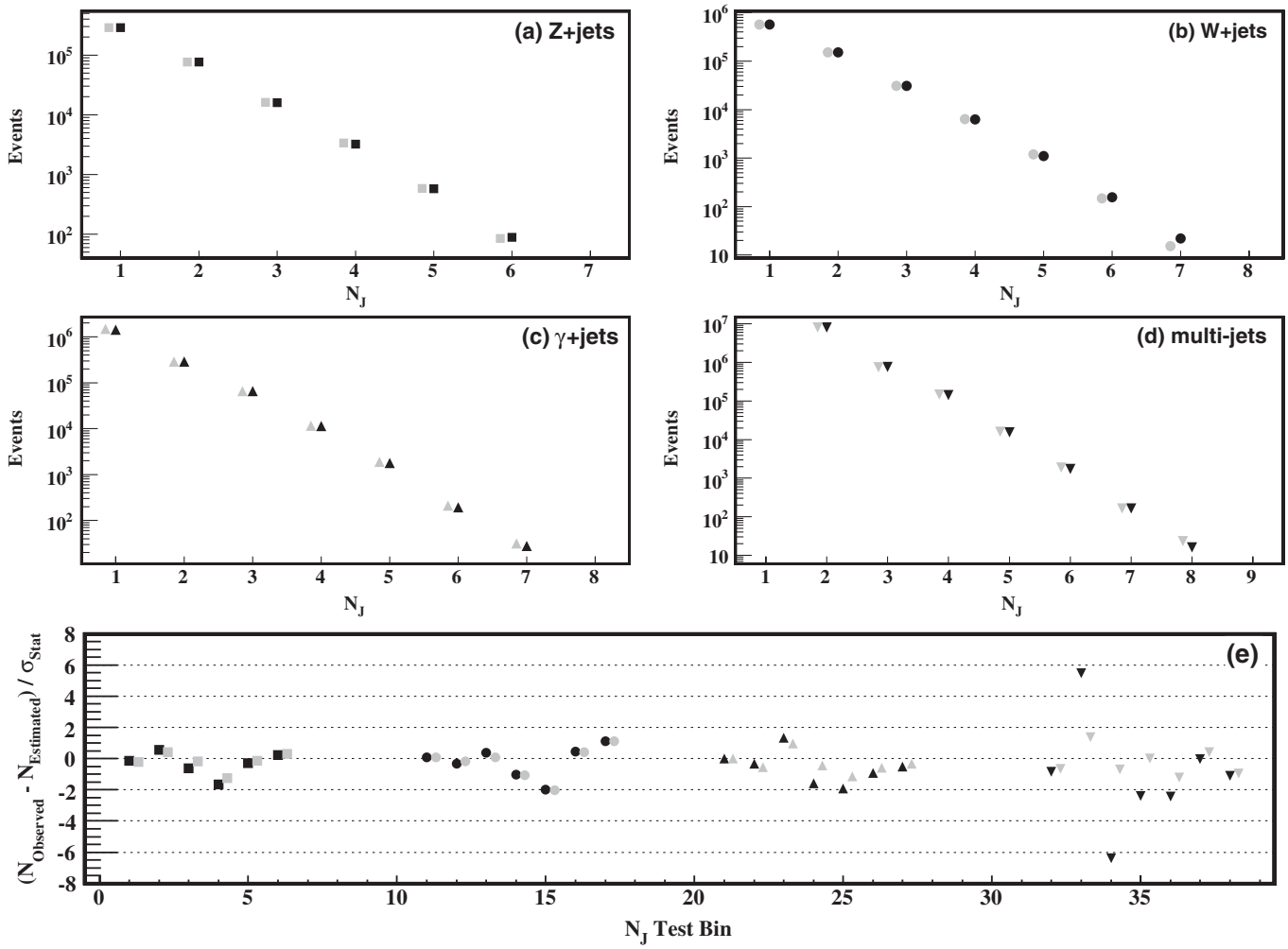

FIG. 4. The N_J distributions for Z + jets [(a), 5.0 fb^{-1}], W + jets [(b), 1.0 fb^{-1}], γ + jets [(c), 400 pb^{-1}], and pure multijets [(d), 1 pb^{-1}]. The backgrounds in the central regions are shown in black markers, its estimate is in shaded markers of the same shape displaced horizontally for visibility. A jet $|\vec{p}_T|$ threshold of 30 GeV is used. Bottom plot: pull distributions for Z + jets (black squares), W + jets (black circles), γ + jets (black triangle-up), and pure multijets (black triangle-down). Here, N_J is offset by 10 between samples for visibility, i.e., N_J = Test Bin mod 10. Shaded markers in the bottom plot show how the pulls change after an addition of a 1% relative systematic uncertainty in each N_J bin.

