



Национальный исследовательский ядерный университет
«МИФИ»



Кафедра физики элементарных частиц №40

Научная исследовательская работа студента на тему:

Применение методов машинного обучения и феноменологические изыскания для разделения электрослабого и КХД процессов рождения Z -бозона с фотоном

Научные руководители
к.ф.-м.н., доцент
Солдатов Евгений Юрьевич

Петухов Александр Максимович

Работа
студента 4-ого курса
Савельева Константина
Михайловича
ИЯФит

г. Москва 2022

Введение

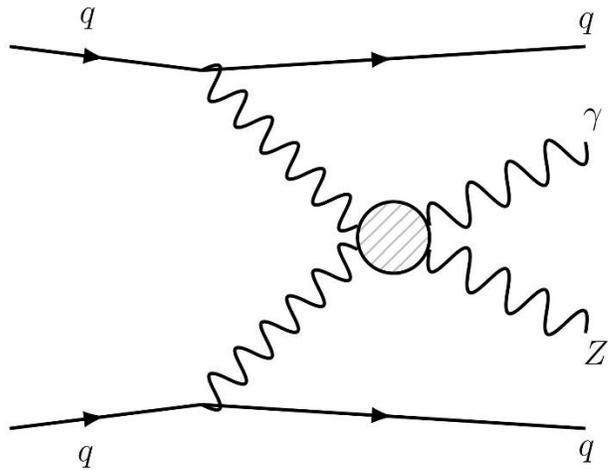


Диаграмма процесса рассеяния векторных бозонов

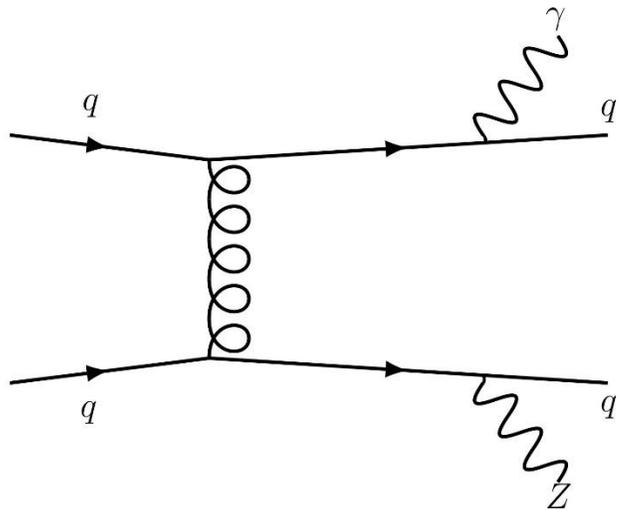


Диаграмма процесса КХД образования состояния $Z\gamma jj$

- Целью работы является изучение редкого процесса рассеяния векторных бозонов с рождением Z -бозона, фотона и двух адронных струй с последующим распадом Z -бозона на нейтрино и антинейтрино.
- Подобные процессы интересны с точки зрения поиска «новой физики» из-за их редкости и высокой чувствительности к отклонениям параметров от Стандартной модели.
- Выделение этого процесса является сложной задачей из-за основного фонового процесса – КХД образования идентичного конечного состояния, сечение которого на два порядка превосходит сечение изучаемого процесса.
- Для вычисления сечения исследуемого процесса с достаточной точностью необходимо эффективно отделять сигнальные события от фоновых. Классические одномерные фиксированные отборы не дают достаточной значимости, поэтому в работе исследовалось применение алгоритмов машинного обучения к разделению событий

Значимость $\sigma = \frac{S}{\sqrt{S+B}}$

Используемые данные

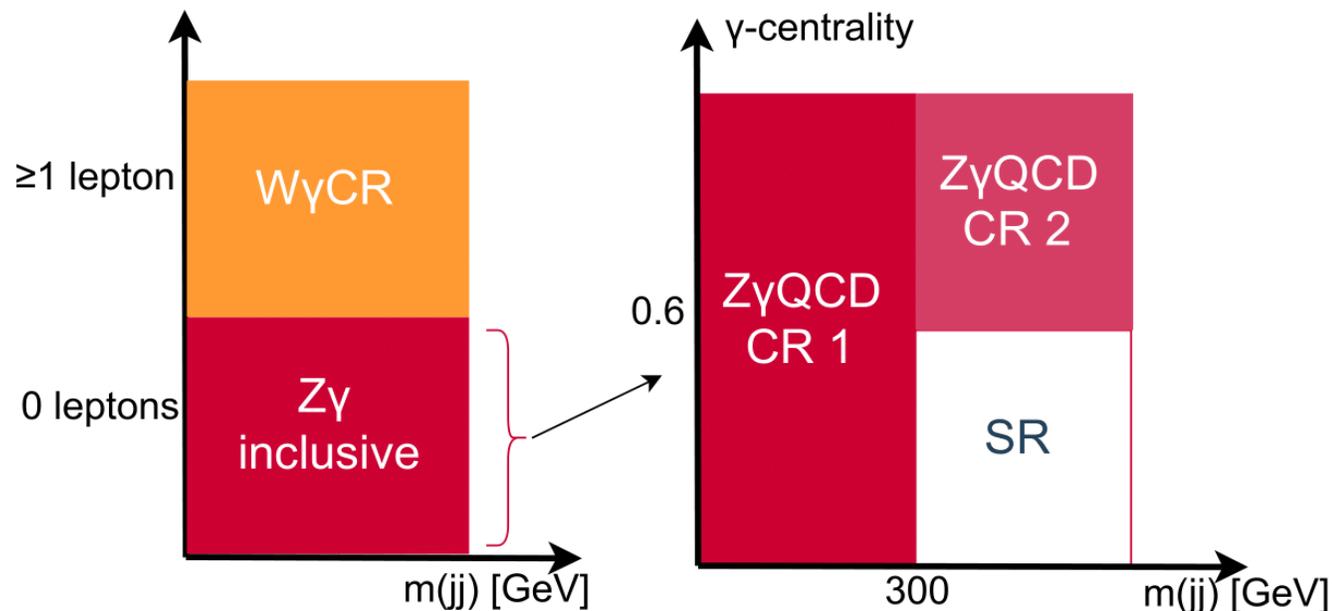
Работа производилась с Монте-Карло моделированными данными протон-протонных столкновений в детекторе *ATLAS* на БАК с энергией 13 ТэВ и интегральной светимостью 139 фб^{-1} и реальными данными, набранными в течение $2015\text{-}2018 \text{ гг.}$

