Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский ядерный университет МИФИ»
(НИЯУ МИФИ)

УДК 539.120.71

ОТЧЁТ
О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

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

Научный руководитель
к.ф.-м.н., доцент _____ Е. Ю. Солдатов

Научный руководитель _____ А. М. Петухов

Студент _____ К. М. Савельев

Москва 2022

Содержание

Введение

Стандартная модель (СМ) даёт достаточно точные качественные и количественные предсказания свойств элементарных частиц и их взаимодействий. Однако, существует ряд явлений, которые не могут быть описаны в рамках Стандартной модели. Одни из самых известных: явление осцилляций нейтрино, из которого следует наличие у него массы, что противоречит СМ; тёмная материя, косвенные признаки наличия которой наблюдаются в большом масштабе в астрономических наблюдениях; факт барионной асимметрии. Всё это свидетельствует о том, что современная СМ не является всеобъемлющей. Это даёт стимул к поискам отклонений, которые могут привести к открытию более совершенной модели для описания взаимодействий элементарных частиц.

Целью исследования является поиск отклонений от предсказаний СМ при рассеянии векторных бозонов, $VV \to VV$, где $V = W/Z/\gamma$. Для исследования был выбран высокочувствительный к отклонениям СМ процесс электрослабого рождения Z-бозона, фотона совместно с двумя адронными струями ($Z\gamma jj$) с последующим распадом Z-бозона на нейтрино и антинейтрино.

Выбор нейтрального канала распада связан с его достаточно большой вероятностью (20%) [1] и возможностью отделения сигнала в отличии от распада по адронному каналу, вероятность которого составляет около 70%. Лептонный канал распада не рассматривался из-за его сравнительно низкой вероятности ($\sim 6.7\%$).

Этот процесс невозможно отделить от других электрослабых процессов с тем же конечным состоянием. Поэтому его изучение возможно только посредством рассмотрения всех процессов электрослабого образования конечного состояния $Z\gamma jj$. Они включают в себя как процессы рассеяния векторных бозонов, чувствительных к изменениям параметров Стандартной модели, так и прочие электрослабые процессы. Пример диаграммы процесса рассеяния представлен на рисунке 1a. Кинематические параметры частиц в конечном состоянии позволяют отделить их от КХД процессов

с тем же конечным состоянием, которые являются основным фоном наряду с экспериментальными фонами и обладают сечением почти в сотню раз превышающим фон изучаемого процесса. Пример диаграммы электрослабого процесса, не являющегося рассеянием, представлен на рисунке 1в, а пример диаграммы КХД процесса представлен на рисунке 1б.

Применение одномерных фиксированных отборов при отделении сигнальных событий от фона при изучении процессов электрослабого рождения Z-бозона, фотона и двух адронных струй не позволяет достаточно хорошо отделить сигнал от фона, из-за чего точность измерения сечения получается низкой. Поэтому проводилось исследование использования алгоритма машинного обучения «Композиция деревьев решений» (Boosted Decision Trees, BDT) [2]. Были изучены подходы к подбору наилучшего набора переменных и были установлены оптимальные настройки моделей машинного обучения. Также была проведена проверка соответствия данных, сгенерированных с помощью метода Монте-Карло (МК), реальным данным.

Эффективное разделение сигнальных и фоновых событий даст возможность определить величину сечения исследуемого процесса с большей точностью.

В предыдущем исследовании, опубликованном коллаборацией ATLAS, использовались данные столкновений с энергией в системе центра масс 8 ТэВ и интегральной светимостью 20.3 фб$^{-1}$, из-за недостаточной чувствительности оно было выполнено только для поиска аномальных вершин [3]. Это исследование нацелено на увеличение значимости измерения сечения процесса. Достижение порога значимости в 5σ позволит подтвердить наблюдение исследуемого процесса. На данный момент опубликовано сечение процесса $Z(\ell\ell)\gamma jj$ со значимостью 4.7σ [4]. Результаты измерения параметров этого процесса могут использоваться для поиска аномалий в вершине $WWZ\gamma$, а также вершин $ZZZ\gamma$, $ZZ\gamma\gamma$ и $Z\gamma\gamma\gamma$ [5–7], запрещённых в СМ.

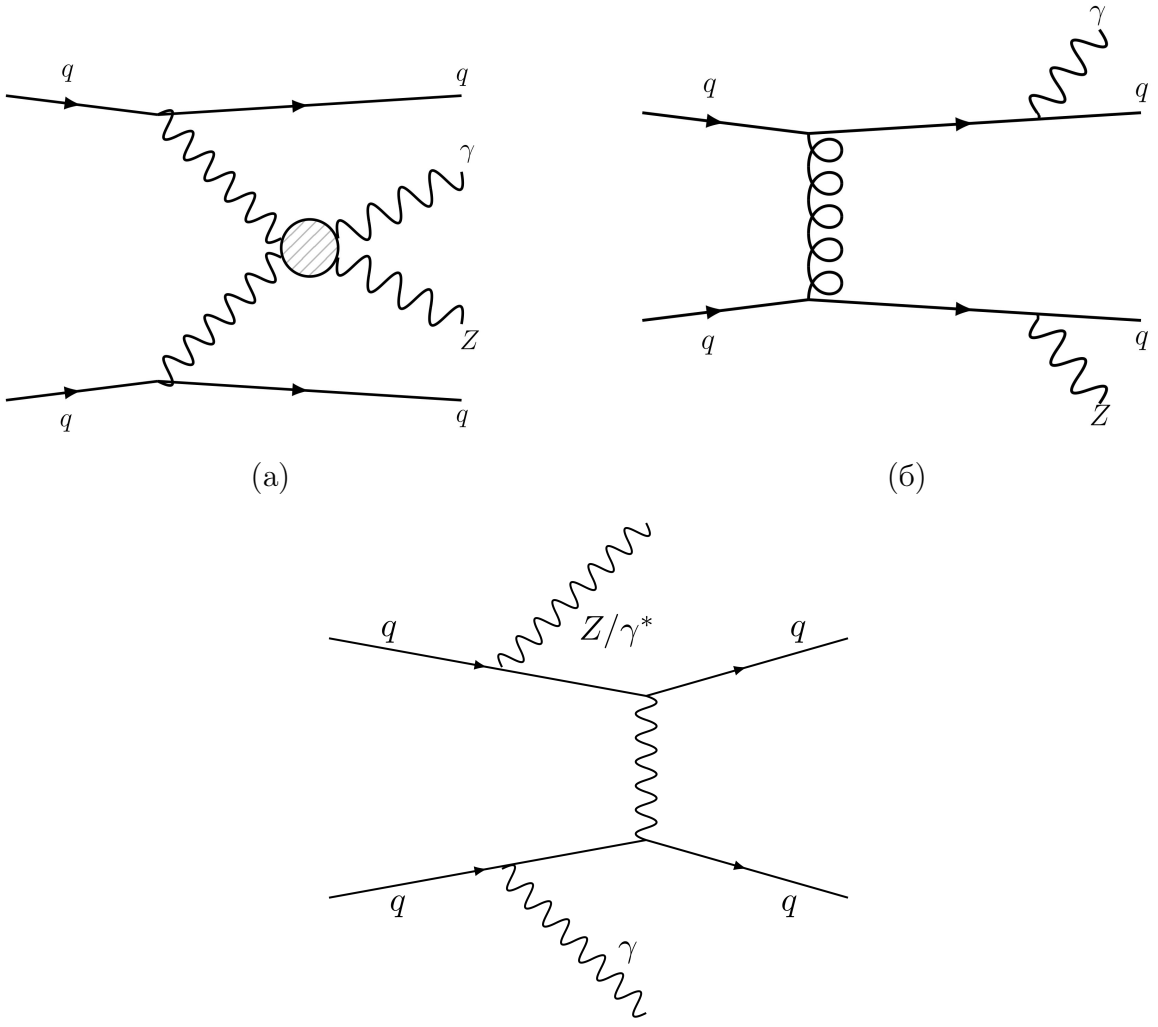


Рисунок 1 — Диаграммы процессов образования посредством рассеяния векторных бозонов (а), КХД (б) и электрослабое образование (в) состояния $Z\gamma jj$.

1 Методы

1.1 Композиция деревьев решений

Композиция деревьев решений, созданная с помощью градиентного бустинга (Boosted Decision Trees, BDT) – это классификатор с бинарной древовидной структурой. Принцип его работы заключается в поочерёдном применении ограничений по различным переменным с построением дерева решений. Это, в отличии от применения фиксированных отборов, позволяет разбить фазовое пространство на множество областей, которые классифицируются как сигнальные или фоновые.

1.1.1 Принцип построения дерева решений

Входные данные попадают в корневой узел дерева, далее производятся отборы по переменным так, чтобы максимизировать коэффициент разделение сигнала и фона. Затем из этих отборов выбирается тот, который обеспечивает максимальное разделение событий. Процесс повторяется для каждого дочернего узла до тех пор, пока количество событий в каком-либо из них не станет меньше установленного. Далее все узлы классифицируются как сигналоподобные или фоноподобные в зависимости от коэффициента чистоты или от преобладания в них сигнальных, либо фоновых событий. Схематичный вид дерева решений изображен на рисунке 2.

