Facultad de Ciencias

Etiquetado de b-jets en el experimento CMS del LHC
(b-tagging in the CMS experiment of the LHC)

Trabajo de Fin de Grado
para acceder al

GRADO EN FÍSICA

Autor: Imobach García Ibañez

Director: Jesús Manuel Vizán García

Febrero - 2023

Agradecimientos:

Quisiera agradecer:

A la Universidad de Cantabria y su personal la formación recibida a lo largo de estos años del grado en Física. No siempre ha sido un camino fácil pero siempre ha merecido la pena.

Al director de este TFG, Jesús Vizán, su orientación, guía y ayuda durante todos estos meses.

En especial a mi madre, padre y hermana, su apoyo constante. Sin vosotros esto no hubiese sido posible. Gracias por ayudarme a llegar hasta aquí.

Resumen:

El acelerador LHC (Gran Colisionador de Hadrones) del CERN (Organización Europea para la Investigación Nuclear) es el acelerador de partículas mas grande construido y el CMS (Solenoide Compacto de Muones) es un detector de propósito general ubicado en él. Su uso ha aportado un extenso conocimiento en el campo de la Física de Partículas. La realización de numerosos análisis en el CMS dependen de la correcta identificación del quark b (bottom) y para ello se emplean algoritmos de alto nivel a las muestras de datos obtenidos pudiendo distinguir entre jets de partículas asociados a un quark b y a sabores mas ligeros. El uso de dichos algoritmos fue empleado en el descubrimiento del bosón de Higgs en el año 2012 por las colaboraciones CMS y ATLAS (el principal modo de decaimiento del bosón de Higgs es en pares de quarks b) y está actualmente involucrado en el estudio de medidas dentro del Modelo Estándar de Partículas - producción de pares de bosones de Higgs, mecanismo de producción del Higgs junto a pares de quarks t... etc. - y mas allá de él, como la búsqueda de supersimetría en estados finales con jets que proceden de un quark b.

La determinación de las eficiencias de los algoritmos de b-tagging - así como la diferencia de sus valores en una muestra de datos reales y en una muestra de datos simulados - es fundamental para que su uso pueda ser empleado satisfactoriamente. Para ello se parte de datos enriquecidos en b-jets los cuales contienen un muón y se utilizan diferentes técnicas de etiquetado de b-jets con bajo nivel de correlación entre sí permitiendo la construcción de un sistema de ocho ecuaciones con ocho incógnitas (System8) cuya resolución permite la determinación de las eficiencias deseadas. La correlación entre las técnicas utilizadas introduce ocho parámetros adicionales en el sistema de ecuaciones, los factores de correlación, cuyo valor es determinado a partir de muestras de datos simulados.

En este trabajo se han calculado las eficiencias de dos algoritmos de b-tagging (DeepCSV y DeepJet) correspondientes a la campaña de datos del segundo ciclo operacional del LHC (2015-2018), para el año 2018, del detector CMS: inicialmente se han recreado los resultados presentados oficialmente en la colaboración CMS por el IFCA (Instituto de Física de Cantabria). Debido a las limitaciones del método numérico usado que se observan - no siempre proporciona soluciones válidas - se ha implementado un método alternativo basado en el ajuste del System8 mediante una minimización. Su uso es prometedor, siendo capaz de recuperar para la campaña de datos considerada soluciones consistentes que el método numérico no halla. Finalmente se ha realizado, por primera vez en el IFCA, un estudio preliminar de la dependencia de las eficiencias de los algoritmos de b-tagging con la incertidumbre en los factores de correlación.

Palabras clave:
Eficiencias, algoritmos, etiquetado, b-jets, CMS, LHC.

Abstract:

The LHC (Large Hadron Collider) of CERN (European Organization for Nuclear Research) is the biggest particle accelerator ever built and the CMS (Compact Muon Solenoid) is a general purpose detector placed within the LHC. Its use has provided a large knowledge in the Particle Physics field. Many analysis carried in the CMS rely on the correct identification of b quarks (bottom) and for that purpose high level algorithms are applied to data samples allowing to distinguish between jets of particles associated to a b quark and to lighter quark flavours. The forementioned algorithms were used in the discovery of the Higgs boson in the year 2012 by the CMS and ATLAS collaborations (the main decay mode of the Higgs boson is a pair of b quarks) and are nowadays applied in measurements within the Standard Model of Particles - pair production of Higgs bosons, production mechanism of the Higgs with a pair of t quarks... etc - and beyond it, like the search for supersymmetry in final states with jets that come from a b quark.