Предотборы, применяемые в анализе:

$E_{\text{T}}^{\text{miss}}$	$> 120 \text{ ГэВ}$
E_{T}^{γ}	$> 150 \text{ ГэВ}$
Кол-во жёстких фотонов	$N_{\gamma} = 1$
Кол-во струй	$N_{\text{jets}} \geq 2$
Лептонное вето	$N_{\text{e}}, N_{\mu} = 0$

Отборы, применяемые в анализе:

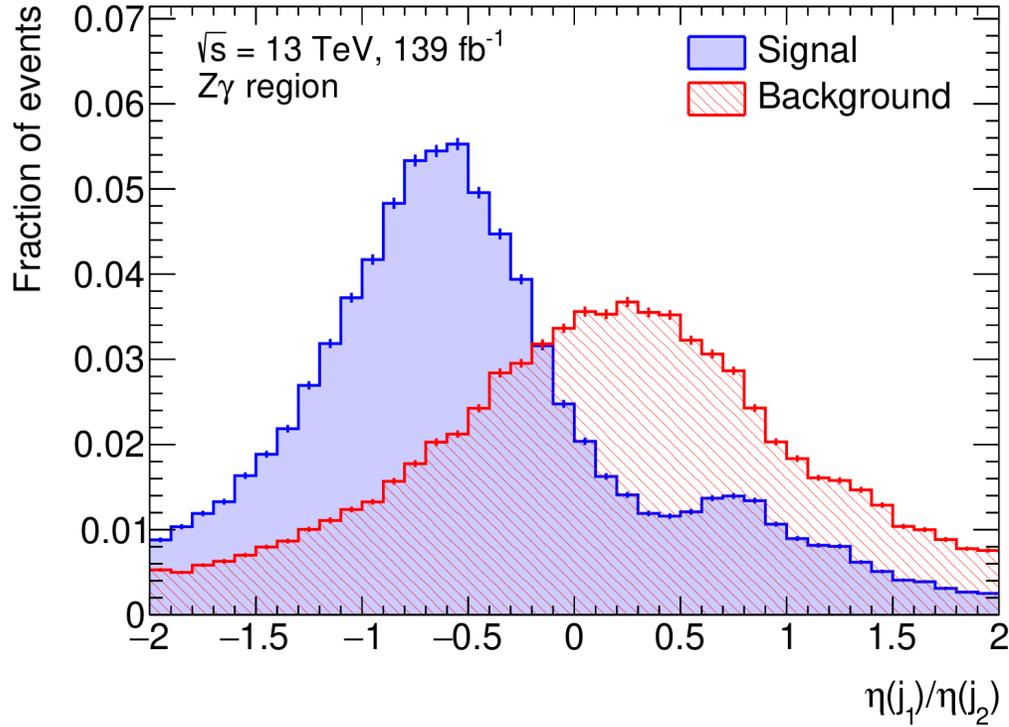
Значимость $E_{\text{T}}^{\text{miss}}$	> 12
$ \Delta\phi(\gamma, \vec{p}_{\text{T}}^{\text{miss}}) $	> 0.4
$ \Delta\phi(j_1, \vec{p}_{\text{T}}^{\text{miss}}) $	> 0.3
$ \Delta\phi(j_2, \vec{p}_{\text{T}}^{\text{miss}}) $	> 0.3
$p_{\text{T}}^{\text{SoftTerm}}$	$< 16 \text{ ГэВ}$



В качестве модели классификатора используется композиция деревьев решений с использованием техники градиентного бустинга (*Boosted Decision Trees, BDT*) из библиотеки машинного обучения *LightGBM*.

Обучение классификаторов производилось в $Z\gamma$ -регионе, оценка значимости производилась в сигнальном регионе.