1.1.2 Бустинг

Недостатком деревьев решений является их чувствительность к флуктуациям в исходных данных. Например, из-за флуктуации одной переменной в тренировочном наборе данных может сильно повлиять на структуру итогового дерева решений.

Этой проблемы можно избежать, прибегнув к бустингу [8]. Суть этого алгоритма заключается в создании леса деревьев решений. При последо-

вательном создании каждого дерева веса событий тренировочного образца изменяются таким образом, чтобы максимизировать влияние на построение дерева тех переменных, которые были неправильно классифицированы на предыдущих шагах. При этом каждому дереву присваивается вес, который отражает его эффективность в разделении событий.

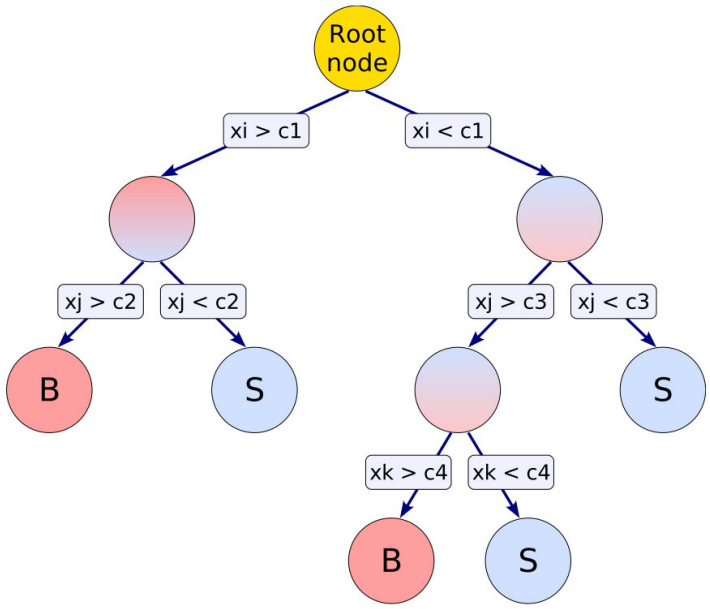


Рисунок 2 — Схематичный вид дерева решений.

При применении классификатора к набору данных, события поступают на вход каждому дереву решений, его отклик равен 1, если событие сигнальное и -1, – если фоновое. Отклик классификатора – непрерывная величина, лежащая в пределах [-1;1] и являющаяся взвешенной суммой откликов всех деревьев в лесу. Распределение по отклику можно использовать для разделения сигнальных и фоновых событий.

1.1.3 Оценка результатов работы алгоритма

Для оценки эффективности работы алгоритма используются параметр значимости при применении ограничений к отклику классификатора (1.1), а также площадь под ROC-кривой, которая является функцией зависимости эффективности отбора сигнала (1.2) и фонового отклонения (1.3) как функций от значения ограничения по отклику. Эффективность сигнала определяется как доля сигнальных событий, которая остаётся по-

сле применения классификатора. Отклонение фона – это доля фоновых событий, исключаемых из исходного набора.

$$\sigma = \frac{S}{\sqrt{S+B}} \tag{1.1}$$

$$\varepsilon = \frac{S}{S_{\text{init}}} \tag{1.2}$$

$$\kappa = 1 - \frac{B}{B_{\text{init}}} \tag{1.3}$$

где S – число сигнальных событий, B – число фоновых событий, S_{init} и B_{init} – число сигнальных и фоновых событий в исходном наборе соответственно.

Также для всех моделей осуществлялась оценка переобучения. Переобучение – это явление при котором модель хорошо проявляет себя при применении к тренировочной выборке, но плохо работает при применении к событиям из тестовой выборки.

С помощью критерия согласия Пирсона оценивалось совпадение распределений отклика при применении модели на тестовой и тренировочной выборках.

1.2 Метод главных компонент

Метод главных компонент [9] (Principal Component Analysis, PCA) – один из основных методов понижения размерности данных, подаваемых на вход классификатора, с потерей минимально возможного количества информации.

Суть метода заключается в нахождении подпространства меньшей размерности, чем у исходного пространства признков, в ортогональной проекции на которые выборочная дисперсия максимальна.

Алгоритм поиска главных компонент сводится к поиску собственных чисел λ_i и собственных векторов ω_i матрицы ковариации для исходных данных $X^T X$ и дальнейшей проекции признаков на них:

$$Z = X \begin{pmatrix} | & | & \dots & | \\ \omega_1 & \omega_2 & \dots & \omega_k \\ | & | & \dots & | \end{pmatrix} \tag{1.4}$$

Графическая интерпретация поиска главных компонент представлена на рисунке 3.

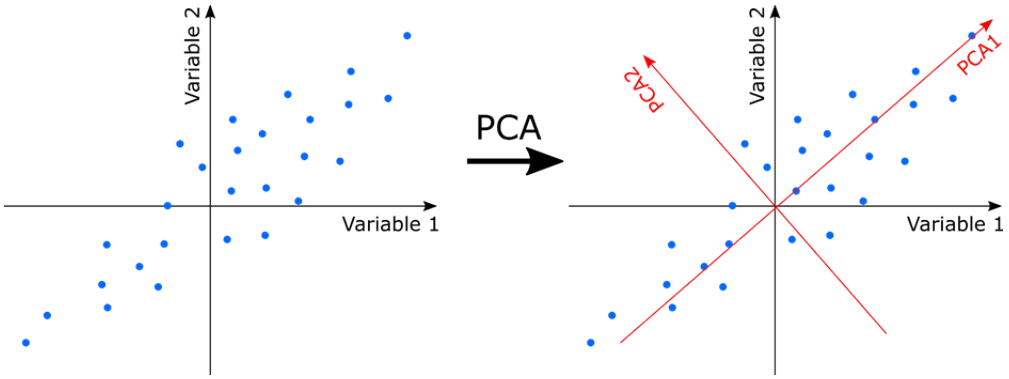


Рисунок 3 — Графическая интерпретация поиска главных компонент.

Собственные числа матрицы ковариации определяют дисперсию точек, спроецированных на соответствующую компоненту. По кумулятивному графику доли объясненной дисперсии, определяемой формулой 1.5,

$$d_i = \frac{\lambda_i}{\sum \lambda_i} \tag{1.5}$$

определяется количество компонент, которые в сумме дают определённый уровень объяснённой дисперсии.

1.3 Библиотека LGBM

Библиотека LightGBM [10] предоставляет реализации большого спектра алгоритмов деревьев решений и использует подход, основанный на гистограммах, который позволяет кратно уменьшить время обучения моделей. Также эта библиотека поддерживает работу с недостающими значениями. Во время разделения по какой либо из переменных при построении дерева решений события с недостающими значениями игнорируются. Далее узел, в который попадают события с недостающим значением переменной выбирается так, чтобы коэффициент разделения был наибольшим.

1.4 Отбор переменных

Для отбора переменных для обучения классификатора используется «N+1». Его принцип заключается в следующем. Изначально список переменных для обучения пуст. В него добавляется переменная-кандидат. На основе текущего списка создаётся классификатор и определяется метрика эффективности разделения событий(в работе в качестве метрики используется значимость). Затем переменная убирается из списка и процесс повторяется для всех остальных переменных-кандидатов. После этого выбирается переменная, которая обеспечивает наибольший прирост метрики разделения. Она окончательно добавляется в список переменных и исключается из переменных-кандидатов. Затем весь алгоритм повторяется до тех пор, пока список переменных-кандидатов не будет пуст.

Далее строится зависимость значения метрики разделения от выбранных переменных. Список переменных выбирается как первые k переменных до момента, когда метрика перестаёт расти.

1.5 Подбор настроек классификаторов

Для подбора оптимальных настроек классификаторов для обучения было создано большое число моделей с случайными значениями гиперпараметров. Для каждой модели строились распределения отклика сигнала и фона при её применении к тестовой и тренировочной выборке для оценки переобучения. Перетренированные модели отбрасывались с помощью критерия Пирсона на уровне значимости 0.05. После этого выбирались модели с наибольшей значимостью разделения событий.

Случайный выбор настроек обеспечивает намного более быстрый подбор оптимальных значений по сравнению с перебором настроек «по сетке».

2 Используемые данные

2.1 Устройство детектора

ATLAS [11] – это многоцелевой 4π-детектор, один из четырёх крупнейших детекторов на Большом адронном коллайдере, расположенном в Европейской организации по ядерным исследованиям CERN, Женева, Швейцария. Он состоит из внутреннего трекового детектора, электромагнитного и адронного калориметров, магнитной системы, а также мюонной системы. Устройство детектора ATLAS изображено на рисунке 4. Трековый детектор предназначен для определения треков заряженных частиц для измерения их импульса. Калориметры необходимы для измерения энерговыделения частиц, мюонная система используется для определения импульса и направления пролёта мюонов. Магнитная система необходима для искривления траекторий заряженных частиц для определения их импульса.