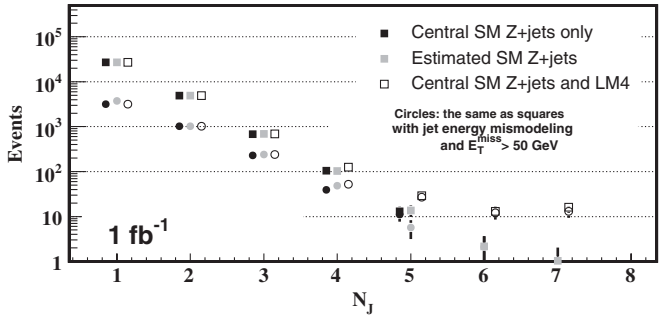

FIG. 5. N_J distributions for Z + jets (black markers) and a mixture of Z + jets and events from the LM4 mSUGRA benchmark (shaded markers: estimated central SM background, open markers: all central events). This comparison is made with a 50 GeV jet $|\vec{p}_T|$ threshold and a sample size corresponding to 1 fb^{-1}. The effect of a $E_T^{\rm miss} > 50$ GeV requirement for a sample with the jet energy mismodeling discussed in section VI is shown by the circles.

ground prediction is seen at large N_J. The integrated luminosity of the data sample in this figure is 1 fb^{-1}, and a jet $|\vec{p}_T|$ threshold of 50 GeV is used. Square markers show N_J distributions without a requirement on missing energy. The effect of a missing energy requirement is discussed in Sec. VI.

V. ROBUSTNESS

The main goal of our method is robustness against imperfections of the SM background modeling and detector simulation. By design, uncertainties in the background cross section are accounted by normalizing to the yield in the forward region. In addition, any systematic effect present in data should be taken into account by the background estimate, as long as the biases in R_{N_J} ratios associated with the effect are a linear or slowly varying function of N_J.

To examine the robustness of our method, we present a few illustrative tests. In each test, a change to the mock data samples is made and the analysis procedure is repeated. The results are presented in the form of pull distributions in Fig. 6, where only statistical uncertainties are used to normalize the differences between observed and estimated numbers of events.

The composition of the SM Z + jets sample, or other samples with a large number of jets, could differ from the ALPGEN predictions. To test the effect of such mismodeling, we separate the Z + jets sample into two subsamples with an even $\{0, 2, 4\}$ and odd $\{1, 3, 5\}$ number of ALPGEN partons and apply the analysis procedure to these subsamples. This is a particularly stringent test as it introduces drastic bin-to-bin variations in the N_J distributions. However, we find that the background is estimated accurately in most bins [Fig. 6 (top, bin range from 0 to 19)]. There are two bins, in W + jets and γ + jets, where the observed and estimated yields differ by about 3 standard

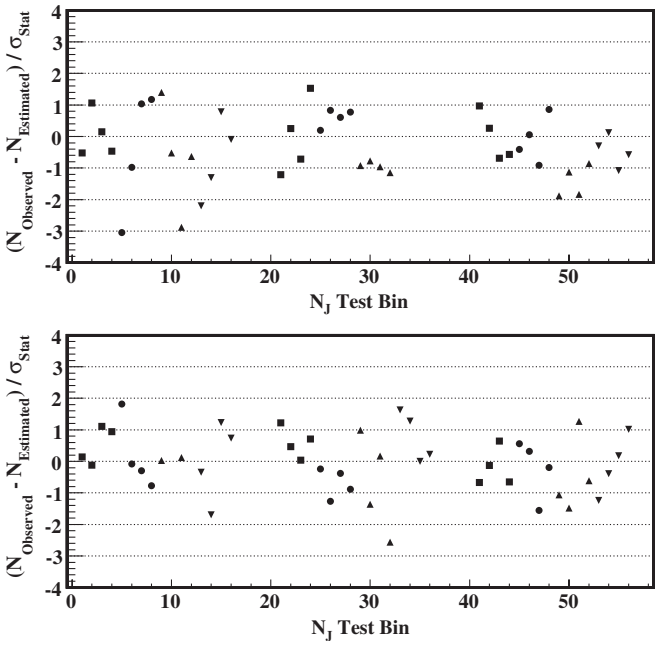

FIG. 6. Pulls between observed and estimated numbers of events for Z + jets (squares), W + jets (circles), γ + jets (triangle-up), and pure multijets (triangle-down) from robustness tests in Sec. V (top) and from tests with a requirement on $E_T^{\rm miss}$ in Sec. VI (bottom). Top: N_J test bins in ranges [0; 19], [20; 39], and [40; 59] correspond to tests without a requirement on $E_T^{\rm miss}$ consisting of changing the composition of the ALPGEN sample ($\{0, 2, 4\}$ and $\{1, 3, 5\}$ partons), lepton/photon efficiencies (over the entire η range and in the forward region), and jet efficiencies (over the entire η range and in the forward region), respectively. Bottom: N_J test bins in ranges [0; 19], [20; 39], and [40; 59] are from tests with an $E_T^{\rm miss}$ or M_T requirement, for different composition of the ALPGEN sample ($\{0, 2, 4\}$ and $\{1, 3, 5\}$ partons), hypothetical holes (over the entire η range and in the forward region), and fluctuations in jet energies (over the entire η range and in the forward region), respectively. In each test, pulls in the two highest N_J bins are plotted. (Note, pulls in these tests are correlated as tests are made using events drawn from the same mock data samples.)

deviations. These biases are attributed to changes in R_{N_J} associated with the migration of events from higher to lower N_J bins. An event with n jets reconstructed in the $(n-1)$ N_J bin has a higher probability to be a forward event, as forward jets are lost more often and the tag rapidity is correlated, although weakly, with the rapidity of the jet system recoiling against the tag.