Отбор переменных

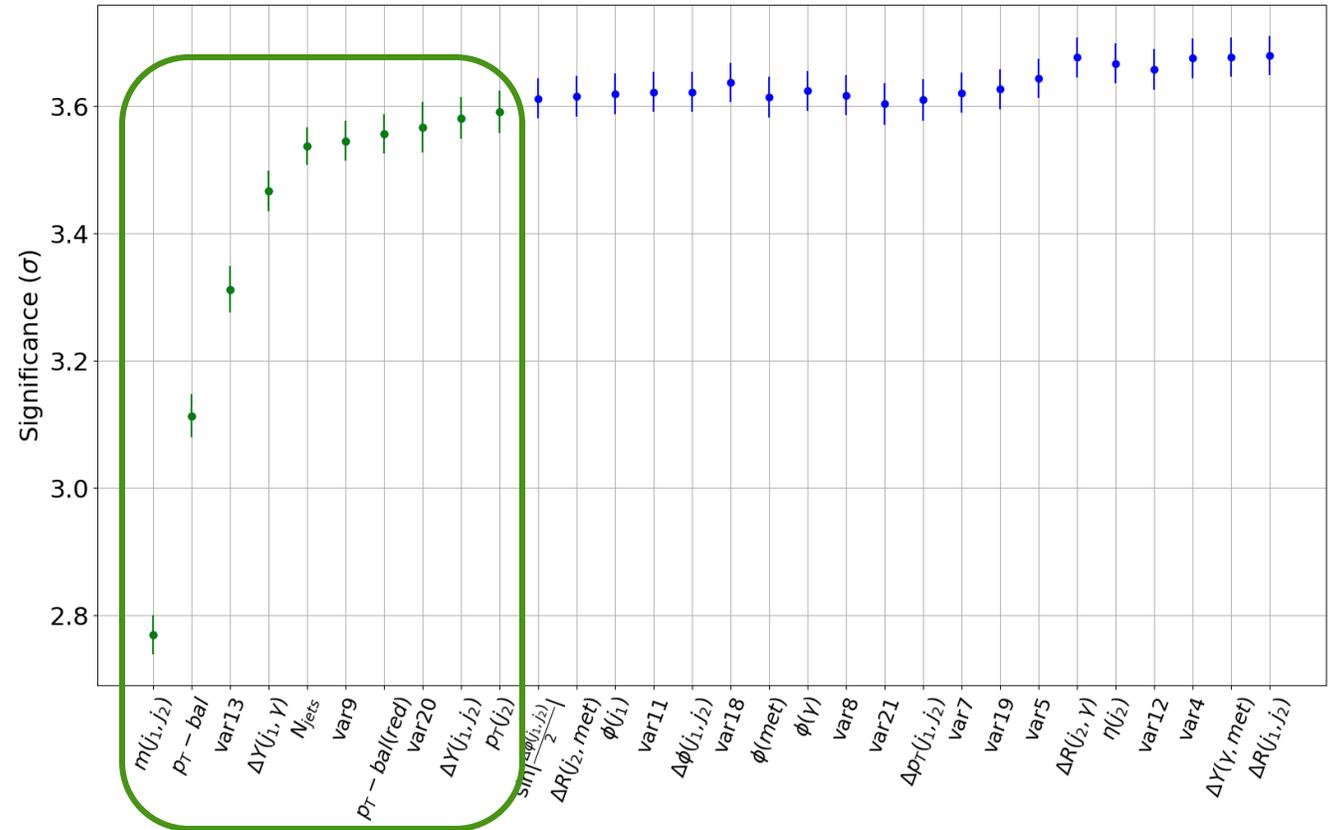


В дополнение к имеющемуся набору были составлены порядка 20 дополнительных переменных.

К примеру, в набор была добавлена переменная отношения псевдобыстрот двух лидирующих струй.

Для сигнала характерны вылеты струй в разные стороны. Также сигнал имеет меньшую дисперсию распределения.

На основании отбора методом "N+1" было отобрано 10 переменных:



1. m_{jj}

2. $p_T - \text{balance} = \frac{|\vec{p}_T^{\text{miss}} + \vec{p}_T^\gamma + \vec{p}_T^{j_1} + \vec{p}_T^{j_2}|}{E_T^{\text{miss}} + E_T^\gamma + p_T^{j_1} + p_T^{j_2}}$

3. $\text{var13} = \sqrt{p_T(j_1)^2 + p_T(j_2)^2}$

4. $\Delta Y(j_1, \gamma)$

5. N_{jets}

6. $\text{var9} = \frac{E_T^{\text{miss}}}{p_T(j_1) + p_T(j_2)}$

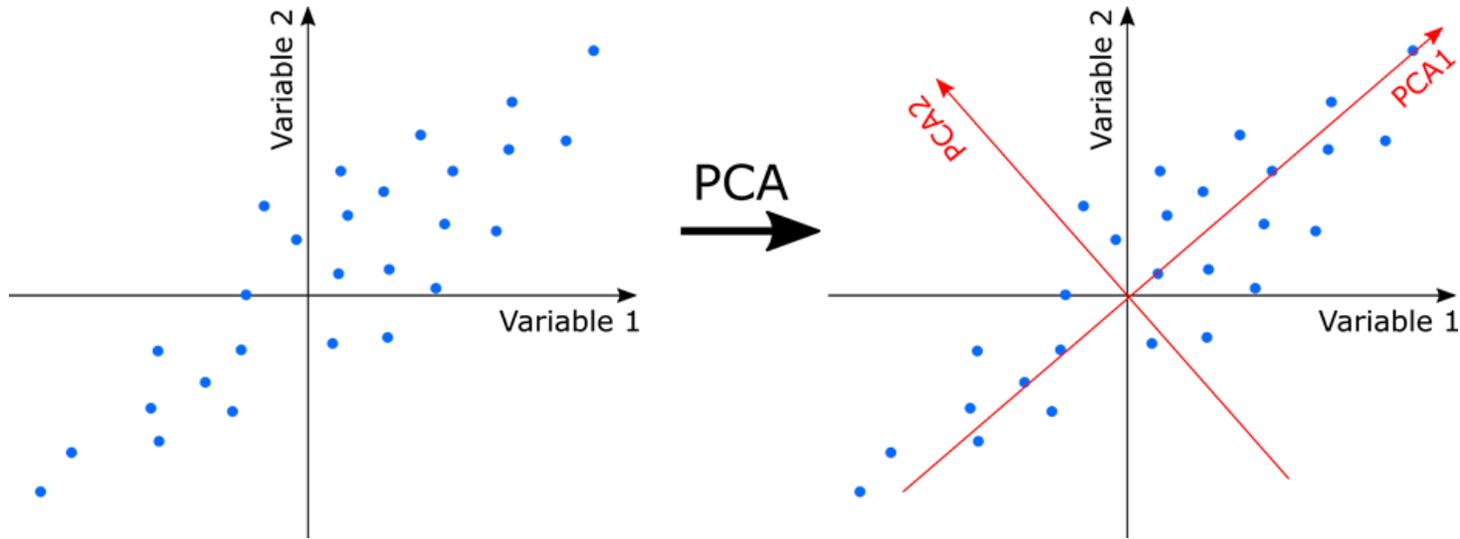
7. $p_T - \text{balance}(\text{red}) = \frac{|\vec{p}_T^\gamma + \vec{p}_T^{j_1} + \vec{p}_T^{j_2}|}{E_T^\gamma + p_T^{j_1} + p_T^{j_2}}$