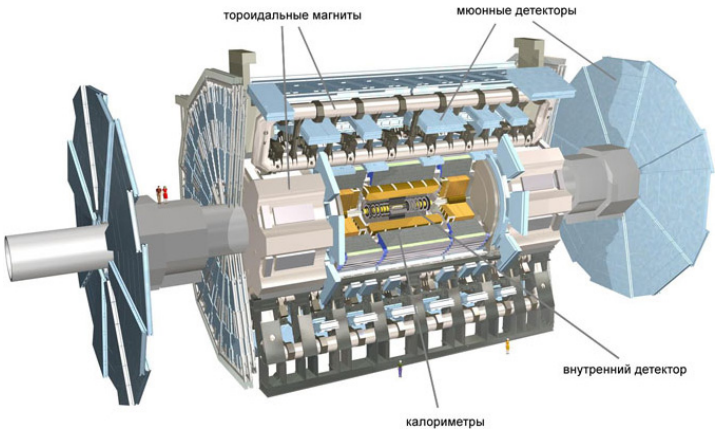


Рисунок 4 — Устройство детектора ATLAS.

Триггерная система детектора состоит из нескольких уровней. Она снижает частоту событий с десятков мегагерц до сотен герц, отбирая события, представляющие интерес для анализа.

Для описания направления вылета частиц, используется цилиндрическая система координат. Азимутальный угол ϕ отсчитывается в плоскости, перпендикулярной оси детектора, полярный θ – от положительного направления оси z, направленной вдоль оси детектора. Обычно вместо угла θ используется величина псевдобыстроты (2.1), дающая более равномерное распределение частиц, рождённых при столкновении.

$$\eta = -\ln \operatorname{tg} \frac{\theta}{2} \tag{2.1}$$

2.2 Исходные данные

Работа проводится с данными, полученными методом МК моделирования протон-протонного столкновения в детекторе ATLAS на Большом адронном коллайдере с энергией в системе ценра масс 13 ТэВ и интегральной светимости 139 фб$^{-1}$. К исходным данным применены ограничения для отбора кандидатов на процессы с конечным состоянием, содержащим Z-бозон, фотон и две адронные струи (Z-бозон распадается на нейтрино и антинейтрино, которые дают недостающий поперечный импульс). Ограничения перечислены в таблице 1. Ограничение накладываемое на E_T^γ обеспечивает эффективность отбора событий 98.5%. Условия на число фотонов, струй соответствует конечному состоянию процесса. Лептонное вето отсеивает процессы с лептонами в конечном состоянии. Угловые ограничения оптимизированы таким образом, чтобы максимально подавлять прочие фоны. Применённые отборы образуют $Z\gamma$ инклюзивную контрольную область.

Фоновыми процессами являются смоделированные с помощью МК процессы:

- $Z(\nu\nu)\gamma$ QCD – КХД образование Z-бозона с фотоном с последующим распадом Z-бозона на нейтрино и антинейтрино.
- $W\gamma$ QCD и $W\gamma$ EWK – КХД и электрослабые процессы рождения W-бозона с фотоном.
- ttγ – образование пары топ-кварков с фотоном.

- $Z(ll)\gamma$ – фон, связанный с потерей электрона или мюона.

А также фоны, оцениваемые из данных, связанные с неправильной идентификацией частиц или неверным измерением параметров:

- $W(e\nu)$, top, $\mathrm{t\bar{t}}$ – фон, связанный с идентификацией электрона как фотона.

- γ + j – фон, связанный с неверным измерением недостающего поперечного импульса.

- Zj, jj – фон, связанный с идентификацией струи как фотона.

Для анализа данные дополнительно разделяются на области фазового пространства. Это ноебходимо для верной оценки некоторых фоновых процессов с помощью реальных данных. Ограничения для областей выбираются таким образом, чтобы в них была как можно большая доля событий того процесса, который необходимо оценить. $W\gamma$ контрольная область используется для оценки фонов $W\gamma$ QCD, $W\gamma$ EWK и $\mathrm{tt}\gamma$. $Z\gamma$ контрольные области используются для оценки фона $Z\gamma$ QCD. Критерии отбора для областей приведены в таблице 2, где E_T^{miss} и E_T^{γ} – потерянная поперечная энергия и поперечная энергия фотона соответственно. Графическое представление критериев отбора для областей представлено на рисунке 5.

Таблица 1 — Критерии отбора событий.

Переменная	Ограничение
E_T^{miss}	>120 GeV
E_T^{γ}	>150 GeV
Число фотонов	$N_\gamma = 1$
Число струй	$N_{jets} \geq 2$
Число лептонов	$N_e = 0, N_\mu = 0$
$\lvert\Delta\phi(\gamma, \vec{p}_T^{\,miss})\rvert$	> 0.4
$\lvert\Delta\phi(j_1, \vec{p}_T^{\,miss})\rvert$	> 0.3
$\lvert\Delta\phi(j_2, \vec{p}_T^{\,miss})\rvert$	> 0.3

Таблица 2 — Определение контрольных и сигнальных областей.

Переменная	Ограничение
$W\gamma$ контрольная область	
N_{leptons}	≥ 1
$Z\gamma$ QCD контрольная область 1	
N_{leptons}	$= 0$
m_{jj}	< 300 GeV
$Z\gamma$ QCD контрольная область 2	
N_{leptons}	$= 0$
m_{jj}	> 300 GeV
$\zeta(\gamma)$	> 0.6
$Z\gamma$ EWK сигнальныая область	
N_{leptons}	$= 0$
m_{jj}	> 300 GeV
$\zeta(\gamma)$	< 0.6

где $\zeta(\gamma)$ – центральность фотона, определяемая формулой

$$\zeta(\gamma) = \left| \frac{\eta_\gamma - \frac{\eta_{j_1}+\eta_{j_2}}{2}}{\eta_{j_1} - \eta_{j_2}} \right| \tag{2.2}$$

где η_γ – псевдобыстрота фотона, η_{j_1}, η_{j_2} – псевдобыстроты струй.

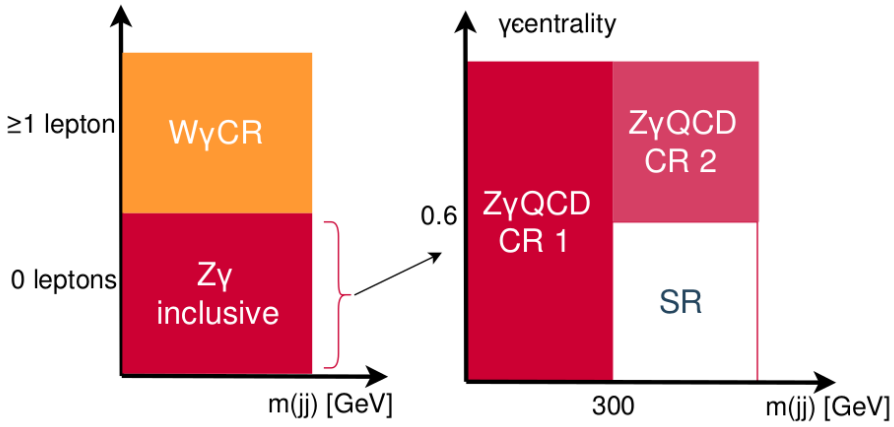


Рисунок 5 — Определение контрольных и сигнальных областей.

Обучение моделей производилось в $Z\gamma$ инклюзивной области. Определение значимости проводилось в $Z\gamma jj$ EWK сигнальной области. Для избежания переобучения моделей из обучающей выборки были исключены фоны, содержащие малое число событий: γ + j; Zj, jj; $Z(ll)\gamma$.

3 Процесс работы и результаты

3.1 Конструирование переменных

Из вида диаграмм сигнального и фонового процессов, изображённых на рисунке 6 можно сделать предположения о виде кинематических распределений.

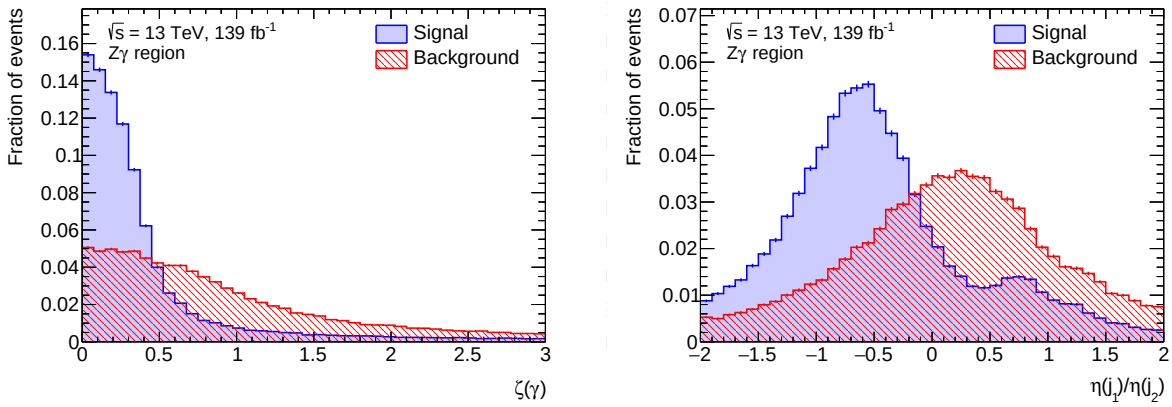
В первую очередь, в фоновом процессе фотон излучается одной из струй и поэтому прижимается к ней. Из этого следует наличие корреляции по псевдобыстроте между фотоном и одной из струй. В отличии от фона, в сигнальном процессе фотон вылетает в промежутке по псевдобыстроте между двумя лидирующими струями. Этот эффект наблюдается на распределении по центральности фотона, изображённом на рисунке 6а.