The measurement of the efficiencies of the b-tagging algorithms - and a comparison of their values for a real data sample and for a simulation data sample - is fundamental for their satisfactory use. Different b-tagging techniques with little correlation between them are applied to samples enriched with b-jets that contain a muon allowing the construction of a system of eight equations with eight unknowns (System8) whose resolution allows the determination of the efficiency of the b-tagging algorithms. The correlation between the techniques used introduce eight parameters in the system of equations - the correlation factors - whose values are determined with simulation data samples.

In this work the efficiencies of two b-tagging algorithms (DeepCSV y DeepJet) corresponding to the data campaign of the CMS of the second run of the LHC (2015-2018), for the year 2018, have been calculated: first the official results presented by the IFCA (Institute of Physics of Cantabria) in the CMS collaborations have been recreated. Due to the fact that limitations in the use of a method based on finding a numeric solution of the System8 have been observed - a solution is not always found - an alternative method based in and adjustment of the System8 by a minimization has been implemented. The new method is promising and can recover coherent solutions for the data campaing considered that the numeric method can not. Finally, for the first time in IFCA, a preliminary study of the dependency of the efficiencies with the uncertainties introduced in the correlation factors has been carried.

Keywords:
Efficiencies, algorithms, tagging, b-jets, CMS, LHC.

Índice

1. Introducción

1.1. El modelo estándar de partículas

El modelo estándar es una teoría mecánico-cuántica y relativista que describe los constituyentes fundamentales del universo y sus interacciones. El inicio de su desarrollo fue en la década de los 70 impulsada por los avances logrados en las décadas previas en las áreas de la mecánica cuántica y la teoría de la relatividad especial y los descubrimientos experimentales de nuevas partículas como el positrón y el muón.

En el modelo estándar las interacciones entre las partículas se deben al intercambio de otras partículas: los bosones gauge. Cada tipo de interacción fundamental esta asociada a una simetría gauge bajo la cual la función de onda de las partículas permanece invariante.

Las partículas fundamentales que constituyen el universo según el modelo estándar son las siguientes (Figura 1):

Por un lado se tienen doce partículas - y sus correspondientes doce antipartículas, con carga opuesta a la partícula y diferentes números cuánticos - de spin 1/2 llamadas fermiones. Se dividen a su vez en dos grupos de 6 partículas cada una: 1) Seis leptones l, agrupados en tres generaciones: el electrón (e^-) y el neutrino electrónico (υ_e), el muón (μ^-) y el neutrino muónico (υ_μ) y el tau (τ^-) y el neutrino tauónico (υ_τ). El e^-, μ^- y τ^- tienen carga y sufren las interacciones electromagnéticas y débiles mientras que los tres neutrinos son partículas neutras y solo sufren interacciones débiles. 2) Seis quarks q, agrupados también en tres generaciones: up (u) y down (d), charm (c) y strangeness (s), top (t) y bottom (b). Se dice que cada quark es de un *sabor* diferente. Los quarks sufren la interacción electromagnética, ya que tienen carga, y la interacción fuerte debido a que poseen color. El color es un nuevo grado de libertad asignado a los quarks. Cada quark puede tener tres colores: *red* (r), *blue* (b) y *green (g)*. Debido a la hipótesis del confinamiento de color los quarks no se pueden encontrar aisladamente como partículas libres, si no en un hadrón (partícula compuesta de quarks) que sea un singlete de color. Los hadrones observables son los únicos que cumplen el confinamiento de color: los mesones (pares $q\bar{q}$) y bariones (tríos qqq o $\bar{q}\bar{q}\bar{q}$).

Por el otro lado se tienen los bosones gauge: el fotón (γ), el gluón (g) y los bosones W^+, W^- y Z^0, asociados a las 3 interacciones fundamentales: electromagnética, fuerte y débil respectivamente. El bosón restante es un bosón de spin 0 asociado al mecanismo que otorga masa a las partículas fundamentales: el bosón de Higgs.