8. $\text{var20} = \frac{E_T^{\text{miss}} - p_T^\gamma}{p_T(j_1) + p_T(j_2)}$

9. $\Delta Y(j_1, j_2)$

10. $p_T(j_2)$

Метод главных компонент

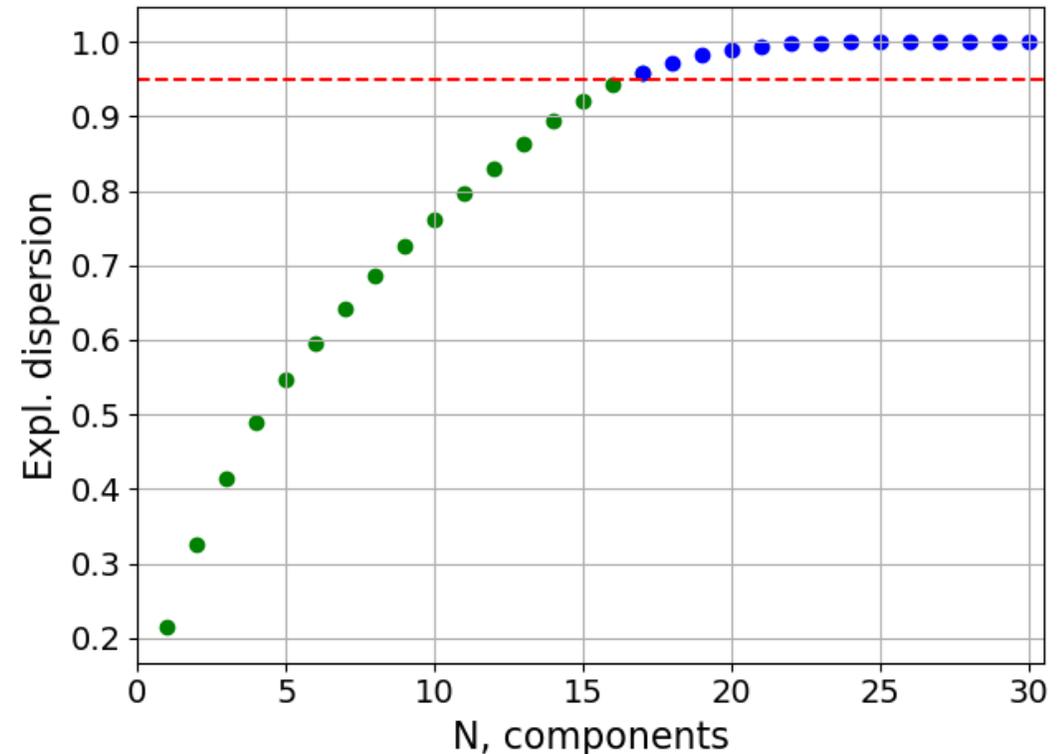


В качестве альтернативы методу "N+1" использовался метод главных компонент.

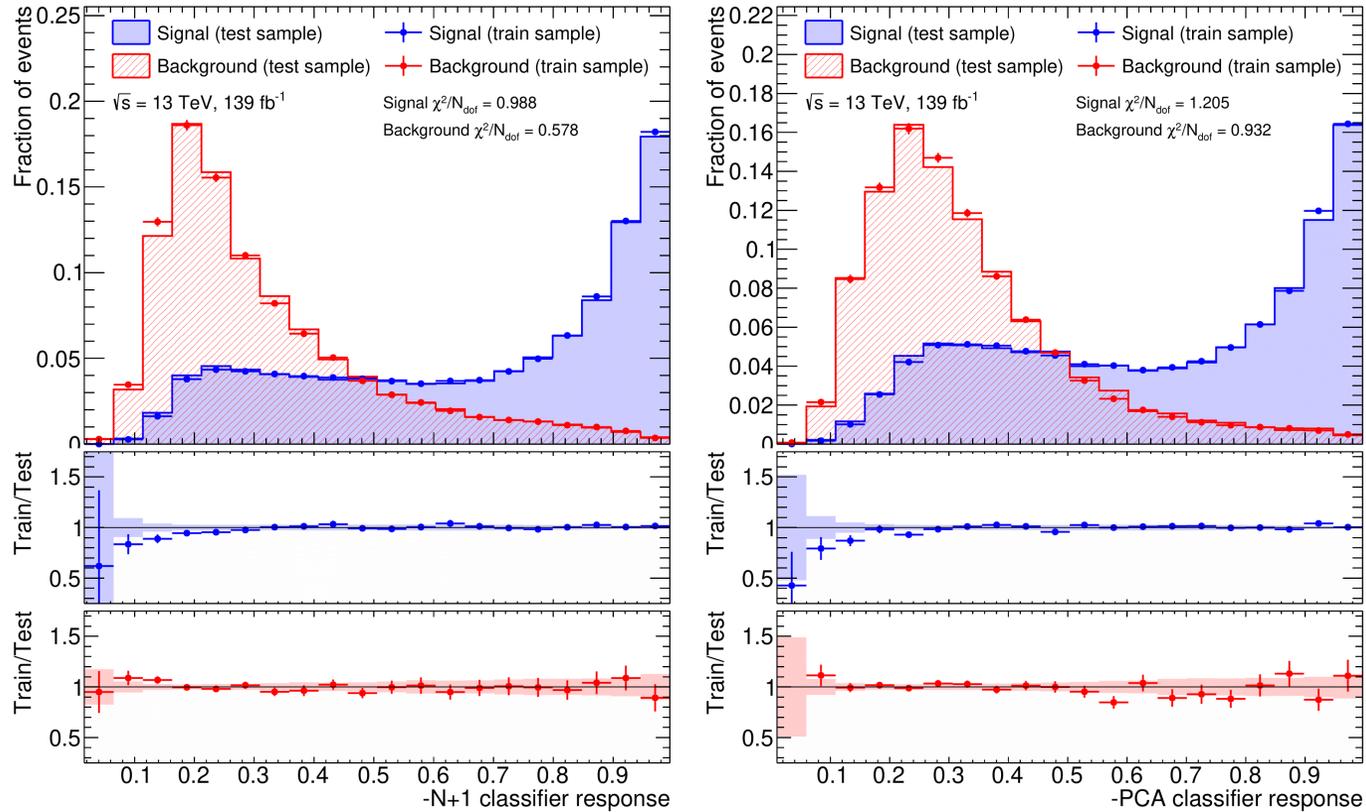
Его смысл заключается в поиске подпространства меньшей размерности, чем исходное пространство признаков с наибольшим значением дисперсии вдоль осей. То есть необходимо найти такие линейные преобразования признаков чтобы количество потерянной информации было наименьшим.