Центральность, определяемая формулой 2.2, показывает как ориентировано направление вылета фотона относительно струй. Она равна нулю если фотон вылетает строго между двумя струями и стремится к 0.5 если фотон прижимается к одной из струй.

Из форм распределений видно, что центральность фотона может являться эффективной переменной для разделения сигнальных и фоновых событий, однако, она используется для определения контрольных регионов, поэтому не может использоваться для обучения классификатора.

Другая переменная – отношение псевдобыстрот двух струй. Из диаграмм можно сделать вывод о том, что оно будет иметь большую дисперсию для фона, так как струи излучают фотон и массивный Z-бозон, в отличии от сигнального процесса, где псебдобыстроты струй более скореллированы. Это наблюдается на распределении на рисунке 6б. Также можно заметить, для большей части сигнальных событий струи вылетают в противоположные стороны, о чём свидетельствует смещение распределения отношения псевдобыстрот струй в отрицательную область.

Также были сконструированы переменные, представляющие из себя различные комбинации параметров частиц.



(а) (б)

Рисунок 6 — Распределения по переменным центральности фотона (а) и отношения псевдобыстрот двух лидирующих струй (б).

3.2 Отбор переменных

Отбор переменных для обучения производился с помощью метода «N+1». График зависимости значимости при отборе по отклику классификатора от используемого для обучения набора переменных представлен на рисунке 7.

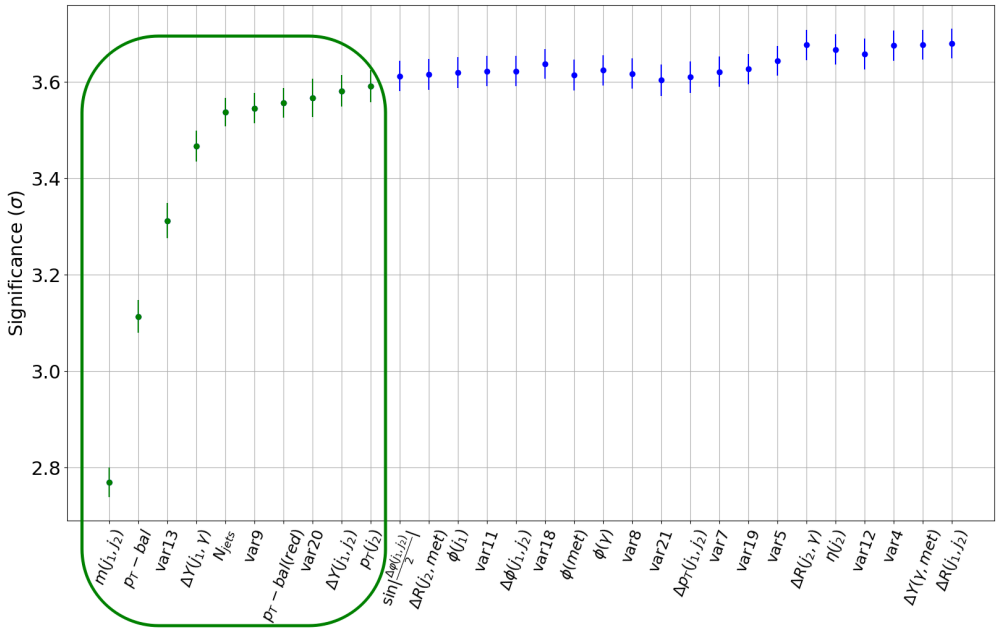


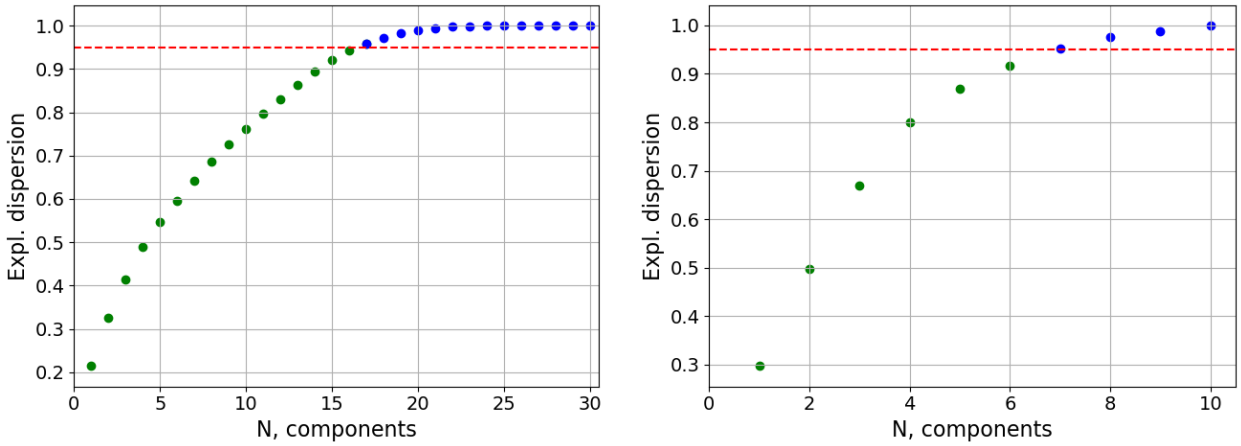
Рисунок 7 — График зависимости значимости при отборе по отклику классификатора от используемого для обучения набора переменных (отобранные переменные отмечены зелёным цветом).

Список отобранных переменных и их распределения представлены в аппендиксе A. Проверка моделирования в контрольных областях представлена в аппендиксе B.

3.3 Использование метода понижения размерности

Была произведена проверка возможности использования метода понижения размерности для увеличения эффективности работы классификатора. Для этого он был применён к 30 исходным переменным, полученным методом «N+1». Далее были отобрыны первые n компонент, обеспечивающих 95% объяснённой дисперсии. На рисунке 8а представлен кумулятивный график зависимости объяснённой дисперсии от количества компонент. Таким обрызом было отобрано 16 компонент для обучения классификатора.

Для проверки также рассматривался случай применения метода главных компонент к отобранным переменным. Из графика зависимости объяснённой дисперсии от количества компонент, изображённого на рисунке 8б, видно, что из набора из 10 переменных можно составить 6 компонент, обеспечивающих 95% объяснённой дисперсии.



(а) (б)

Рисунок 8 — График зависимости объяснённой дисперсии от количества компонент.

3.4 Тренировка и применение моделей

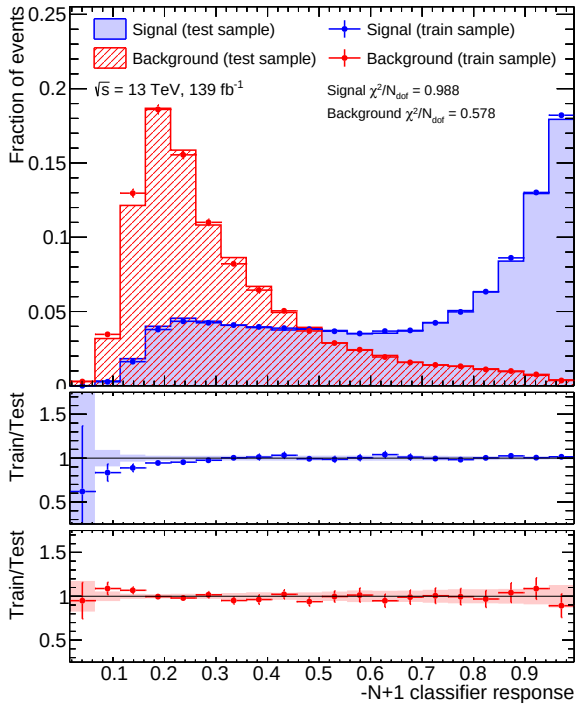
Для каждого из трёх случаев был создан большой набор классификаторов с случайными значениями настроек и выбраны лучшие. Для каждого классификатора были построены распределения по отклику для сигнала и фона при применении к тренировочной и к тестовой выборкам с целью проверки переобучения. Они изображены на рисунке 9.

Сводные характеристики работы классификаторов приведены в таблице 3.