Las interacciones fundamentales son: interacción electromagnética, interacción débil e interacción fuerte. Sus características básicas son:

1. Interacción electromagnética:
Esta mediada por el fotón, partícula de masa nula que provoca que el rango de la interacción sea infinito. El fotón se acopla a la carga por lo que tanto los seis quarks, el electrón, el muón, el tau - y sus antipartículas - como los bosones W interaccionan con él. Las partículas sin carga no interaccionan con los fotones. Los hadrones que decaen por interacción electro-

magnética tienen un tiempo de semivida de $\sim 10^{-16}$ s.

2. Interacción débil:
En la interacción débil se distingue aquella mediada por los bosones W^+ y W^- y la mediada por los bosones Z, llamadas corrientes neutras, con carga nula. Todos ellos son masivos. La interacción débil se acopla a los leptones del modelo estándar. Su fuerza intrínseca es similar a la electromagnética pero la masa de los bosones W y Z hacen que parezca mas débil. Los hadrones que decaen mediante interacción débil tienen un tiempo de semivida relativamente largo de $\sim 10^{-10} - 10^{-8}$ s. La interacción débil y la electromagnética están acopladas. Durante la década de los setenta se desarrollo la teoría electrodébil, que agrupaba ambas interacciones. En cualquier proceso en el que se pueda intercambiar un fotón también es posible que lo haga un bosón Z.

3. Interacción fuerte:
Los gluones, con masa nula, son las partículas que median la interacción fuerte, la cual se acopla al color. Los quark y los gluones tienen color y son las únicas partículas que sufren la interacción fuerte. A nivel fundamental es una interacción similar a la electromagnética pero el hecho de que los gluones interaccionen con otros gluones, al tener carga de color introduce ciertas notables diferencias, la principal siendo la llamada libertad asintótica. Como consecuencia los quarks se encuentran relativamente libres a pequeñas distancias pero al aumentar la separación entre ellos la fuerza se hace mas intensa evitando su separación o dando lugar a un proceso de hadronización con el quark. En la hadronización, entre los quarks que se separan, se genera debido a la libertad asintótica un tubo de densidad de energía muy alta en el que se crean gluones que decaen en pares quark-antiquark. Estos pares se recombinan con otros quarks para dar lugar a otras partículas compuestas que cumplen el confinamiento de color y experimentalmente se observan entonces jets de partículas.

4. El bosón y campo de Higgs:
Además de las tres interacciones fundamentales se tiene que mencionar el bosón de Higgs. Este bosón escalar de spin 0, predicho en el año 1964 por Peter Ware Higgs [1] y descubierto en el LHC por las colaboraciones ATLAS [2] y CMS [3] en el año 2012 es el responsable de la masa de las partículas en el modelo estándar. Surge por una ruptura espontánea del grupo de simetría local y global $SU(2)_L U(1)_Y$ (grupo de simetría electrodébil). Previa a la ruptura de dicha simetría se tendrían 3+1 bosones no masivos: los bosones W_1, W_2, W_3 y B. La combinación de los bosones W_1 y W_2 se identifican con los bosones W^+ y W^- mientras que la combinación de los bosones W_3 y B, mezclados con un ángulo de mezcla θ_W, dan lugar a los bosones observables Z y γ. Un desarrollo de la interacción del bosón de Higgs muestra como los bosones W^+, W^- y Z^0 adquieren su masa mientras que el fotón γ no interacciona con el campo creado por él y se queda con masa nula. El bosón de Higgs, en sus interacciones, se acopla a la masa de las partículas.

