Usage of machine learning for the separation of electroweak and strong $Z\gamma$ production at the LHC experiments

A M Petukhov and E Yu Soldatov
National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), Kashirskoe highway 31, Moscow, 115409, Russia
E-mail: aleksandr.petukhov@cern.ch, EYSoldatov@mephi.ru

Abstract. Separation of electroweak component from strong component of associated $Z\gamma$ production on hadron colliders is a very challenging task due to identical final states of such processes. The only difference is the origin of two leading jets in these two processes. Rectangular cuts on jet kinematic variables from ATLAS/CMS 8 TeV $Z\gamma$ experimental analyses were improved using machine learning techniques. New selection variables were also tested. The expected significance of separation for LHC experiments conditions at the second data-taking period (Run2) and 120 fb$^{-1}$ amount of data reaches more than 5$\sigma$. Future experimental observation of electroweak $Z\gamma$ production can also lead to the observation physics beyond Standard Model.

1. Introduction
After completion of the Standard Model (SM) in 2012 by the experimental observation of the Higgs boson in ATLAS and CMS experiments at the LHC [1, 2], the main focus of experimental high energy physics moved to the search of the manifestations of the physics beyond the Standard Model (BSM).

Direct search of the BSM physics basically implies the observation of the brand new resonances, which is constrained by the collision energy limitations on the modern accelerators. Indirect search implies the precision study of the rare SM processes and observation of any deviations, caused by the new interraction or the new resonance in the inaccessible energy region.

In this work we study such kind of rare process - the electroweak Z-boson production with association of the photon. The electroweak (EWK) $Z\gamma$ production consists of processes with fourth-order electroweak coupling. It contains vector boson scattering (VBS) as well as other non-VBS processes, e.g. initial- or final-state vector boson radiation. The VBS has high interference with other EWK-production processes and therefore can not be studied separately. Vector boson scattering is a $VV \rightarrow VV$ process with $V = Z/W/\gamma$, that allows to probe the vector boson self-couplings, namely triple and quartic gauge-boson couplings (TGCs and QGCs). The latter is of particular interest, because new phenomena could generate additional contributions to QGCs with respect to the Standard Model predictions [3, 4]. In the experiment, such process (see figure 1 (a)) can be deduced from events with $Z, \gamma$ and two or more jets in the final state.
Figure 1. Feynman diagrams for VBS (a) and QCD (b) production of $Z\gamma$.

Same final state could be produced by QCD mediated processes — called QCD-production in this paper — with second-order electroweak and second-order strong couplings. These processes include gluon mediated scattering as well as initial- and final-state radiated gluons. The most important diagram of this process is given in figure 1 (b). The problem is that QCD process cross section is approximately 2 order of magnitude higher than EWK, which makes an EWK component extraction really complicated.

The main difference between EWK and QCD is the kinematic parameters of the jets. In the VBS signal process, the two jets are produced from valence quarks and recoil against each other due to the emission of bosons. Jets are then well separated and bosons emitted in the centre of jet rapidity gap. This is not the case for the QCD process where the bosons are radiated from fermions lines.

The ATLAS and CMS experiments recently studied the electroweak production of $Z\gamma$ final state [5, 6], but were not able to observe the VBS process with significance larger than 3 standard deviations.

In this study, simulated data of proton-proton collisions at the Run2 LHC center-of-mass energy $\sqrt{s} = 13$ TeV was used to estimate the significance of EWK- and QCD-production processes separation in the $Z \rightarrow \nu\bar{\nu}$ channel. Results achieved with rectangular cuts were compared with results for Boosted Decision Tree (BDT) algorithms.

2. Simulated samples
The Monte Carlo event samples were generated with MadGraph5_AMC@NLO MC generator [7] interfaced to PYTHIA6 and PYTHIA8 [8] for parton showering and Delphes [9] for the detector simulation.

The EWK- and QCD-production samples were generated separately, at the Leading-order and the Next-to-leading-order accuracy respectively. The interference between these processes was previously found to be less than 10% [5], decreasing with the di-jet invariant mass, and was not taken into account.

The generated data was normalized to the integral luminosity of 36.1 fb$^{-1}$ corresponding to the amount of data collected by the ATLAS experiment in 2015-2016. The expected luminosity of whole ATLAS/CMS Run2 equal to 120 fb$^{-1}$ was also used for to estimate the significance.

3. Significance estimates
3.1. Event preselection
In this study, only events with one photon, at least two jets and high missing transverse momentum were considered. The latter is the signature of the Z-boson decaying to the neutrino-antineutrino pair.

Basing on the [5,6] and due to the fact, that single photon trigger threshold increased for both ATLAS and CMS in Run2, the initial experimental selections to suppress non-$Z\gamma$ backgrounds are:
• Electron and muon veto in the event;
• $p_T(\gamma) > 150$ GeV;
• $p_T^{\text{miss}} > 150$ GeV;
• $\Delta\varphi(\gamma, p_T^{\text{miss}}) > 2.1$.

3.2. Rectangular cuts
Basing on the ATLAS 8 TeV $Z\gamma$ analysis [5], to estimate the significance achieved with simple rectangular cuts the following kinematic variables were used:

- $m_{jj}$, di-jet invariant mass;
- $\Delta Y_{jj}$, absolute rapidity difference between two jets;
- $\zeta_{\gamma} = |y(\gamma) - (y(j_1)+y(j_2))/2|$, $\gamma$-centrality;
- $p_T^{\text{balance}} = \frac{|p_T^{\text{miss}}+p_T^{j_1}+p_T^{j_2}|}{|p_T^{\text{miss}}|+|p_T^{j_1}|+|p_T^{j_2}|}$.