Для определения необходимого количества компонент используется график объяснённой дисперсии.

С помощью метода главных компонент, применённого к первым 30 переменным, которые были отобраны методом "N+1", было найдено 16 компонент, обеспечивающих 95% объяснённой дисперсии.

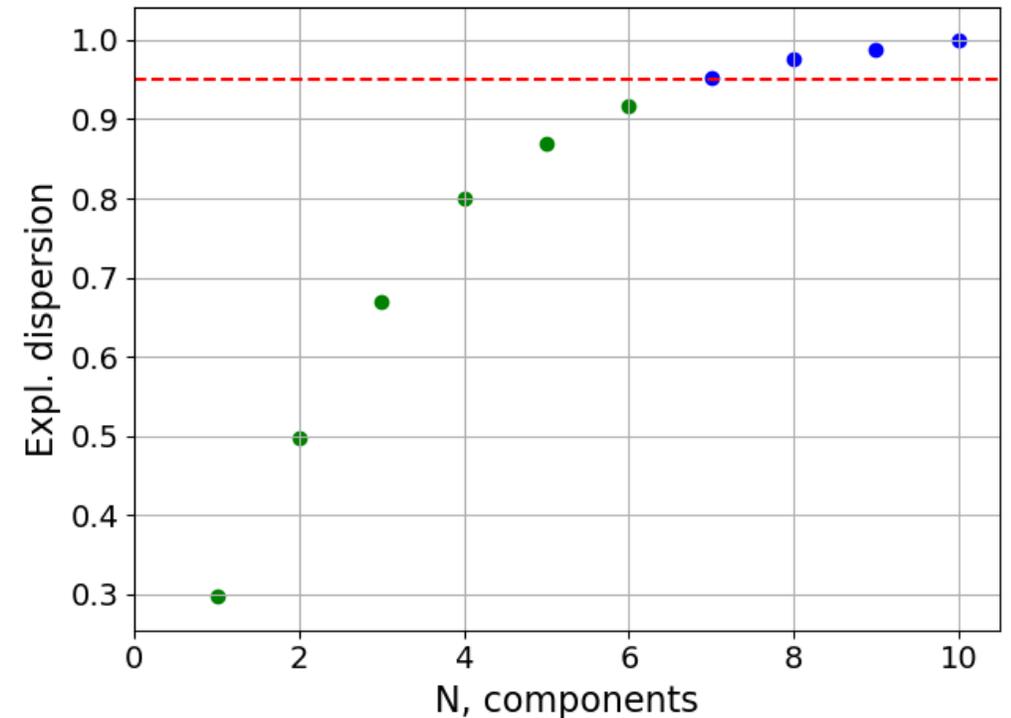


Настройка классификаторов и результаты



Использование метода главных компонент значительно ухудшает значение значимости.

Была произведена проверка возможности применения метода главных компонент к набору переменных, полученному с помощью метода "N+1".



	Вход. сигнал	Вход. фон	Кол-во сигнала	Кол-во фона	Значимость, σ
До отборов	90035	86902	46.7 ± 0.2	304.9 ± 4.6	2.49 ± 0.02
N+1	51734	9563	26.9 ± 0.1	28.6 ± 0.9	3.61 ± 0.03
PCA	53615	11513	27.9 ± 0.1	37.8 ± 1.4	3.44 ± 0.04
PCA & N+1	50874	10411	26.4 ± 0.1	31.7 ± 1.1	3.47 ± 0.03

