An improved selection optimization method used for the measurement of ZZ production under conditions of ATLAS experiment during LHC Run II.

National Research Nuclear University «MEPhI»

Diana Pyatiizbyantseva, Evgeny Soldatov, Dmitriy Zubov

LXXI International Conference NUCLEUS

24.09.21

Introduction

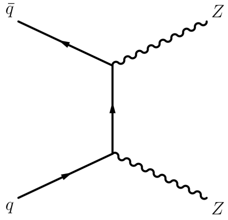
- At the LHC, a large number of particles are produced in p-p collisions.
- Precision measurements and rare process studies require a strict selection of events with intense background suppression.
- The selection of events in the LHC experiments, and in particular in the ATLAS experiment, can be divided into online and offline ones.
- There are various methods for optimizing event selection offline, and one of the simplest and most well-interpreted methods is an optimization using a rectangular cut.
- The work presents consistent and improved multivariate optimization methods and their comparison.

The signal and background

The final state of the $pp \to ZZ \to ll\nu\nu$ process consists of a pair of high-pT isolated leptons and significant missing transverse momentum (E_T^{miss}).

Initial selection:

- Two same flavour opposite-sign leptons (e+e- OR mu+mu-), leading lepton p_T>30 GeV, subleading lepton p_T>20 GeV
- Veto on any additional lepton
- $76 < M_{ll} < 106$ GeV



Main background sources:

- Z+jet
- WZ
- $t\bar{t}$

Monte Carlo event generator:

- MadGraph5_aMC@NLO

Showering and hadronization:

- Pythia8

Detector response simulation:

- Delphes3

Consistent optimization

The goals of optimization:

- search for variables sensitive to background suppression;
- calculation of the optimal threshold of the selected variables to suppress the background with the maximum possible signal retention.

Signal significance Z is a criterion for calculating the best signal-to-background ratio:

$$Z = \sqrt{2 \times [(S + B) \times \ln(1 + (S/B)) - S]}$$

The simplest optimization can be done by finding the threshold of the variable at which the maximum value of Z is reached.
Optimal values for each variable are searched one by one.

Variables for optimization

The variables are optimized in the following order:

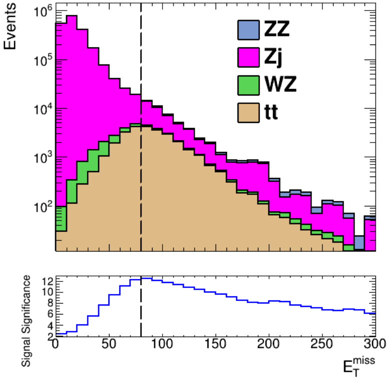
- E_T^{miss} — missing transverse momentum
- $\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$ — the azimuthal angle difference between the E_T^{miss} and dilepton system
- ΔR_{ll} — the angular distance between the two leptons
- $E_T^{miss}/\sqrt{H_T}$
- E_T^{miss}/H_T
- N_{b-jets} — number of jets containing b-hadrons.

H_T is the scalar sum of transverse momenta of selected leptons and jets.

Consistent optimization. E_T^{miss} & $\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$

Preselection:

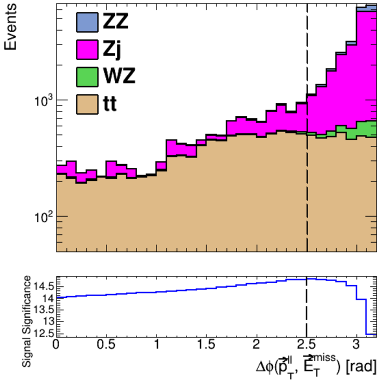
- Initial selection



Cut	E_T^{miss}>80 GeV
Signal signif.	2.45→12.5
Signal	3550→2790
Background	$2.09\cdot10^6$ →49126

Preselection:

- Initial selection + E_T^{miss}>80 GeV

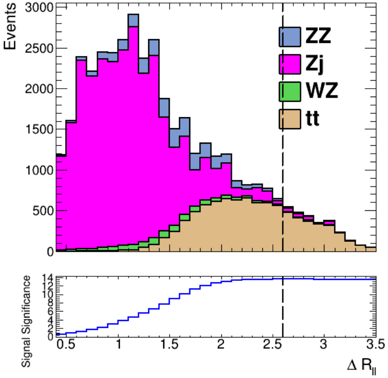


Cut	$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$>2.5
Signal signif.	12.5→13.5
Signal	2790→2636
Background	49136→37483