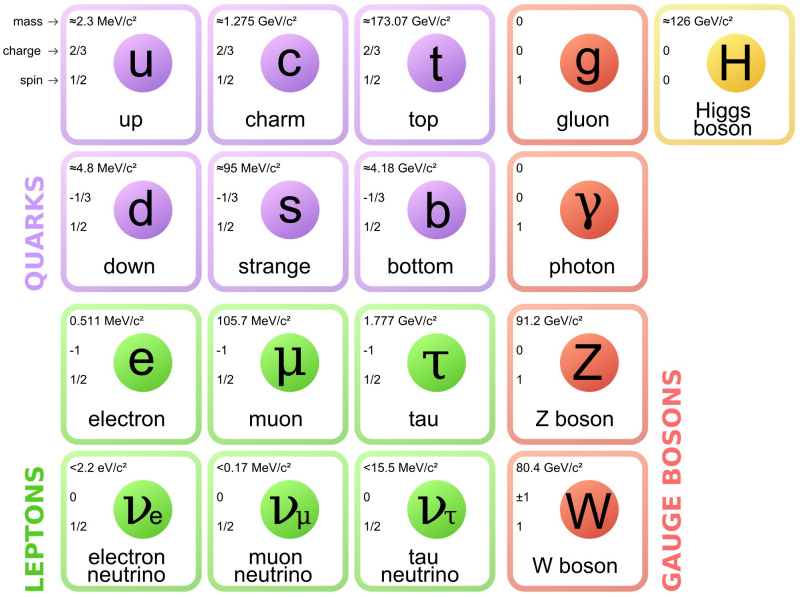

Figura 1: Partículas fundamentales del modelo estándar: 12 fermiones divididos en dos grupos de 6 quarks y 6 leptones y 6 bosones divididos en 5 bosones de spin unidad y un bosón escalar: el bosón de Higgs.

1.2. Problemas abiertos en el modelo estándar

Todas las mediciones experimentales realizadas hasta la fecha no difieren de manera significativa respecto a las previstas por el modelo estándar de partículas. Aun así todavía falta un elevado porcentaje de comprobaciones experimentales por realizar que podrían arrojar desviaciones significas relativas al modelo estándar o cualquier resultado inesperado y en cualquier caso a día de hoy existen varias cuestiones que el modelo estándar no es capaz de explicar por si solo como son las siguientes:

1. La gravedad no esta incluida en el modelo estándar: No existe aun una teoría cuántica desarrollada completamente para la gravedad. Su débil fuerza intrínseca, considerablemente inferior al resto de interacciones permiten que la gravedad se pueda despreciar en los experimentos de física de partículas.
2. Materia y energía oscura: Las observaciones de las curvas de rotación de las galaxias en función de su distancia infieren que existe una gran cantidad de materia en los halos galácticos. Sin embargo esta materia no parece emitir o absorber radiación (materia oscura). Diversas teorías intentan explicar la naturaleza de esta materia, como que esté formada por partículas fuera del modelo estándar.
3. Prevalencia de materia sobre antimateria: La conjunción de carga C y la paridad P se violan en la interacción débil y también lo hace levemente su producto CP. Sin embargo esta leve violación no explica la prevalencia, ordenes de magnitud mayor, de la materia sobre la antimateria.
4. Unificación de la fuerza electrodébil y la fuerza fuerte: Al igual que la fuerza electromagnética y la débil están unificadas cabe pensar que la fuerza fuerte también lo pueda a estar. Sin embargo el modelo estándar no lo considera y la unificación en cualquier caso sería

a energías muy superiores a las logradas actualmente en cualquier experimento de física de partículas.

Nuevas teorías tratan de explicar estas y otras cuestiones relativas al modelo estándar, como lo son las Teorías de Gran Unificación, que acoplan las tres fuerzas fundamentales y que requieren además para funcionar la supersimetría o la teoría de cuerdas, que considera que las partículas del modelo estándar no son mas que estados vibracionales diferentes de un elemento básico: una cuerda. No obstante sigue sin haber actualmente ninguna evidencia experimental para estas teorías.

1.3. Gran Colisionador de Hadrones (LHC)

La mayor parte de los experimentos en el campo de física de partículas requieren energías muy elevadas (o longitudes de onda muy cortas) para producir partículas de especial interés como el quark t, el bosón de Higgs o las predichas en teorías mas allá del modelo estándar. Para lograr energías altas en el sistema del centro de masas que puedan dar lugar a dichas partículas se utilizan aceleradores de partículas. Por otra parte la detección de las partículas producidas requiere de detectores especializados que exploten los fundamentos físicos de la interacción de estas partículas con la materia.

El CERN (Organización Europea para la Investigación Nuclear) dedica la mayor parte de su infraestructura a la investigación en el campo de la física de partículas. Ubicado en la frontera entre Suiza y Francia alberga el acelerador LHC (Gran Colisionador de Hadrones) y el detector CMS (Solenoide Compacto de Muones), dispositivos usados para generar y recopilar los datos que se han empleado en este TFG.