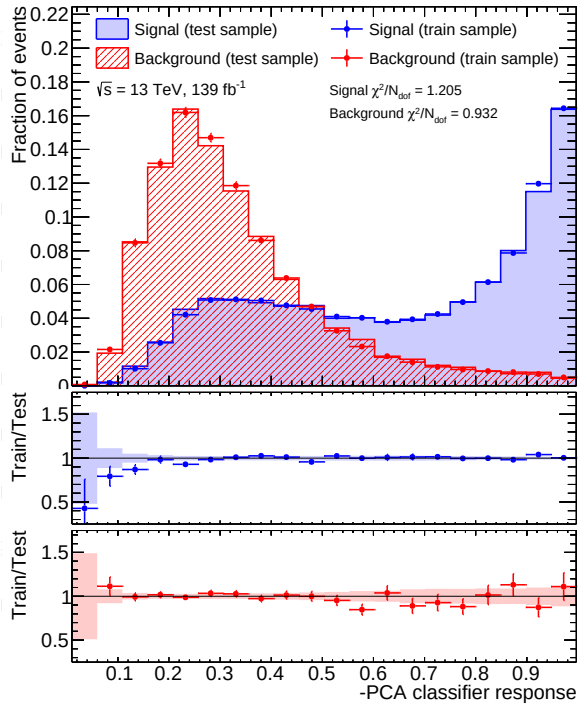
Таблица 3 — Сводные результаты работы классификаторов для трёх методов отбора переменных.

	Вхожд. сигнал	Вхожд. 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

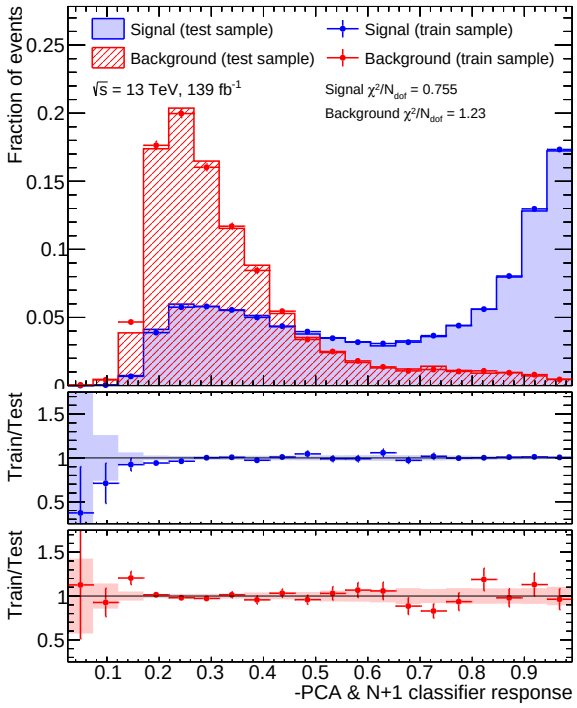
Из результатов следует, что прямое использование метода главных компонент значительно более плохой результат, чем подбор методом «N+1». Также наблюдается падение эффективности классификатора при применении метода главных компонент к переменным, отобранным методом «N+1».



(а)



(б)



(в)

Рисунок 9 — Распределения по отклику классификатора для случаев получения переменных для обучения классификатора методом «N+1»(а), методом понижения размерности (б) и применения метода понижения размерности к набору переменных, отобранных методом «N+1»(в).

Заключение

В процессе работы были сконструированы новые переменные в дополнении к имеющемуся набору, из них был произведён отбор оптимальных переменных для обучения классификаторов методом «N+1», методом главных компонент и комбинированным методом. Для каждого из наборов переменных были оптимизированны настройки классификатора. В результате максимальный прирост значимости был получен при использовании алгоритма «N+1» – с $(2.49 \pm 0.02)\sigma$ до $(3.61 \pm 0.03)\sigma$. Наилучший результат для метода главных компонент был получен при его совместном использовании с методом «N+1» и составляет $(3.47 \pm 0.03)\sigma$, что значительно ниже, чем при использовании метода «N+1» и доказывает неоправданность его использования.

Для всех отобранных переменных было проверено качество их согласованности с данными. По её результатам можно сказать, что модели, обученные на используемом в исследовании наборе МК данных, могут применяться для классификации событий в реальных данных.

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

Список используемых источников

1. *Group P. D.* [et al.]. Review of Particle Physics // Progress of Theoretical and Experimental Physics. — 2020. — Aug. — Vol. 2020, no. 8. — eprint: https://academic.oup.com/ptep/article-pdf/2020/8/083C01/34673722/ptaa104.pdf ; — 083C01.
2. *Hoecker A.* [et al.]. TMVA - Toolkit for Multivariate Data Analysis. — 2007. — arXiv: physics/0703039 [physics.data-an].
3. *Aaboud M.* [et al.]. Studies of $Z\gamma$ production in association with a high-mass dijet system in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector // Journal of High Energy Physics. — 2017. — July. — Vol. 2017, no. 7.
4. *Khachatryan V.* [et al.]. Measurement of the cross section for electroweak production of $Z\gamma$ in association with two jets and constraints on anomalous quartic gauge couplings in proton–proton collisions at s=8 TeV // Physics Letters B. — 2017. — Vol. 770. — P. 380–402.
5. *Éboli O. J. P.*, *Gonzalez-Garcia M. C.*, *Lietti S. M.* Bosonic quartic couplings at CERN LHC // Phys. Rev. D. — 2004. — May. — Vol. 69, issue 9. — P. 095005.
6. *Éboli O. J. P.*, *Gonzalez–Garcia M. C.* Classifying the bosonic quartic couplings // Phys. Rev. D. — 2016. — May. — Vol. 93, issue 9. — P. 093013.
7. *Baak M.* [et al.]. Study of Electroweak Interactions at the Energy Frontier. — 2013.
8. *Friedman J. H.* Greedy function approximation: A gradient boosting machine // Ann. Stat. — 2001. — Vol. 29, no. 5. — P. 1189–1232.
9. *Jolliffe I. T.*, *Cadima J.* Principal component analysis: a review and recent developments. — 2016.

10. *Ke G.* [et al.]. LightGBM: A Highly Efficient Gradient Boosting Decision Tree // Advances in Neural Information Processing Systems. Vol. 30 / ed. by I. Guyon [et al.]. — Curran Associates, Inc., 2017.

11. *Collaboration T. A.* [et al.]. The ATLAS Experiment at the CERN Large Hadron Collider // Journal of Instrumentation. — 2008. — Aug. — Vol. 3, no. 08. — S08003–S08003.

A Распределения по переменным

Были построены распределения по переменным, отобранным для обучения классификатора методом «N+1».

- m_{jj} – инвариантаная масса двух струй
- $p_T - \text{balance} = \frac{|\vec{p}_T^{\,miss} + \vec{p}_T^{\,\gamma} + \vec{p}_T^{\,j1} + \vec{p}_T^{\,j2}|}{E_T^{miss} + E_T^{\,\gamma} + p_T^{\,j1} + p_T^{\,j2}}$ – баланс поперечных импульсов
- $var13 = \sqrt{p_T(j_1)^2 + p_T(j_2)^2}$
- $\Delta Y(j_1, \gamma)$ – разность быстрот лидирующей струи и фотона
- N_{jets} – число адронных струй
- $var9 = \frac{E_T^{miss}}{p_T(j_1) + p_T(j_2)}$
- $p_T - \text{balance(reduced)} = \frac{|\vec{p}_T^{\,\gamma} + \vec{p}_T^{\,j1} + \vec{p}_T^{\,j2}|}{E_T^{\gamma} + p_T^{\,j1} + p_T^{\,j2}}$
- $var20 = \frac{E_T^{miss} - p_T^{\gamma}}{p_T(j_1) + p_T(j_2)}$
- $\Delta Y(j_1, j_2)$ – разность быстрот двух струй
- $p_T(j_2)$ – поперечный импульс второй струи

Лидирующая и вторая струя – это струи, расположенные в порядке возрастания поперечного импульса. В качестве фоновых событий использовалась сумма всех фонов, рассматриваемых в этом анализе. На рисунках 10-11 представлены распределения по отобранным переменным.