Distributions of the EWK and QCD production processes for these variables are shown in figure 2.

![Figure 2](image)

**Figure 2.** Distributions of the EWK (Signal) and QCD (Background) $Z\gamma$ production components versus $m_{jj}$ (a), $\Delta Y_{jj}$ (b), $\zeta_{\gamma}$ (c) and $p_T^{\text{balance}}$ (d). The vertical green dashed line shows the value for the best rectangular cut applied.

The resulting significance is given in table 1. The main source of significance uncertainty comes from systematic uncertainty of the cross section calculation caused by the PDF uncertainty and the choice of the renormalization/factorization scales.
Table 1. Statistical significance \( (S/\sqrt{S+B}) \) for different separation algorithms, with normalization to 36.1 fb\(^{-1}\) of the integrated luminosity. \( S \) is the number of EWK process events, \( B \) is the number of the QCD process events.

| Algorithm                        | \( S/\sqrt{S+B} \) |
|----------------------------------|---------------------|
| Rectangular cuts                 | 2.65 ± 0.04         |
| 4 variables, AdaBoost            | 2.99 ± 0.04         |
| 4 variables, Gradient Boost      | 2.89 ± 0.04         |
| Additional variables, AdaBoost   | 3.08 ± 0.04         |
| Additional variables, Gradient Boost | 2.90 ± 0.04     |

3.3. Boosted decsicion trees
To increase the significance of signal separation the ROOT-based machine learning kit TMVA [10] was used. Namely, the Boosted Decision Trees classifier with two boost algorithms — AdaBoost and Gradient Boost. It was chosen because it allows for a better signal separation than simple linear classifiers while not requiring laborious tuning like Artificial Neural Network algorithms.

Two sets of input variables were used — the same as in the section 3.2 and with the following variables added:

- \( \eta_{j_1}, \eta_{j_2} \), where \( \eta \) is the pseudorapidity of the jet;
- \( \Delta R_{jj} \), where \( \Delta R = \sqrt{(\Delta \phi)^2 + (\Delta \eta)^2} \);
- \( \Delta R_{j\gamma} \);
- \( \Delta \rho_{jj} \);
- \( \sin |\Delta \phi(JJ)| \).

These variables were chosen among the rest within the framework of the conducted optimization study as the most powerful ones in discrimination of the processes under consideration.

The \( \gamma \)-centrality upper cut was found to increase the significance of the separation. The cause of it is that the jet rapidity difference in the denominator makes it possible for some events to have \( \zeta \), few orders of magnitude larger than the bulk of the measurements. This makes it harder for the BDT algorithm to make a proper cut on \( \zeta \gamma \) and in result it leads to less effective classifier. The cut providing the maximum significance was found to be \( \zeta \gamma < 1.5 \).

The figure 3 shows the classifier response distributions for signal and background samples (EWK and QCD-production respectively) for the extended set of variables. The figure 4 shows dependence of the significance on the value of the cut on the classifier response. The resulting significances are given in table 1.

The AdaBDT provides separation with higher significance, although its response seems to be less stable than the one of the Gradient Boost.

Significance was also estimated for \( L = 120 \) fb\(^{-1}\) - the expected integrated luminosity of the whole ATLAS/CMS Run2. The full results are given in table 2.
Figure 3. The classifier response distributions for signal and background (EWK and QCD-production respectively) samples for the extended set of variables: (a) AdaBoost (b) Gradient Boost.

Figure 4. The dependence of significance on the value of the cut on the classifier response for the extended set of variables: (a) AdaBoost (b) Gradient Boost.

Table 2. Statistical significance ($S/\sqrt{S+B}$) for different separation algorithms, with normalization to 120 fb$^{-1}$ of the integrated luminosity. S is the number of EWK process events, B is the number of the QCD process events.

| Algorithm                              | $S/\sqrt{S+B}$     |
|----------------------------------------|--------------------|
| Rectangular cuts                       | 4.84 ± 0.08        |
| 4 variables, AdaBoost                  | 5.45 ± 0.08        |
| 4 variables, Gradient Boost            | 5.27 ± 0.08        |
| Additional variables, AdaBoost         | 5.62 ± 0.08        |
| Additional variables, Gradient Boost   | 5.28 ± 0.08        |
4. Conclusions
This work demonstrates the real possibility to distinguish the EWK component of $Z\gamma$ production from the dominant QCD one. A set of discriminative variables was chosen after the optimization study and tests. Separation was done using simple rectangular cuts approach and using BDT algorithms AdaBoost and Gradient Boost. BDT application gives increase in significance of the order of 12-13%.

Maximal reached significance for 36.1 fb$^{-1}$ is $3.08 \pm 0.04 \sigma$ and for 120 fb$^{-1}$ is $5.62 \pm 0.08 \sigma$. Thus classifier developed in this study can lead to the experimental observation of the electroweak $Z\gamma$ production.

5. Acknowledgements
We gratefully acknowledge the financial support from Russian Science Foundation grant No.17-72-10021.

References
[1] Aad G et al. 2012 Science 338 1576-82
[2] Chatrchyan S et al. 2012 Science 338 1569-75
[3] Eboli O J P et al. 2004 Phys. Rev. D 69 095005
[4] Eboli O J P et al. 2006 Phys. Rev. D 74 073005
[5] Aad G et al. 2017 JHEP 07 107
[6] Chatrchyan S et al. 2017 Phys. Lett. B 770 380-402
[7] Alwall J et al. 2014 JHEP 07 079
[8] Sjöstrand T et al. 2008 Comp. Phys. Comm. 178 852-67
[9] de Favereau J et al. 2014 JHEP 02 057
[10] Hoecker A et al. 2007 arXiv:physics/0703039