El Gran Colisionador de Hadrones - Large Hadron Collider, LHC - es el mayor acelerador de partículas construido. Inicia su funcionamiento en el año 2008 y para su construcción se aprovecho parte de la entonces existente infraestructura del acelerador LEP (Gran Colisionador de Electrones y Positrones) [4]. Se ubica a unos 100 metros bajo tierra en la frontera entre Suiza y Francia y junto a la ciudad de Ginebra. Tiene un recorrido circular de 27 km en el cual paquetes de protones son acelerados actualmente a energías de $\sqrt{s} = 13,6$ TeV medidos en el sistema de referencia de su centro de masas.

Los paquetes de protones son generados y acelerados inicialmente por otros sistemas y aceleradores del complejo de aceleradores del CERN para ser posteriormente inyectados en el LHC (Figura 2). Los protones después viajan en tubos paralelos del LHC en los cuales se ha logrado hacer el vacío con una gran precisión. Son guiados en su recorrido circular por un intenso campo magnético logrado mediante electroimanes superconductores enfriados a una temperatura de -271.3^{o}C con helio líquido principalmente [5]. Los protones en su recorrido circular en el LHC alcanzan una frecuencia de $\sim$ 11000 revoluciones por segundo.

Los paquetes de protones se encuentran y colisionan en varios puntos del recorrido del LHC donde se encuentran los diferentes detectores. En el segundo ciclo operacional del LHC cada paquete de protones incluía $\sim 10^{11}$ protones y cada par de paquetes colisionaba en los detectores cada 25 nanosegundos. La luminosidad integrada alcanzada en todo el ciclo fue de

160 fb^{-1} [6] y la energía en el centro de masas de $\sqrt{s} = 13$ TeV (En 2018, año del cual se utilizan los datos en este TFG).

Los cuatro detectores mas relevantes son los siguientes: El detector CMS, el detector ATLAS (Aparato Toroidal del LHC), ALICE y LHcB. Tanto CMS como ATLAS son detectores generales y aunque emplean materiales diferentes comparten una estructura similar que permite la comparación de los resultados obtenidos por ambos detectores. LHcB y ALICE están enfocados a propósitos mas concretos, el primero en el estudio de la diferencia entre la cantidad de materia y antimateria en el universo a través del estudio del decaimiento de mesones B y el segundo en el estudio del plasma quark-gluón y las interacciones fuertes en energías y densidades altas de materia.

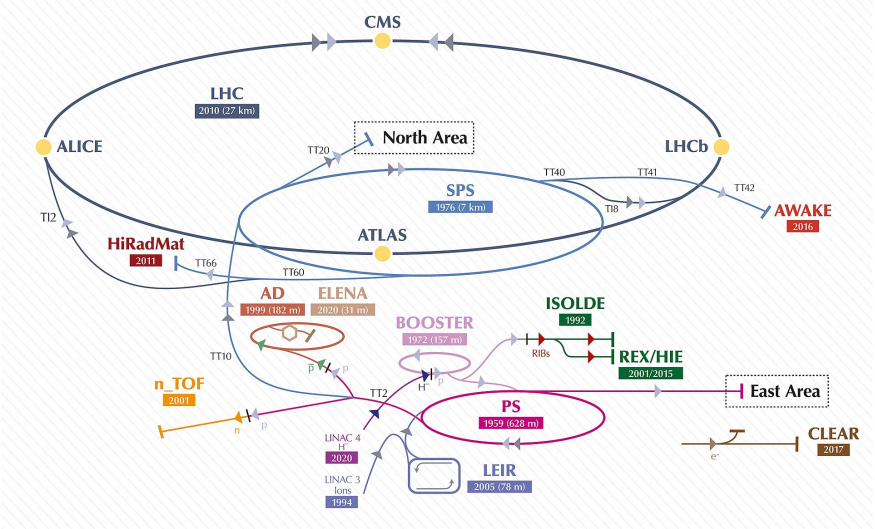

Figura 2: Complejo de aceleradores del CERN, incluyendo el LHC y la ubicación de sus detectores mas relevantes.