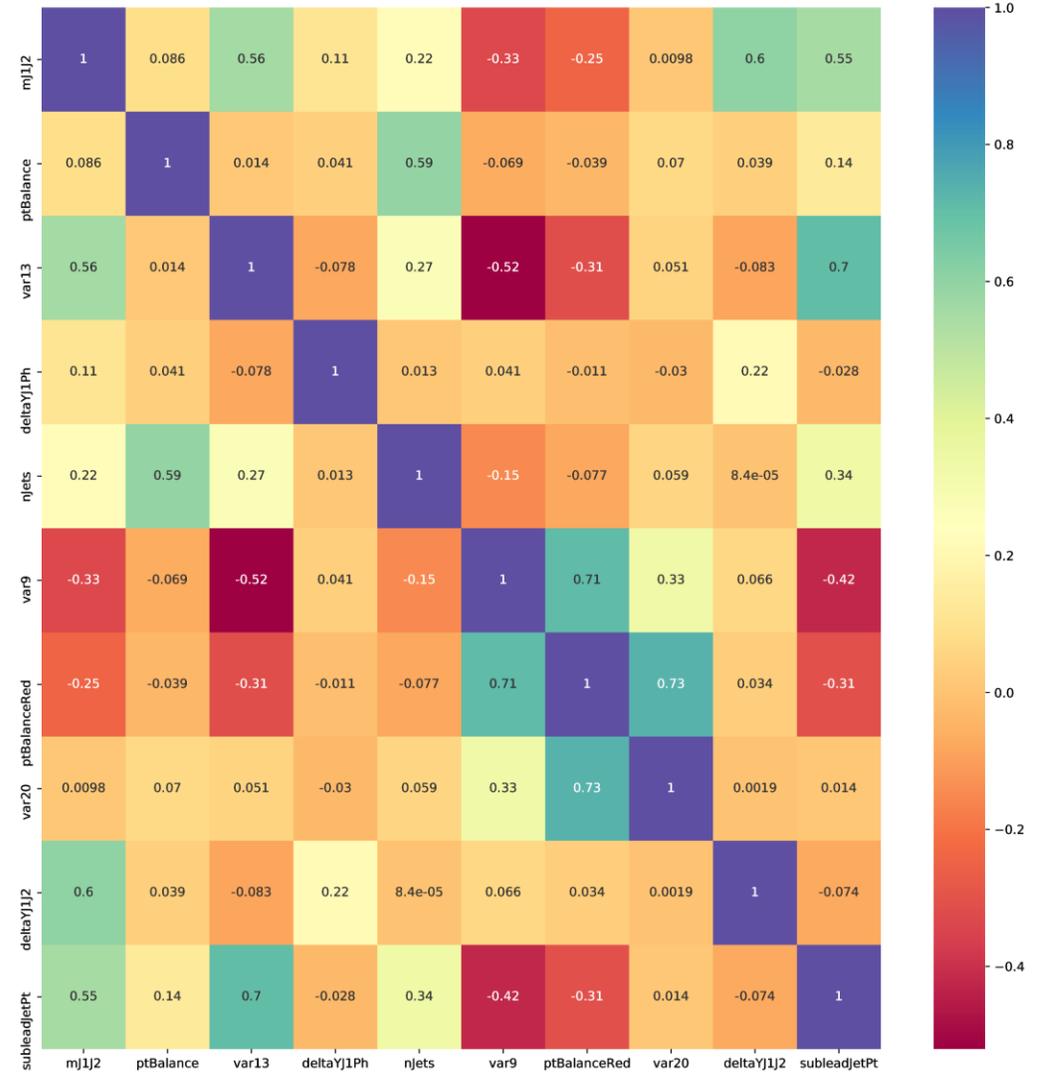
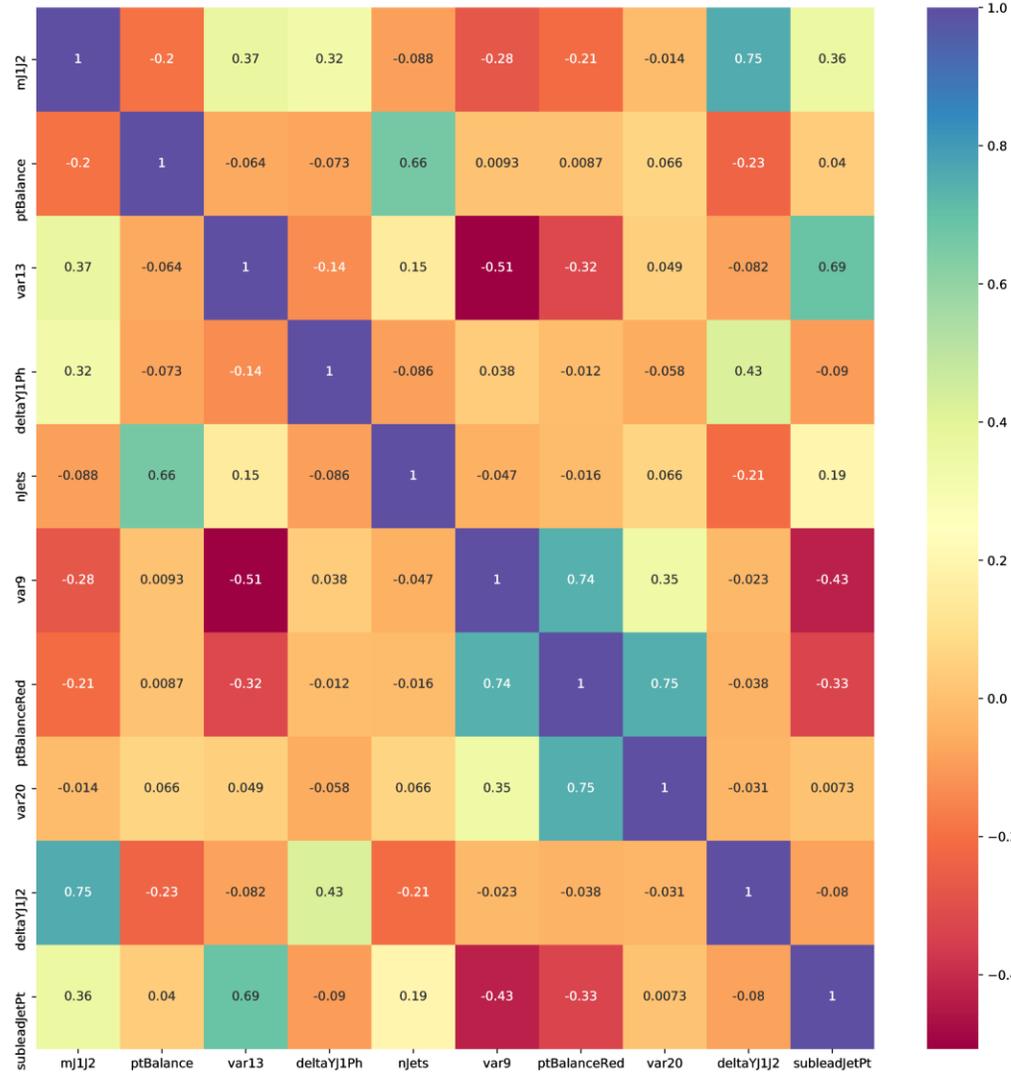
Заключение

- Составлены новые переменные для обучения классификатора, из них отобраны лучшие, в числе которых, например, среднеквадратичное значение поперечных импульсов струй, хорошо показавшее себя при классификации;
- Был получен прирост значимости при применении классификатора с использованием переменных, отобранных методом "N+1" с значения значимости $(2.49 \pm 0.02)\sigma$ до $(3.61 \pm 0.03)\sigma$;
- Была осуществлена проверка метода главных компонент, позволившего получить значение значимости $(3.47 \pm 0.03)\sigma$, что показывает неоправданность использования данного метода;