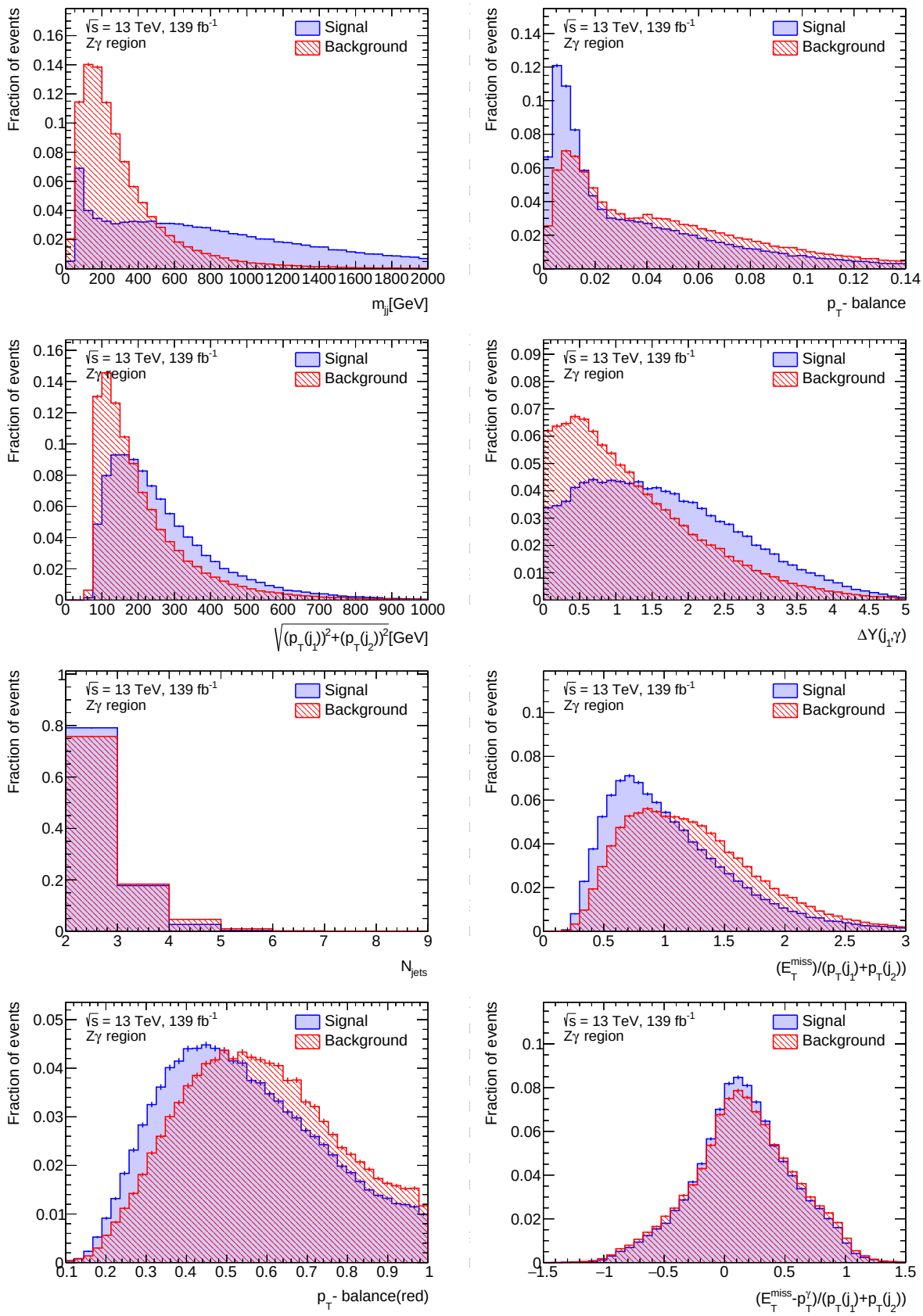


Рисунок 10 — Распределения по отобранным переменным, нормированные на полное число событий, для сигнала и фона.

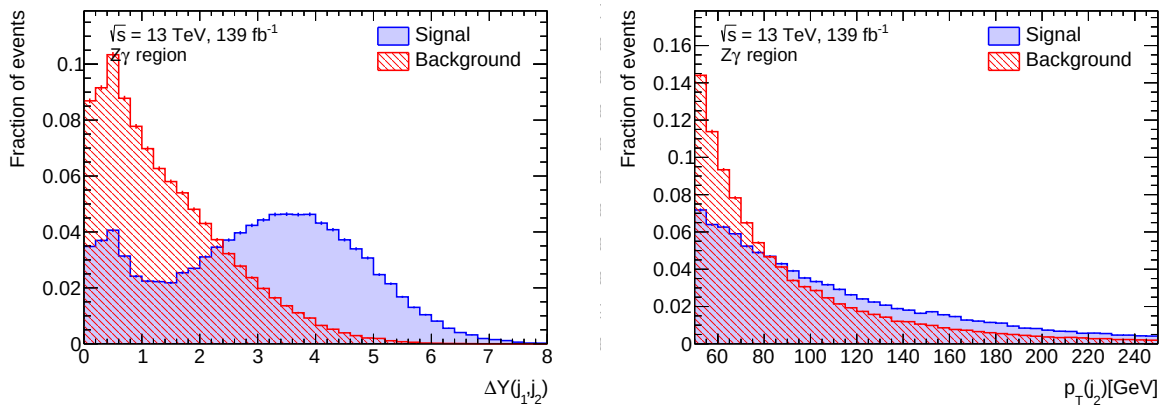


Рисунок 11 — Распределения по отобранным в работе переменным, нормированным на полное число событий, для сигнала и фона.

B Проверка моделирования переменных

Распределения переменных для МК-моделирования и реальных данных в трёх контрольных областях, исследуемых в анализе, представлены на рисунках 12-18. По ним можно видеть, что МК смоделированные данные довольно хорошо согласуются с реальными данными. Поэтому можно сказать, что модели, обученные на смоделированных данных, могут применяться к реальным данным для классификации событий.

B.1 Первая Zγ QCD контрольная область

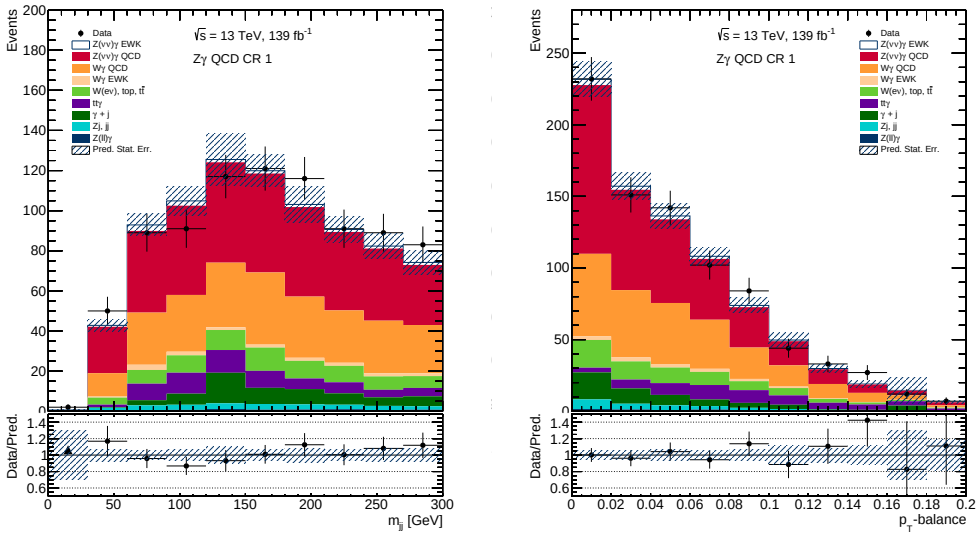


Рисунок 12 — Сравнение распределений переменных реальных и смоделированных данных для первой $Z\gamma$ QCD контрольной области. Штриховкой обозначена статистическая погрешность смоделированных данных.

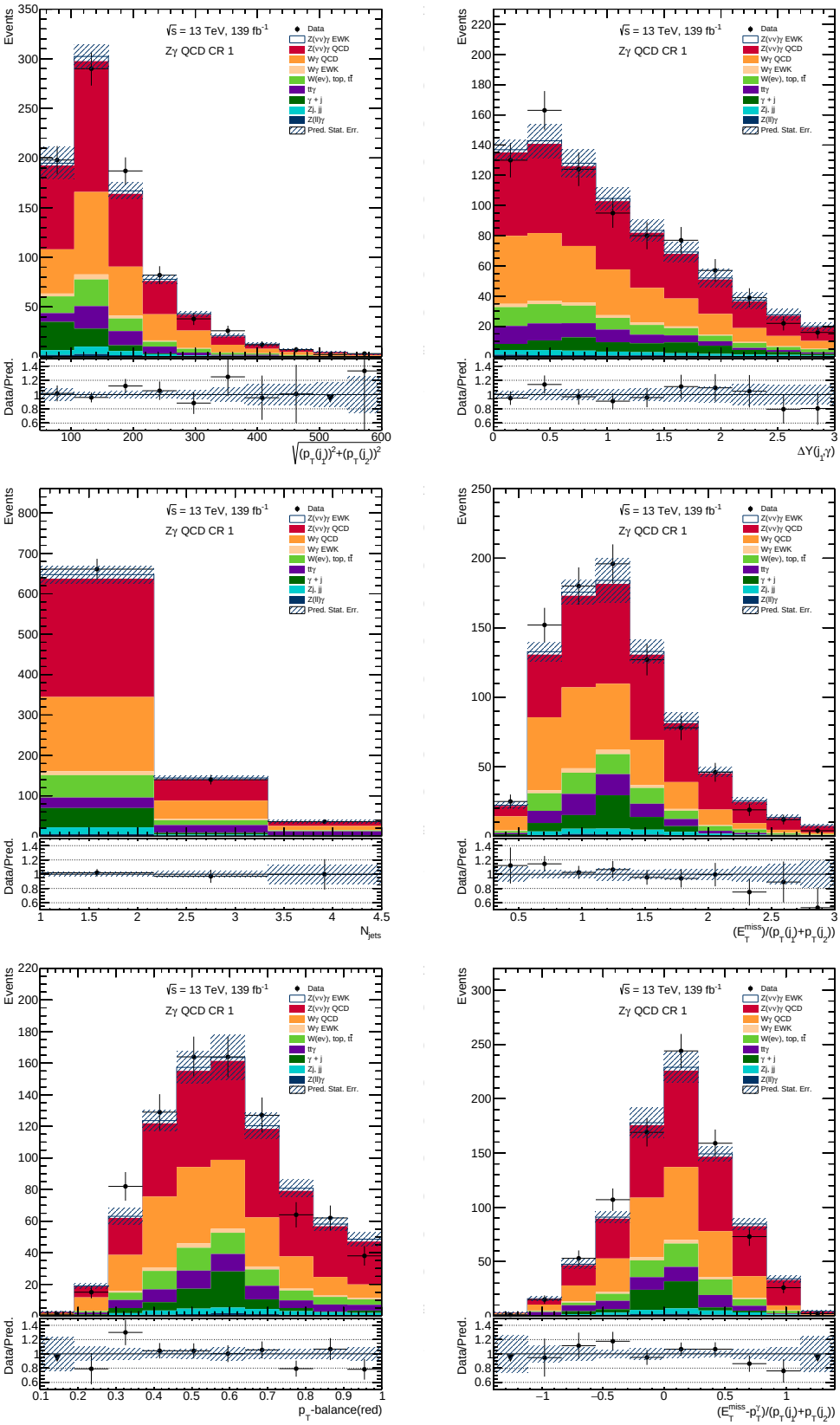


Рисунок 13 — Сравнение распределений переменных реальных и смоделированных данных для первой $Z\gamma$ QCD контрольной области. Штриховкой обозначена статистическая погрешность смоделированных данных.