El LHC, en concreto las colaboraciones ATLAS y CMS, están detrás de uno de los más recientes y mayores hitos en la F. de Partículas: el descubrimiento experimental del bosón de Higgs [2][3], anunciado el 4 de julio de 2012. Era la ultima partícula fundamental que faltaba por observar del modelo estándar y además se determinó por primera vez su masa (m=125 GeV). Su producción puede darse en las colisiones p-p a través de diferentes mecanismos, como la fusión de gluones o de pares de bosones vectoriales. Debido a su reducido tiempo de semivida no se puede detectar de manera directa si no mediante la detección de los productos finales de su desintegración. En concreto el decaimiento preferente del bosón de Higgs es en pares de quarks b (Figura 3).

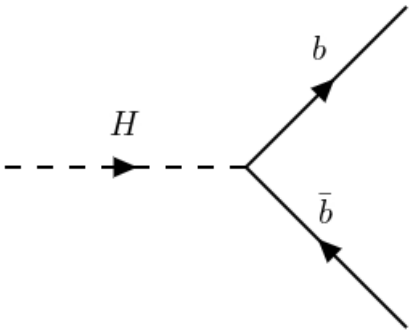

Figura 3: Diagrama de Feynman del decaimiento del bosón de Higgs en un un par quark-antiquark b.

1.4. Importancia de los b-jets y los algoritmos de b-tagging

El bosón de Higgs tiene un canal de desintegración en pares b. Debido a los procesos de hadronización dichos quarks, integrados en una partícula compuesta por el confinamiento de color, aparecen junto a otro grupo de bariones: se observan jets de partículas. La correcta identificación de jets de partículas asociados a un quark b y no generados por la hadronización de otros quarks de sabores mas ligeros (el quark t no se considera ya que su tiempo de semivida es tan corto que decae antes de que pueda dar lugar el proceso de hadronización) fue fundamental para su correcta detección. Para ello se usan algoritmos de alto nivel de etiquetado de b-jets (b-tagging) que explotan las características propias del quark b para su reconocimiento.

Actualmente los algoritmos de b-tagging siguen jugando un importante papel en análisis que involucran el bosón de Higgs como la producción de pares de Higgs [7] para la medición de el acoplamiento consigo mismo o el mecanismo de producción del bosón con un par de quarks top para la determinación del acoplamiento con dicho quark [8]. Los algoritmos también se emplean en la búsqueda de física mas allá del modelo estándar, como la supersimetría, en la cual se predice que partículas supersimétricas puedan decaer en pares de quarks b [9]. Su uso también es usado para la eliminación de fondo de procesos que no involucran b-jets.

2. Dispositivo experimental: CMS

2.1. Solenoide Compacto de Muones (CMS)

El CMS - Solenoide Compacto de Muones - es uno de los cuatro detectores situados en el LHC. Es un detector de propósito general que se encuentra cercano a la localidad de Cessy y en él colisionan los haces de protones acelerados en sentidos contrarios y con una energía en el sistema del centro de masas de 13 TeV en el Run 2 del LHC.

Su forma es cilíndrica y tiene unas dimensiones de 21 metros de largo y 15 metros de diámetro [10]. Dispone de diferentes capas concéntricas diseñadas para detectar las partículas mediante diferentes tipos de técnicas y detectores. Solo ciertas partículas pueden detectarse de manera directa mientras que el resto se detectan posteriormente de forma indirecta una vez se han reconstruido los datos obtenidos en el CMS. Las capas están diseñadas de forma específica para soportar altos niveles de radiación que no dañen los dispositivos utilizados y ofrecer una respuesta electrónica rápida. El detector esta divido en dos partes diferenciadas: las tapas y el barril. Las capas, ordenadas de interiores a exteriores son las siguientes: tracker, calorímetro electromagnético, calorímetro hadrónico, solenoide y cámaras de muones y aislamiento de hierro (Figura 4). El CMS cuenta con un complejo sistema electrónico integrado en las diferentes capas, el trigger, que permite filtrar los eventos detectados.