Efficiencies for forward and central leptons are different. One might account for these differences by applying efficiency corrections measured from data, but these corrections will have significant uncertainties in early data taking. To test the robustness of the method against the mismodeling of lepton reconstruction efficiencies, we change forward or central efficiencies by 30%. We find that the background estimate remains accurate [Fig. 6 (top, bin range from 20 to 39)] [15].

Similarly, lepton fakes introduce background in the Z + jets and W + jets channels, and photon fakes in the γ + jets channel. Because the lepton and photon fake rates are expected to be a slowly varying function of N_J, background from such fakes should be accounted for accurately in our method. For example, a 5% QCD background contribution to Z + jets introduces a less than 1% discrepancy.

Significant uncertainties in the jet reconstruction efficiencies are expected during early data taking. To test the robustness of the method against such inefficiencies, jets are removed randomly with 30% probability. We find that the background estimate remains accurate [Fig. 6 (top, bin range from 40 to 59)]. More demanding tests related to jet reconstruction efficiency and jet energy mismeasurements are presented below in Sec. VI.

We have confirmed that effects associated with uncertainties in PDFs are accommodated by our method and do not bias the background prediction. The algorithm was also found to be robust in other tests not discussed here.

VI. PERFORMANCE WITH $E_T^{\rm miss}$

In the results presented above, no requirement is made on missing transverse energy, $E_T^{\rm miss}$. Requiring large $E_T^{\rm miss}$ could significantly suppress SM backgrounds, and it is expected to be efficient in a large class of new physics models, e.g., R-parity conserving SUSY searches [12,13]. It is challenging to rely solely on $E_T^{\rm miss}$ in analyses of early data, because $E_T^{\rm miss}$ is particularly difficult to model. However, it could be useful as an additional discriminator against SM backgrounds in the context of our algorithm.

Unphysical sources of $E_T^{\rm miss}$ include those associated with jet energy fluctuations, noise and inefficient regions of the calorimeters, which could all be larger in the forward region. Our method is expected to work well with a $E_T^{\rm miss}$ requirement, nonetheless. The rapidity of the tag is only weakly correlated with the rapidity of the jet system recoiling against the tag due to the boost along the beam line in the laboratory frame. As a result, the $E_T^{\rm miss}$ in the tag recoil system tends to be averaged over the entire rapidity coverage. Remaining effects can be accounted by low N_J bin fits to R_{N_J}.

We have made a set of robustness tests with a requirement on $E_T^{\rm miss}$ by introducing mismeasurements and evaluating the consistency of the method's predictions. We require $E_T^{\rm miss} > 50$ GeV [16] for Z + jets, γ + jets, and multijets. In W + jets, the undetected neutrino is a source of genuine $E_T^{\rm miss}$, and requiring $E_T^{\rm miss} > 50$ GeV would have little effect. Instead, we impose a requirement on the transverse mass, M_T, which is constructed from $E_T^{\rm miss}$ and the lepton's transverse momentum. Requiring $M_T > M_W + x$ GeV, where M_W is the W mass, is approximately equivalent in suppressing SM W + jets to requiring $E_T^{\rm miss} > x$ GeV for SM Z + jets. For robustness tests in the W + jets sample, we require $M_T > M_W + 50$ GeV. In all four channels, the angle between the highest $|\vec{p}_T|$ jet and the missing transverse momentum in the transverse plane is required to be larger than 0.15.

We repeat tests related to the ALPGEN composition of the mock data samples with a requirement on $E_T^{\rm miss}$. To emulate the effect of holes in the detector coverage, we completely remove jets that fall within a cone of $\Delta R \equiv \sqrt{\Delta\eta^2 + \Delta\phi^2} < 0.7$ around three points in the detector, at $\eta = 0$ and $\eta = \pm 2$, each at $\phi = 0$. The energy of each jet is varied according to the hypothetical probability density function shown in Fig. 2 which includes wide non-Gaussian tails. Pulls between the observed and estimated numbers of events in high N_J bins from these tests are shown in Fig. 6 (bottom). Good consistency between estimated and observed yields is seen. In these tests, the predictions are made based on only two N_J bins: $2 \le N_J \le 3$ for Z + jets and W + jets, and $3 \le N_J \le 4$ for γ + jets and multijets. We find that R_{N_J} values in $N_J = 1$ for Z + jets and W + jets, and $N_J = 2$ for γ + jets and multijets tend to decrease after an additional requirement on missing energy for the reason already discussed in Sec. V. These bins are excluded from the background prediction procedure. Events reconstructed in higher N_J bins are less sensitive to this effect since the correlation between $E_T^{\rm miss}$ and tag rapidities is weaker in events with multiple jets.