Consistent optimization. ΔR_{ll} & $E_T^{miss}/\sqrt{H_T}$

Preselection:

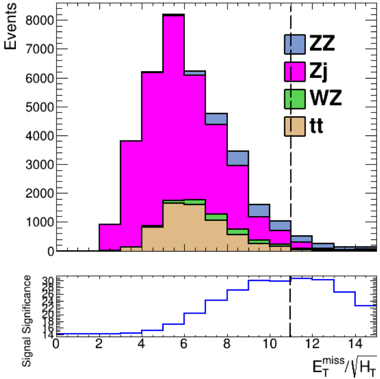
- Initial selection + E_T^{miss}>80 GeV + $\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$>2.5



Cut	ΔR_{ll}<2.6
Signal signif.	13.5→13.6
Signal	2636→2560
Background	37483→34854

Preselection:

- Initial selection + E_T^{miss}>80 GeV + $\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$>2.5 + ΔR_{ll}<2.6

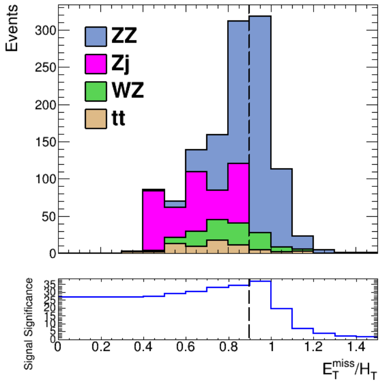


Cut	$E_T^{miss}/\sqrt{H_T}$>11
Signal signif.	13.6→27.1
Signal	2560→723
Background	34854→506

Consistent optimization. E_T^{miss}/H_T & N_{b-jets}

Preselection:

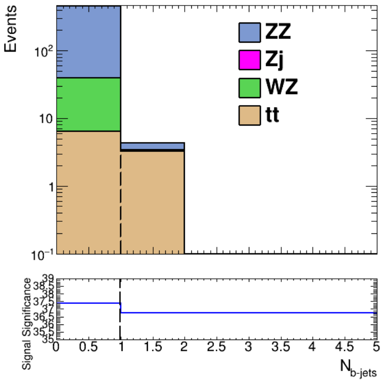
- Initial selection + E_T^{miss}>80 GeV + $\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$>2.5 + ΔR_{ll}<2.6 + $E_T^{miss}/\sqrt{H_T}$>11



Cut	E_T^{miss}/H_T>0.9
Signal signif.	27.1→36.8
Signal	723→418
Background	506→43

Preselection:

- Initial selection + E_T^{miss}>80 GeV + $\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$>2.5 + ΔR_{ll}<2.6 + $E_T^{miss}/\sqrt{H_T}$>11 + E_T^{miss}/H_T>0.9



Cut	N_{b-jets}<1
Signal signif.	36.8→37.4
Signal	418→417
Background	43→40

Consistent optimization. Results.

During the optimization process, a significant suppression of the background was achieved, the signal significance increased by an order of magnitude.

Disadvantages of the consistent method:

- The result depends on the sequence in which the variables are optimized
- Result depends on correlations between variables
- Method doesn't cover the full range of possible solutions

Variable	Before	After
$E_T^{miss}/\sqrt{H_T}$, $\sqrt{\text{GeV}}$	—	>11
E_T^{miss}, GeV	—	>80
ΔR_{ll}	—	<2.6
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$	—	>2.5
N_{b-jets}	—	<1
E_T^{miss}/H_T	—	>0.9

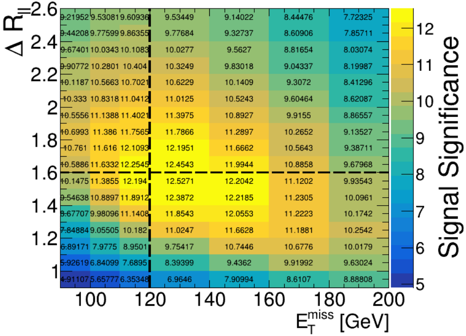
Signal signif.	2.45 ±0.12	37 ±2
S/B	$(1.697 \pm 0.011)\cdot 10^{-3}$	11 ±2
Signal		
ZZ	3550 ±16	417±5
Total signal	3550±16	417±5
Background		
Zj	$(2057\pm9)\cdot 10^3$	0±8
WZ	5859±35	33±2
tt	2880 ±215	6±3
Total bkg.	$(2092\pm9)\cdot 10^3$	40±9