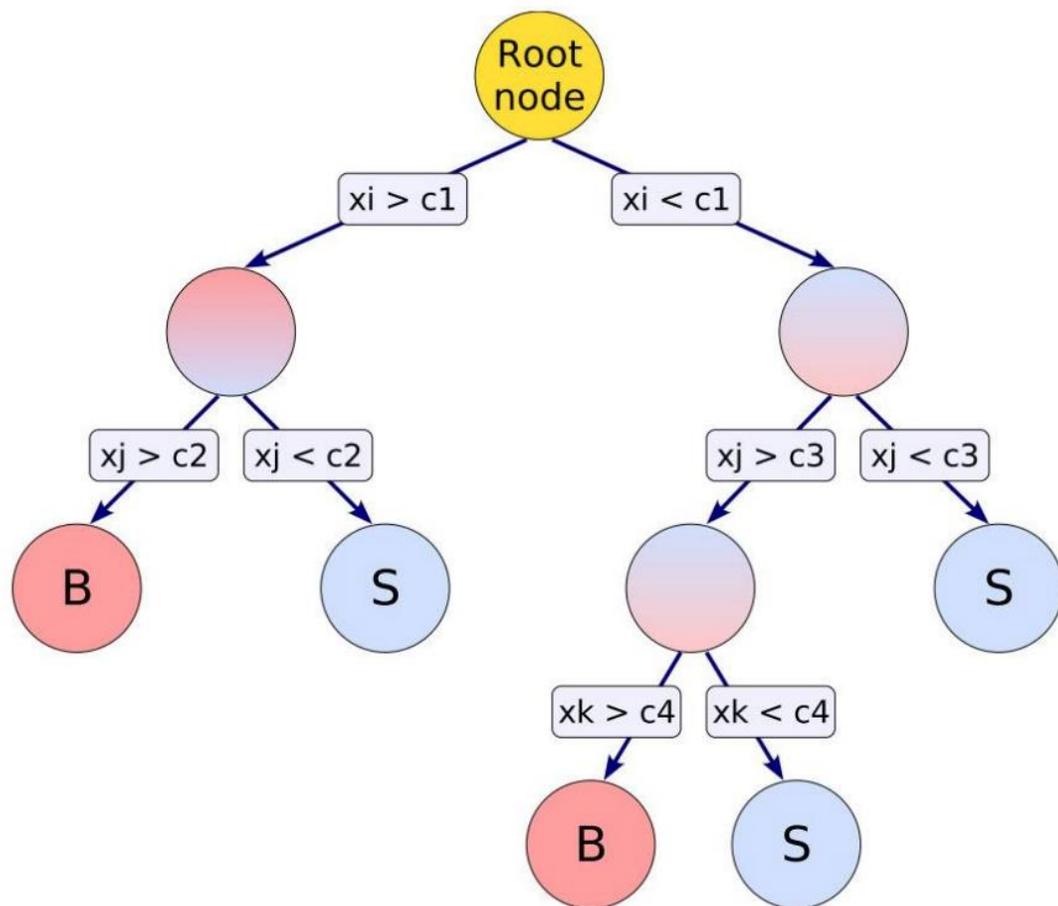
В дальнейшем планируется использовать дополнительные параметры модели классификатора при настройке, использовать другие метрики при оптимизации настроек классификатора. Также планируется объединить текущие результаты с результатами по использованию информации о третьей по поперечному импульсу струе в событии.

Спасибо за внимание!

Back-up



Композиция деревьев решений



Композиция деревьев решения, созданная с помощью градиентного бустинга (Boosted Decision Trees, BDT) – это классификатор с бинарной древовидной структурой, позволяющий разбивать фазовое пространство на множество областей.

Основные характеристики метода:

- небольшое время обучения
- не склонен к переобучению из-за
- не требует подготовки переменных

$$1. \frac{\eta_{j_1}}{\eta_{j_2}} \frac{1-2\eta_\gamma}{\eta_{j_1}+\eta_{j_2}}$$

$$2. \left| \frac{\eta_{j_1}-\eta_{j_2}}{\eta_{j_1}+\eta_{j_2}} \right|$$

$$3. \left| \frac{\eta_{j_1}+\eta_{j_2}+\eta_\gamma}{\eta_{j_1}-\eta_{j_2}} \right|$$

$$4. \left| \frac{\eta_\gamma}{\sqrt{|\eta_{j_1}\eta_{j_2}|}} \right|$$

$$5. \left| \frac{\eta_\gamma - \sqrt{|\eta_{j_1}\eta_{j_2}|}}{\eta_{j_1}-\eta_{j_2}} \right|$$

$$6. \frac{E_T^{\text{miss}}}{P_T^\gamma}$$

$$7. \frac{P_T^\gamma}{P_T^{j_1}+P_T^{j_2}}$$

$$8. \frac{P_T^\gamma}{\sqrt{P_T^{j_1}P_T^{j_2}}}$$

$$9. \frac{E_T^{\text{miss}}}{P_T^{j_1}+P_T^{j_2}}$$

$$10. \frac{E_T^{\text{miss}}}{\sqrt{P_T^{j_1}P_T^{j_2}}}$$

$$11. \frac{P_T^{j_1}-P_T^{j_2}}{P_T^{j_1}+P_T^{j_2}}$$

$$12. \frac{P_T^\gamma}{\sqrt{(P_T^{j_1})^2+(P_T^{j_2})^2}}$$

$$13. \sqrt{(P_T^{j_1})^2+(P_T^{j_2})^2}$$

$$14. \left| \frac{(\eta_\gamma)^2 - \left(\frac{\eta_{j_1}+\eta_{j_2}}{2}\right)^2}{(\eta_{j_1}-\eta_{j_2})^2} \right|$$