The effect of a $E_T^{\rm miss} > 50$ GeV requirement on a search in the Z + jets sample with the jet energy mismodeling over the entire rapidity coverage is shown in Fig. 5 in round markers. The $E_T^{\rm miss}$ requirement suppresses the SM Z + jets rate, but the suppression is a function of N_J. Nonetheless, our method continues to predict the background accurately, and a signal excess is clearly apparent above the background prediction.

VII. SM $t\bar{t}$

A search in the W + jets sample is complicated by the top quark. The $t\bar{t}$ process, with one of the top quarks decaying semileptonically and the other hadronically, produces the same signature as that of W + jets. Because of the large top quark mass, the W bosons from top decays tend to be produced at small rapidities, and they increase R_{N_J} ratios over that of W + jets.

Figure 7 shows results of the analysis procedure applied to a sample of W + jets and $t\bar{t}$ events, where the fit to the R_{N_J} distribution is made in $1 \le N_J \le 2$. The central yield is higher than the background prediction because of the top contribution; the pull distribution in the right column shows the significance of the $t\bar{t}$ excess. This demonstrates that the method works in revealing decays of massive particles, and it could be used to measure the $t\bar{t}$ cross section. However, $t\bar{t}$ complicates the search for other massive particles.

One approach to searching beyond $t\bar{t}$ would be to subtract the $t\bar{t}$ contribution, either using a prediction for its cross section, or an independent measurement. Another

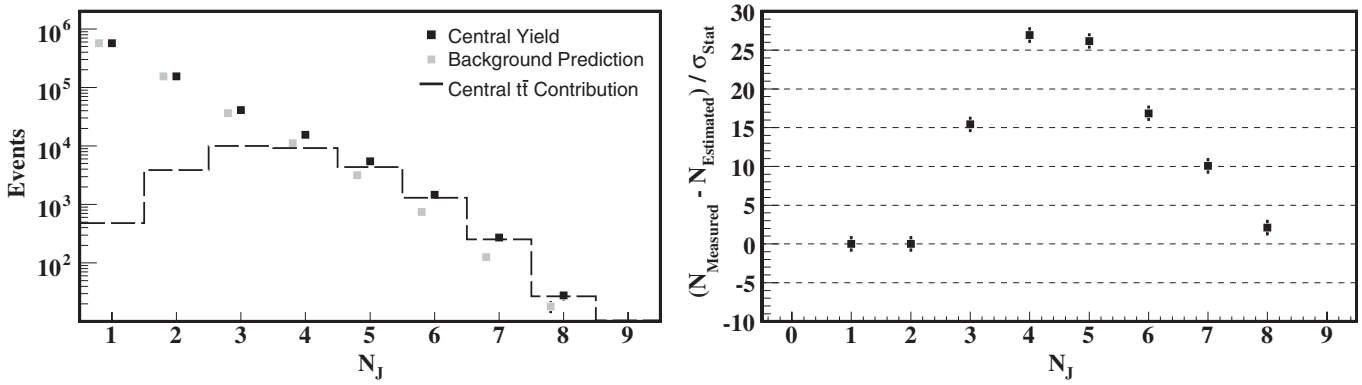

FIG. 7. Results of the analysis procedure applied to the combined W + jets and $t\bar{t}$ sample for selection criteria defined in Sec. III. Left: N_J distributions for the combined W + jets and $t\bar{t}$ sample. Right: pull distributions for the plots in the left column.

approach is to include the $t\bar{t}$ background in the fit. At high N_J, shifts in R_{N_J} caused by $t\bar{t}$ are a slowly varying function of N_J, so that the method should accommodate the combined W + jets and $t\bar{t}$ contribution in the background prediction.

Low mass mSUGRA models are challenging for searches in N_J as they produce N_J distributions peaking in the region where the $t\bar{t}$ contribution is maximal. Figure 8 illustrates this by comparing the central yield and prediction with and without a signal contribution. The LM6 mSUGRA benchmark is used and the comparison is made for a sample size corresponding to 1 fb^{-1}. A jet threshold of 50 GeV is used, and a transverse mass requirement of $M_T > M_W + 150$ GeV is applied to suppress SM backgrounds. There is a large signal contribution at $N_J \geq 4$, but it is not easily discernible above the central prediction made using $2 \leq N_J \leq 3$. The prediction is biased due to the residual $t\bar{t}$ contribution bridging between the W + jets dominated low N_J region and the signal dominated high N_J region. The $t\bar{t}$ and signal contributions together are large enough to bias the prediction. We discuss an alternative approach in the next section.