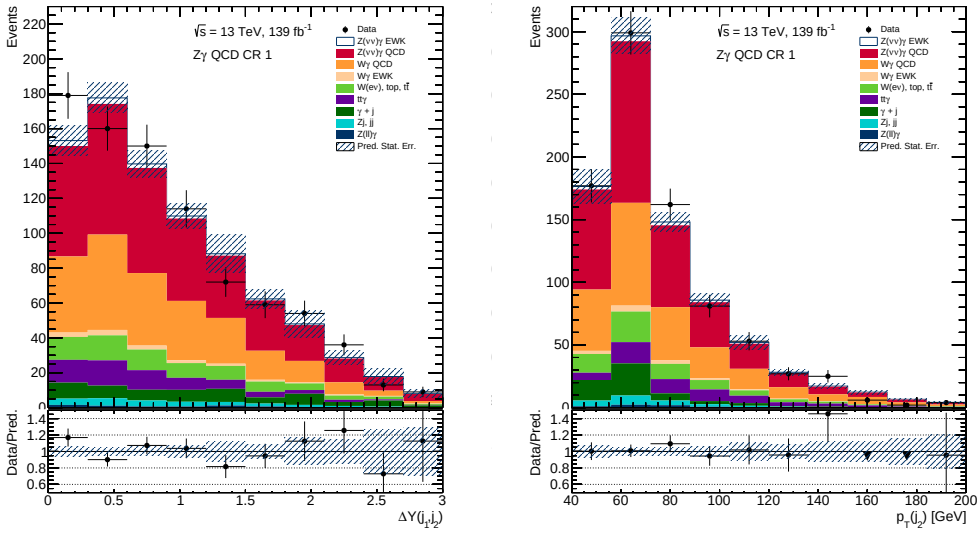


Рисунок 14 — Сравнение распределений переменных реальных и смоделированных данных для первой $Z\gamma$ QCD контрольной области. Штриховкой обозначена статистическая погрешность смоделированных данных.

B.2 Вторая Zγ QCD контрольная область

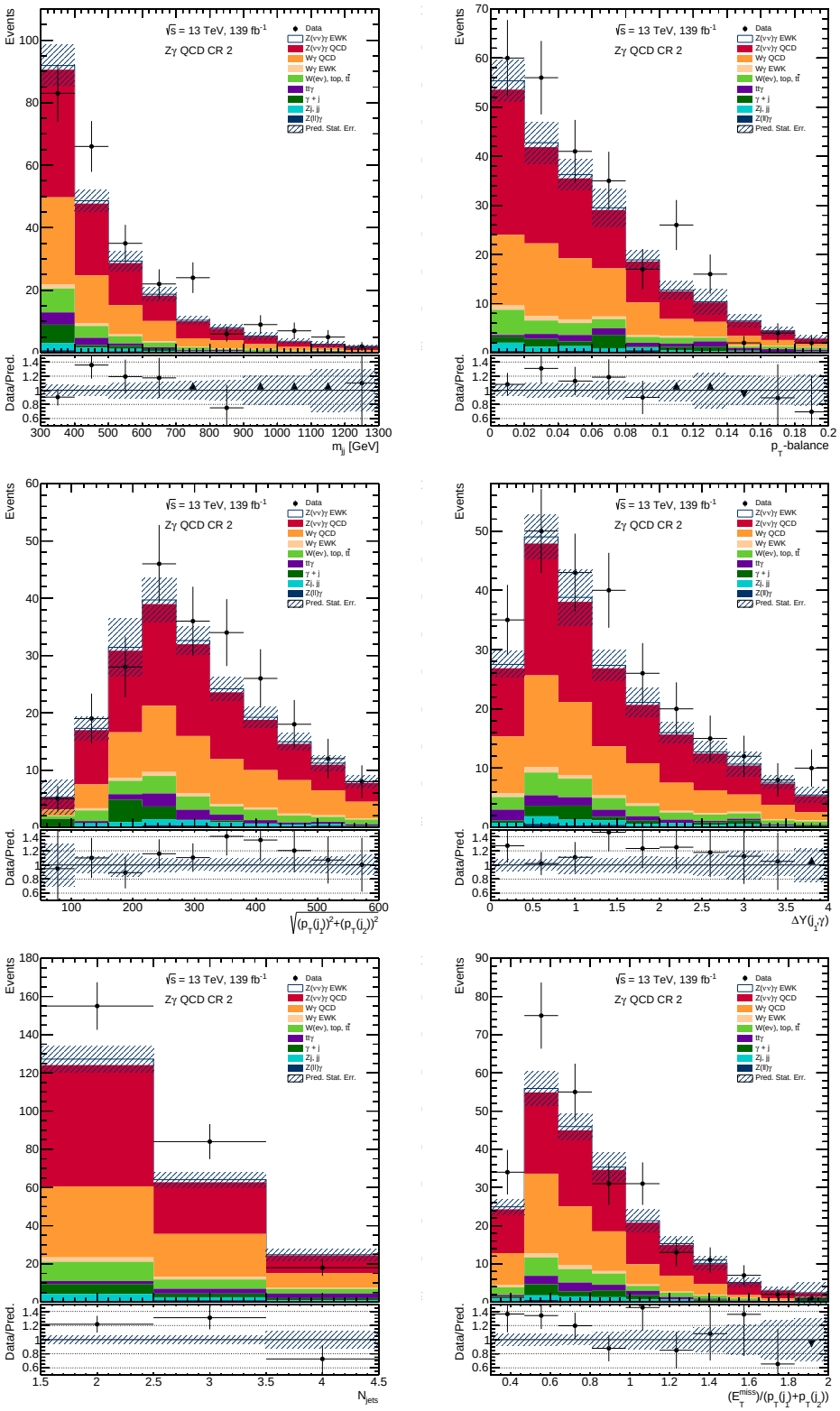


Рисунок 15 — Сравнение распределений переменных реальных и смоделированных данных для второй $Z\gamma$ QCD контрольной области. Штриховкой обозначена статистическая погрешность смоделированных данных.