An improved selection optimization method

- All described disadvantages can be avoided in a multivariate method (MV), considering the signal significance Z as a function of the thresholds on the variables.
- The MV optimization method finds the vector of thresholds on variables at which maximum Z is reached.
- MV method is implemented on the basis of the root package using the class of multidimensional histograms ThnSparse.
- This implementation makes it possible to check all possible combinations of selections, which guarantees a solution with the maximum Z.
- The MV method allows the application of additional conditions that satisfy the found solution.

An improved selection optimization method

- The signal significance is calculated for each bin of a multidimensional histogram containing signal and background events.
- Bining determines the accuracy and the phase space area in which the solution will be sought.
- The number of bins determines the time and memory consumption.



- MV method is difficult to visualize when the number of variables is greater than two. The figure illustrates the method in a two-dimensional case.

An improved selection optimization method. Results.

Initial selection:

- Two same flavour opposite-sign leptons (e+e- OR mu+mu-), leading lepton p_T>30 GeV, subleading lepton p_T>20 GeV
- Veto on any additional lepton
- $76 < M_{ll} < 106$ GeV

The method went through all the variations of the cuts for the ZZ process and found the best threshold vector.

Variable	Before	After
$E_T^{miss}/\sqrt{H_T}$, $\sqrt{\text{GeV}}$	—	>6
E_T^{miss}, GeV	—	—
ΔR_{ll}	—	<2.2
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$	—	>2.4
N_{b-jets}	—	<1
E_T^{miss}/H_T	—	>0.7

Signal signif.	2.45 ±0.12	45 ±3
S/B	$(1.697 \pm 0.011)\cdot 10^{-3}$	1.7 ±0.3
Signal		
ZZ	3550 ±16	1732±11
Total signal	3550±16	1732±11
Background		
Zj	$(2057\pm9)\cdot10^3$	640±160
WZ	5859±35	326±8
tt	2880 ±215	48±9
Total bkg.	$(2092\pm9)\cdot10^3$	1014±160

Consistent and multivariate methods. Comparison.

Variable	Before	Consistent	Multivariate
$E_T^{miss}/\sqrt{H_T}, \sqrt{\text{GeV}}$	—	>11	>6
E_T^{miss}, GeV	—	>80	—
ΔR_{ll}	—	<2.6	<2.2
$\Delta\phi(\vec{E}_T^{miss}, \vec{p}_T^{ll})$	—	>2.5	>2.4
N_{b-jets}	—	<1	<1
E_T^{miss}/H_T	—	>0.9	>0.7
Signal signif.	2.45 ±0.12	37 ±2	45 ±3
S/B	$(1.697 \pm 0.011)\cdot 10^{-3}$	11 ±2	1.7 ±0.3
Total signal	3550±16	417±5	1732±11
Total bkg.	$(2092\pm 9)\cdot 10^3$	40±9	1014±160

- The consistent optimization method leads to larger background suppression and a higher S/B ratio.
- The MV method results in more relaxed selections to be found with a higher signal significance value and a larger number of signal events.
- Taking into account all the factors, the MV method can be considered the best, since the main goal of the study is to find the maximum signal significance.

Conclusion

- The improved multivariate optimization method compared to the consistent method results in a higher signal significance and a larger number of signal events.
- The MV method has been described by optimizing the $ZZ \to ll\nu\nu$ process, but it can be used to optimize event selection for many other high energy physics processes.
- The MV optimization method can also be used for pre-optimization before using machine learning algorithms.

Backup

The ATLAS detector and coordinate system

Main components of the detector:

- Inner-detector
  - Pixel
  - SCT
  - TRT
- Calorimeter system
  - ECal
  - HCal
- Muon spectrometer

Cylindrical coordinates:

- ϕ — the azimuthal angle around the beam pipe
- θ — the polar angle
- $\eta = -\ln\tan\frac{\theta}{2}$ — the pseudorapidity
- $p_T = |\vec{p}|\sin\theta$ — the transverse momentum
- $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ — the angular distance