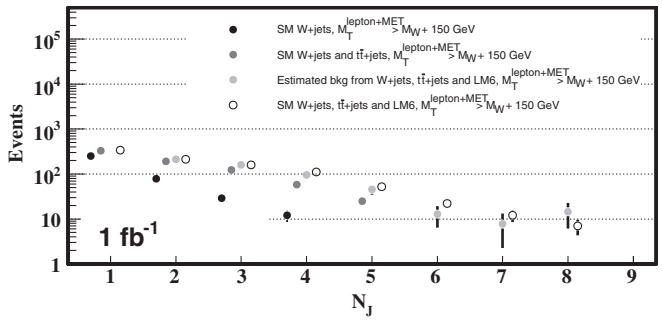

FIG. 8. N_J distributions for W + jets (black markers), W + jets and $t\bar{t}$ (shaded markers), and a mixture of W + jets, $t\bar{t}$ and events for the LM6 mSUGRA benchmark (open markers). Selection criteria on M_T are given in the legend.

VIII. SEARCH FOR NEW PHYSICS IN R_{N_J}

In the preceding discussion, we used fits to R_{N_J} to obtain a background prediction for the high N_J distribution in central events and searched for an excess signal there. Alternatively, we can search for new physics solely in the R_{N_J} distributions. The R_{N_J} ratios for heavy new particles are larger than that for SM processes, and a search for enhancements in the high N_J bins could reveal new phenomena or provide generic bounds on it.

Figure 9 shows the R_{N_J} distributions for a number of LHC processes. A distribution for minimum bias, i.e., low $|\vec{p}_T|$ scattering, events is shown for illustration purposes, where instead of jets, tracks with $|\vec{p}_T|$ above 3 GeV are used with the highest $|\vec{p}_T|$ track providing the rapidity tag. Distributions for SM processes studied in this paper, Z + jets, W + jets, γ + jets, and QCD jets, appear approximately in the middle of the available R_{N_J} range not far from that of the minimum bias events. The $t\bar{t}$ process contributes at higher R_{N_J} due to the large top quark mass. Distributions for LM4 and LM6 mSUGRA benchmarks in the Z + jets and lepton + jets + E_T^{miss} channels appear at higher R_{N_J} of about 0.8.

The Z + jets channel has little background, so identification of a new physics signal within it could be unambiguous. This is illustrated in Fig. 10(a), where the R_{N_J} distributions for SM Z + jets, with and without a new physics contribution (LM4 mSUGRA benchmark), are presented. The same threshold on jet $|\vec{p}_T|$ of 50 GeV as in Fig. 5 is used. Black markers show the SM Z + jets R_{N_J} distribution. It is reproduced accurately in a sample with LM4 by requiring $E_T^{\text{miss}} < 50$ GeV as shown in shaded markers. Alternatively, the SM Z + jets R_{N_J} shape in the sample with LM4 can be obtained based on $1 \leq N_J \leq 3$, where the relative contribution from LM4 is negligible. The new physics signal stands out clearly at $N_J \geq 5$ without any requirements on E_T^{miss}.

The W + jets channel is complicated by the $t\bar{t}$ contribution, as discussed in Sec. VII. Figure 10(b) shows the R_{N_J}

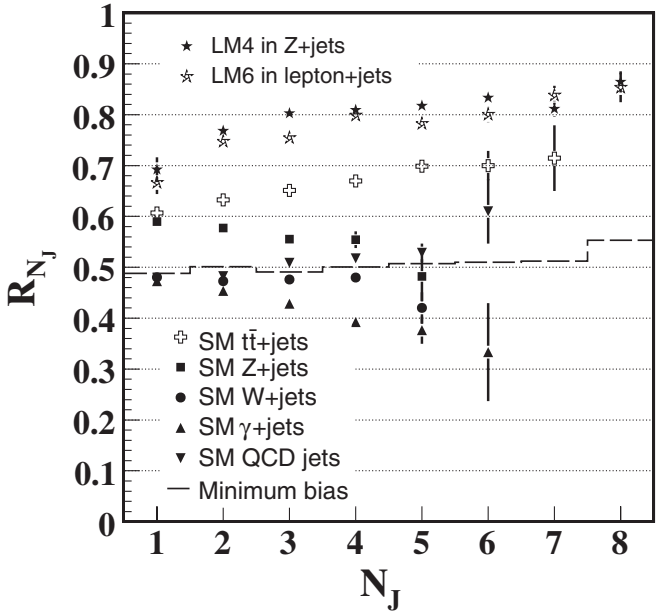

FIG. 9. R_{N_J} distributions for minimum bias events (track based, see the text), $t\bar{t}$ (crosses), Z + jets (squares), W + jets (circles), γ + jets (triangles-up), QCD jets (triangles-down), and new physics signals (stars).