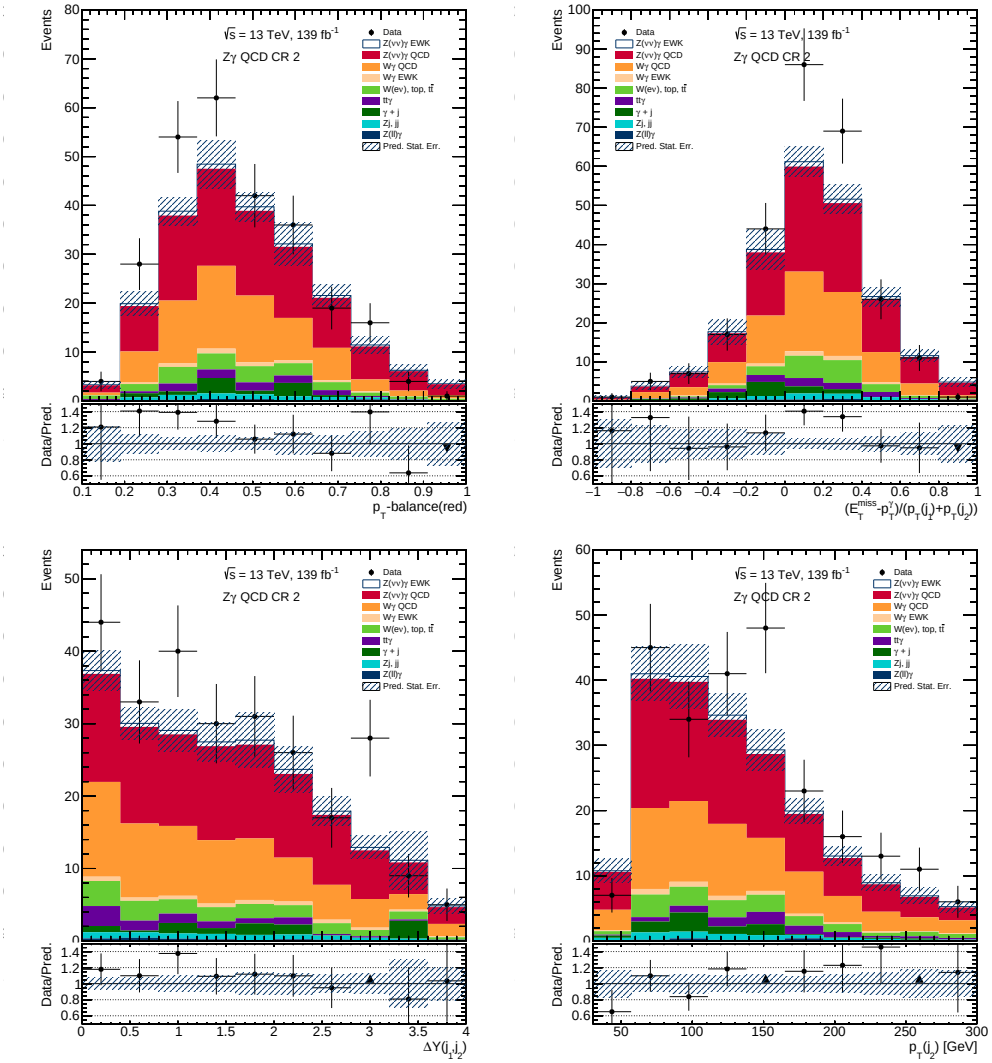


Рисунок 16 — Сравнение распределений переменных реальных и смоделированных данных для второй $Z\gamma$ QCD контрольной области. Штриховкой обозначена статистическая погрешность смоделированных данных.

B.3 Wγ контрольная область

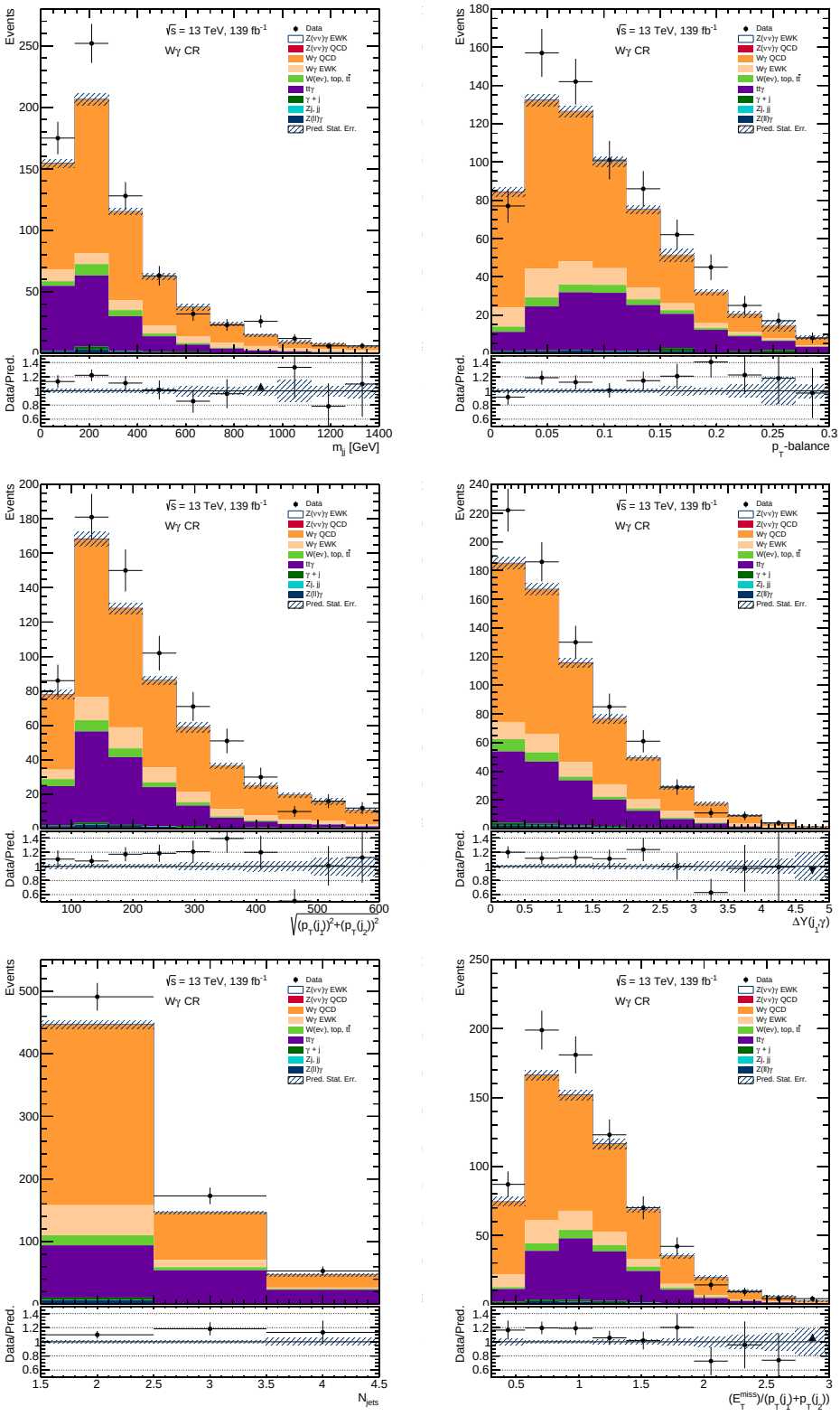


Рисунок 17 — Сравнение распределений переменных реальных и смоделированных данных для $W\gamma$ QCD контрольной области. Штриховкой обозначена статистическая погрешность смоделированных данных.

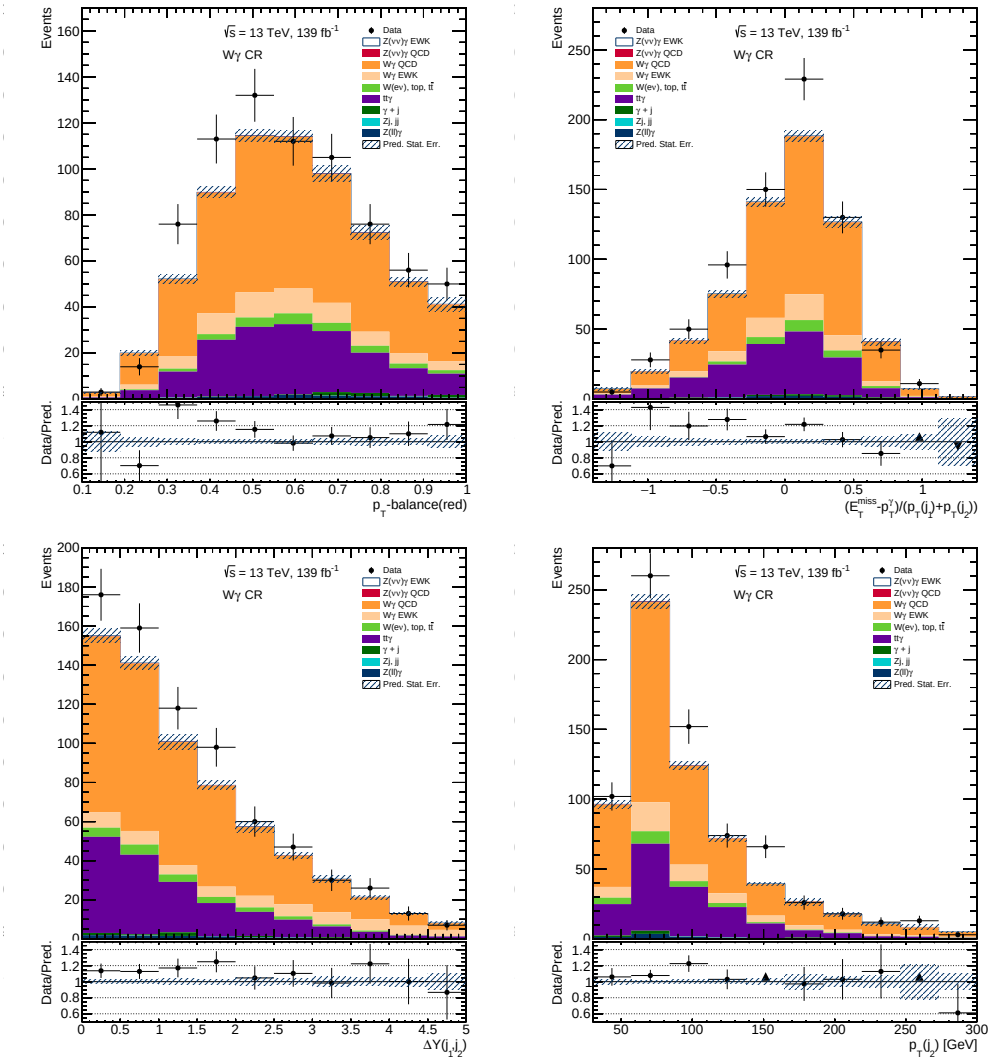


Рисунок 18 — Сравнение распределений переменных реальных и смоделированных данных для $W\gamma$ QCD контрольной области. Штриховкой обозначена статистическая погрешность смоделированных данных.