$$15. \frac{E_T^{\text{miss}}+P_T^\gamma}{P_T^{j_1}+P_T^{j_2}}$$

$$16. \frac{\eta_\gamma}{\eta_{j_1}+\eta_{j_2}}$$

$$17. \left| \frac{E_T^{\text{miss}}-P_T^\gamma}{P_T^{j_1}-P_T^{j_2}} \right|$$

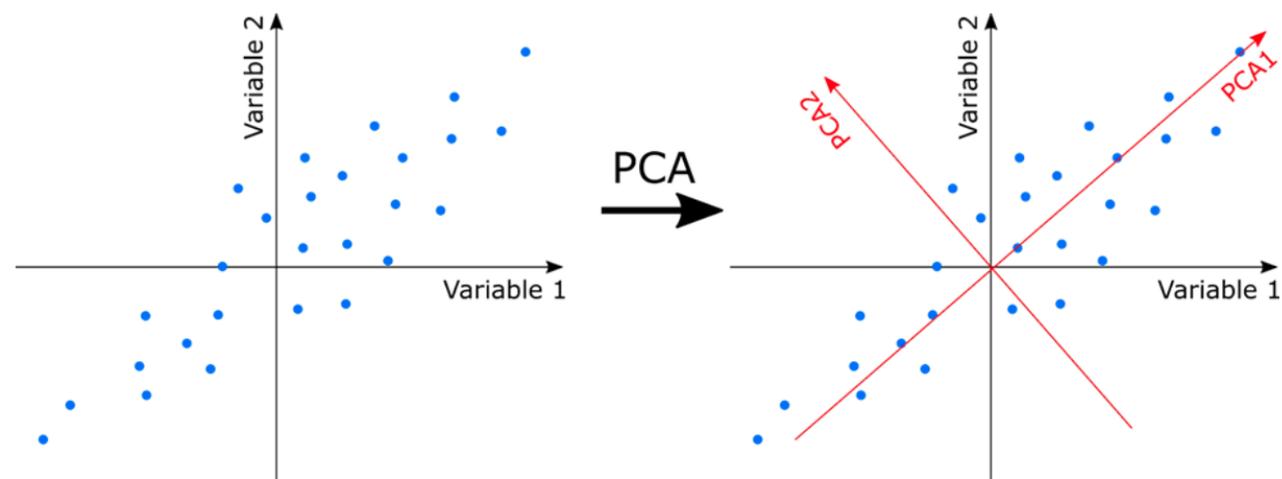
$$18. \frac{P_T^{j_2}}{P_T^{j_1}}$$

$$19. \frac{E_T^{\text{miss}}}{\sqrt{(P_T^{j_1})^2+(P_T^{j_2})^2}}$$

$$20. \frac{E_T^{\text{miss}}-P_T^\gamma}{P_T^{j_1}+P_T^{j_2}}$$

$$21. \frac{E_T^{\text{miss}}+P_T^\gamma}{P_T^{j_1}-P_T^{j_2}}$$

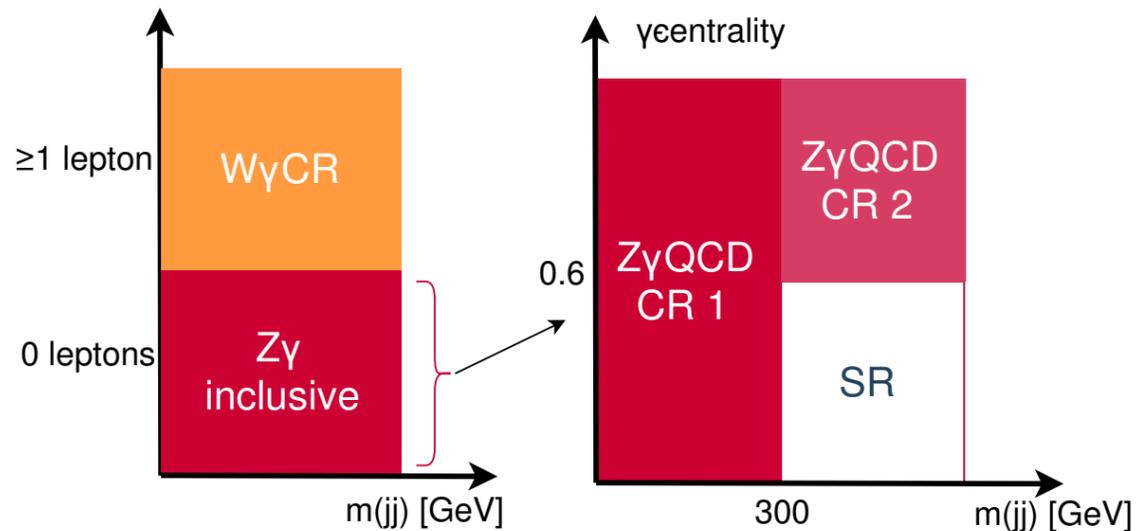
$$22. \frac{\eta_{j_1}}{\eta_{j_2}}$$



$$Z = X \begin{pmatrix} | & | & \dots & | \\ \omega_1 & \omega_2 & \dots & \omega_k \\ | & | & \dots & | \end{pmatrix}$$

$$d_i = \frac{\lambda_i}{\sum \lambda_i}$$

$W\gamma$ control region	
N_{leptons}	≥ 1
$Z\gamma jj$ QCD control region 1	
N_{leptons}	$= 0$
m_{jj}	$< 300 \text{ GeV}$
$Z\gamma jj$ QCD control region 2	
N_{leptons}	$= 0$
m_{jj}	$> 300 \text{ GeV}$
γ -centrality	> 0.6
$Z\gamma jj$ EWK signal region	
N_{leptons}	$= 0$
m_{jj}	$> 300 \text{ GeV}$
γ -centrality	< 0.6



Selections	Cut Value
E_T^{miss}	$> 120 \text{ GeV}$
E_T^γ	$> 150 \text{ GeV}$
Number of tight isolated photons	$N_\gamma = 1$
Number of jets	$N_{\text{jets}} \geq 2$
Lepton veto	$N_e = 0, N_\mu = 0$
E_T^{miss} significance	> 12
$ \Delta\phi(\gamma, \vec{p}_T^{\text{miss}}) $	> 0.4
$ \Delta\phi(j_1, \vec{p}_T^{\text{miss}}) $	> 0.3
$ \Delta\phi(j_2, \vec{p}_T^{\text{miss}}) $	> 0.3
p_T^{SoftTerm}	$< 16 \text{ GeV}$