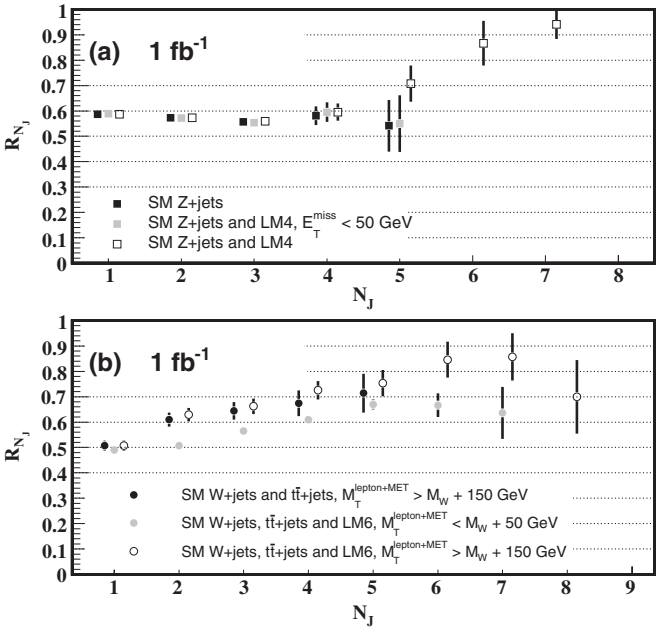

FIG. 10. Plot (a): R_{N_J} distributions for SM Z + jets (black markers) and a mixture of Z + jets and events for the LM4 mSUGRA benchmark (shaded markers: estimated central SM background, open markers: all central events). Plot (b): R_{N_J} distributions for W + jets (black markers), W + jets and $t\bar{t}$ (shaded markers) and a mixture of W + jets, $t\bar{t}$ and events for the LM6 mSUGRA benchmark (open markers). In both plots, a jet threshold of 50 GeV is used; selection criteria on E_T^{miss} or M_T are given in the legend.

distribution for a combined W + jets and $t\bar{t}$ sample, without (black) and with (shaded and open) a LM6 mSUGRA signal. As in Fig. 8, a jet $|\vec{p}_T|$ threshold of 50 GeV is used and M_T is required to be greater than $M_W + 150$ GeV to suppress SM backgrounds. The integrated luminosity of the data sample is 1 fb^{-1}. Similarly to the search in Z + jets, the SM reach in R_{N_J} at high N_J can be constrained by using the sample with LM6 and requiring $M_T < 50$ GeV as shown in shaded markers. There is a large signal excess at $N_J \geq 4$, but the discriminating power of the search in R_{N_J} in the lepton + jets + E_T^{miss} signature for low mass mSUGRA models is limited by the residual $t\bar{t}$ contribution. The identification of new physics in R_{N_J} producing larger number of jets compared to low mass mSUGRA models could be possible.

The search in R_{N_J} is based on the distribution of tags in (pseudo-)rapidity in events from the same N_J bin. One can include additional information in the search from event yields in neighboring bins. At sufficiently high N_J additional jets are produced via higher order QCD processes so that the N_J distributions fall steeply in that regime. Selection criteria imposed on object $|\vec{p}_T|$ thresholds and E_T^{miss} can significantly modify the N_J spectra. However, a very general expectation is that the SM N_J yields fall approximately exponentially at high N_J, while new physics can modify it. We can use that expectation without relying heavily on the shape of the N_J spectrum.

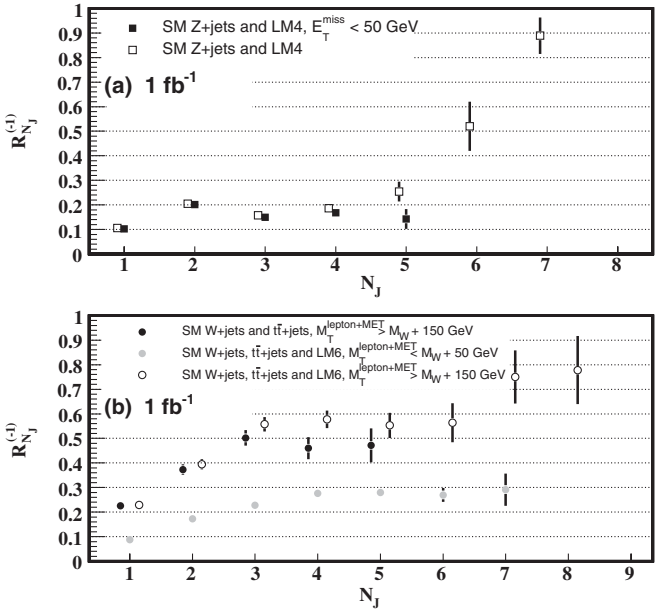

FIG. 11. Plot (a): $R_{N_J}^{(-1)}$ distributions for Z + jets (black markers) and a mixture of Z + jets and events for the LM4 mSUGRA benchmark (open markers). Plot (b): $R_{N_J}^{(-1)}$ distributions for W + jets and a mixture of W + jets, $t\bar{t}$ and events for the LM6 mSUGRA benchmark. In both plots, a jet threshold of 50 GeV is used; selection criteria on E_T^{miss} or M_T are given in the legend.

To that end, we consider another observable $R_{N_J}^{(-1)} \equiv Y_{N_J}^{\text{Central}}/(Y_{N_J-1}^{\text{Forward}} + Y_{N_J}^{\text{Central}})$, where Y_{N_J} is the event yield in the N_J bin. It is identical to R_{N_J} but in the denominator the forward yield in the $N_J - 1$ bin is used. Similarly, one can define $R_{N_J}^{(-2)}$, where the denominator includes the forward yield in the $N_J - 2$ bin. Figs. 11 and 12 show $R_{N_J}^{(-1)}$ and $R_{N_J}^{(-2)}$ for the $Z +$ jets and $W +$ jets samples using the previously described selection. The signal excess is clear and enhanced in the $Z +$ jets sample. For the $W +$ jets sample, the signal shape also has better separation from the background shape than in Fig. 10. These variables are less robust than R_{N_J}, but they have higher discriminating power against the background.

Using quantities like R_{N_J}, $R_{N_J}^{(-1)}$, or $R_{N_J}^{(-2)}$ could allow direct comparison across several signatures, those considered in this paper as well as others, such as, same-sign or opposite-sign di-leptons, jets, and E_T^{miss}. As such, they could be used to quickly perform a comprehensive search for new physics across multiple signatures in a few simple distributions. This search is most effective in early data when precise MC-based background predictions are unavailable. For example, in $Z +$ jets, with the integrated luminosity of about 200 pb^{-1}, the method can be used to predict backgrounds in the $N_J = 4$ bin, where new physics contributions can be significant. Searches in $R_{N_J}^{(-1)}$ and $R_{N_J}^{(-2)}$, though less robust, could be useful at even smaller integrated luminosities.

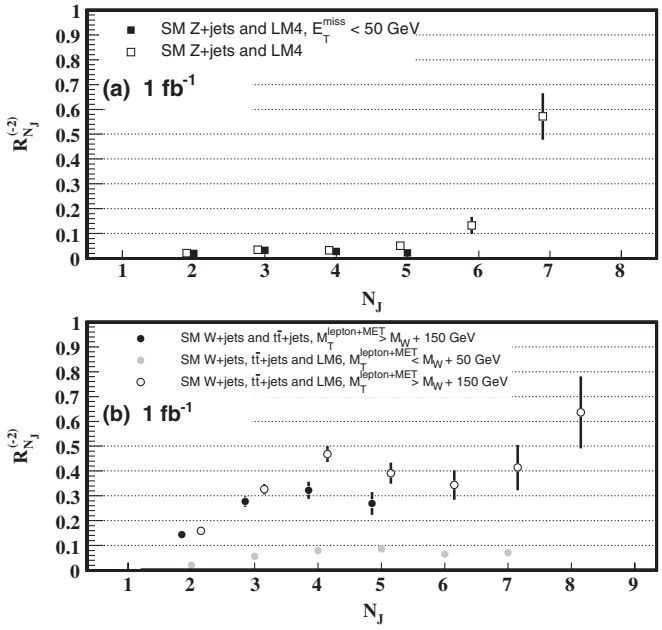

FIG. 12. Plot (a): $R_{N_J}^{(-2)}$ distributions for $Z +$ jets (black markers) and a mixture of $Z +$ jets and events for the LM4 mSUGRA benchmark (open markers). Plot (b): $R_{N_J}^{(-2)}$ distributions for $W +$ jets and a mixture of $W +$ jets, $t\bar{t}$ and events for the LM6 mSUGRA benchmark. In both plots, a jet threshold of 50 GeV is used; selection criteria on E_T^{miss} or M_T are given in the legend.

IX. SYSTEMATIC UNCERTAINTIES

The background estimation method discussed in this paper is not subject to the theoretical and experimental systematic uncertainties usually associated with MC simulation, since the background shapes and normalization are measured from data. Instead, systematic uncertainties come from the statistical precision for extrapolating event yields from large to small rapidity and from uncertainties in the validity of a linear extrapolation in R_{N_J}. There are several sources for an extrapolation bias.

SM processes in which jets are produced via a mechanism other than initial or final state radiation could bias the background prediction. The effect of $t\bar{t}$ discussed above is an extreme example. Di-boson production is another, e.g., WZ with a hadronic W boson decay peaks at $N_J \approx 2$ in the $Z +$ jets channel. The cross-sections for di-boson processes can be measured, but even if not, they are sufficiently small so that their contributions are negligible.

A linear extrapolation in R_{N_J} is valid only approximately. Large correlations between N_J and the rapidity dependence of the tag can lead to a bias. For example, for $N_J = 1$ in the $\gamma +$ jets sample, the $|\vec{p}_T|$ of the γ used for the rapidity tag is directly correlated with the $|\vec{p}_T|$ of the recoiling jet. The effect of correlations can be measured by varying the threshold and identification requirements for jets, leptons, photons, and E_T^{miss}. Lowering thresholds will suppress sensitivity to massive new particles and result in a wider N_J range that is background dominated. Such background samples could be used for systematic studies such as comparison of alternative, i.e., nonlinear parametrizations and different N_J fit ranges. Varying the η ranges used to define forward and central events would have similar utility.

The usage of different, *in situ* control samples is important to optimize and validate the final algorithm with data, and quantify its systematic biases. We expect that dominant systematic uncertainties will be associated with statistical uncertainties in such control samples.

X. CONCLUSION

We have presented a new method to predict SM backgrounds within the context of a search for new phenomena in final states with multiple jets: $Z +$ jets, $W +$ jets, $\gamma +$ jets, and multijets. The fraction of central events, measured in events with few jets, is used to extrapolate the backgrounds measured in the forward region into the central region for events with many jets. This fraction of central events is identified as a new discriminator between SM and heavy new particles and it could be useful in any new physics search at LHC.

The method performs well in robustness tests without and with a requirement on the presence of significant missing transverse energy. We have discussed systematic uncertainties associated with the method and procedures to estimate them. The usage of a ratio cancels many experimental uncertainties, and the data-driven procedure avoids theoretical uncertainties. This analysis could be performed without recourse to MC in early LHC data, when robustness against imperfections of background modeling and detector simulation can be a key to the discovery of new phenomena.

[1] Rapidity of a particle (or a jet) is defined as $y = \frac{1}{2}\ln(\frac{E+p_z}{E-p_z})$, where E and p_z are the particle's energy and the momentum component along the beam line. Pseudorapidity is $\eta = -\ln[\tan(\theta/2)]$, where θ is the particle's polar angle to the beam line.

[2] In the W + jets and pure multijet signatures, the SM $t\bar{t}$ background becomes dominant due to the large top quark mass. It is discussed separately in Sec. VII.

[3] For a recent review see J. M. Campbell, J. W. Huston, and W. J. Stirling, Rep. Prog. Phys. **70**, 89 (2007); C. Anastasiou, L. J. Dixon, K. Melnikov, and F. Petriello, Phys. Rev. D **69**, 094008 (2004).

[4] A. D. Martin, R. G. Roberts, W. J. Stirling, and R. S. Thorne, Eur. Phys. J. C **14**, 133 (2000).

[5] T. Aaltonen *et al.* (CDF Collaboration), Phys. Rev. D **76**, 072006 (2007).

[6] B. Abbott *et al.* (D0 Collaboration), Phys. Rev. Lett. **82**, 2457 (1999); A. Affolder *et al.* (CDF Collaboration), Phys. Rev. D **64**, 012001 (2001).

[7] Other definitions of tags can be made. For example, in Z + jets, W + jets, and γ + jets, the highest $|\vec{p}_T|$ jet could alternatively be used as a tag, which has an advantage due to the large jet η coverage.

[8] G. Aad *et al.* (ATLAS Collaboration), JINST **3**, S08003 (2008); S. Chatrchyan *et al.* (CMS Collaboration), JINST **3**, S08004 (2008).

[9] M. L. Mangano, M. Moretti, F. Piccinini, R. Pittau, and A. D. Polosa, J. High Energy Phys. 07 (2003) 001; S. Mrenna and P. Richardson, J. High Energy Phys. 05 (2004) 040.

[10] CTEQ5L, a default PDF set in ALPGEN, was used to produce mock data samples. In addition, we used CTEQ6L and MRST2001 in Z + jets to test sensitivity to PDF uncertainties as described in Sec. V.

[11] T. Sjöstrand, S. Mrenna, and P. Skands, J. High Energy Phys. 05 (2006) 026.

[12] J. Wess and B. Zumino, Nucl. Phys. **B70**, 39 (1974).

[13] A. H. Chamseddine, R. Arnowitt, and P. Nath, Phys. Rev. Lett. **49**, 970 (1982).

[14] The Minimal Supergravity is a restricted model of supersymmetry characterized by only five free parameters defined at the grand unification scale: m_0, $m_{1/2}$, A_0, $\tan\beta$, $\mathrm{sign}(\mu)$. For LM4 (LM6) benchmarks they are set to 210 (85) GeV, 285 (400) GeV, 0(0), 10(10), +(+). The total LM4 cross section is 25.1 pb (NLO), with $\sigma[Z(l^+l^-) + \text{jets})] \sim 0.6$ pb. The total LM6 cross-section is 4.9 pb (NLO), with $\sigma(l + \text{jets}) \sim 2.3$ pb. For more information see G. L. Bayatian *et al.* (CMS Collaboration), Report No. CMS TDR 8.2, CERN/LHCC 2006-021.

[15] Moreover, it is not necessary to apply any lepton efficiency corrections in our procedure as they will be accounted for in the fit to R_{N_J} as explained earlier. The example is given to illustrate the insensitivity of our method to uncertainties in the lepton reconstruction efficiencies.

[16] The small threshold for E_T^{miss} is used to retain sufficiently high yields to illustrate the performance of the method. In Sec. VIII, we present results with realistic tighter selection criteria on E_T^{miss} and jet $|\vec{p}_T|$.