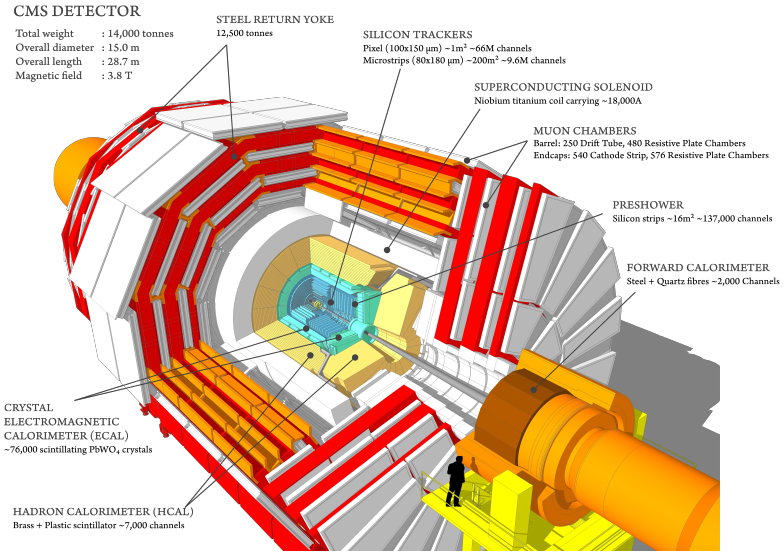

Figura 4: Estructura del CMS

1. Tracker:

La capa mas interna del CMS es el detector de trazas. Es capaz de identificar las trayectorias de las partículas cargadas que se producen en las interacciones p-p. Está cubierta de pixels y microstrips de silicio que actúan como detectores de estado sólido, explotando la pérdida por ionización de las partículas incidentes. En la parte mas interna del tracker - donde se lidia con la mayor densidad de radiación - se tienen los pixels de silicio. Hay aproximadamente unos 124 millones de ellos y tienen unas dimensiones muy reducidas de 100x150 μm [11].

Rodeando a los pixels y hasta una distancia de un metro desde el centro del CMS se tienen los strips de silicio. Cuando una partícula cargada atraviesa los pixels o strips de silicio se registra un hit y a partir de ellos se reconstruyen las trayectorias de las partículas cargadas.

2. Calorímetro electromagnético:

El calorímetro electromagnético (ECal) es la segunda capa mas interna del CMS. Los calorímetros permiten medir la energía y dirección de las partículas incidentes por absorción total. En el calorímetro electromagnético se detectan principalmente los electrones y fotones incidentes y son detenidos por completo. El ECal es un calorímetro homogéneo que usa de material cristales de tungstato de plomo, el cual sirve a su vez para que los fotones y electrones que lo atraviesen produzcan una cascada electromagnética y como centelleador. La luz emitida es proporcional a la energía de la partícula incidente y se detecta con fotodetectores. Con ellos se genera una señal eléctrica que se amplifica y se analiza. El ECal esta armado en el barril central donde se encuentran la mayor parte de los cristales de tungstato de plomo y en dos tapas entre el tracker y el calorímetro hadrónico con un número menor de cristales. Previo a las tapas del ECal se tiene además un detector basado en strips de silicio y un absorbente de plomo. Este detector permite distinguir entre fotones de alta energía y aquellos de baja energía producidos en la cascada electromagnética en los cristales de tungstato de plomo.

3. Calorímetro hadrónico:

En el calorímetro hadrónico (HCal) los hadrones incidentes producen una cadena de reacciones a través de interacciones inelásticas produciendo una variedad de partículas. La cascada de partículas producida tiene una componente electromagnética y otra puramente hadrónica. El HCal se divide en cuatro secciones: Hadron Barrel (HB), Hadron Outer (HO), Hadron Endcap (HE) y Hadron Forward (HF). A diferencia del calorímetro electromagnético, que era homogéneo, las secciones HE y HB son calorímetros inhomogéneos: Se tienen placas anchas de latón, que funcionan como material activo - el material usado es mas denso que el del ECal para poder detener por completo los hadrones incidentes - interconectados por centelleadores de plástico (estos conectados a los fotodetectores). La sección HO está instalada para detectar correctamente las cascadas de hadrones con una distancia de interacción larga que han atravesado los calorímetros ECal y las secciones del HCal HE y HB. La sección HF (calorímetro de hierro y cuarzo) esta instalada para poder cubrir la medida del momento transverso de las partículas en toda la geometría del HCal.

4. Solenoide:

Para lograr el intenso y uniforme campo magnético de 4 Teslas que permite conseguir una buena resolución en el momento de las partículas detectadas en las capas interiores se tiene un potente solenoide. En sus fibras superconductoras, enfriadas a 268,5ºC, circula una corriente de 18500 amperios [12]. Es el solenoide superconductor mas grande y potente construido. La medición de momento se obtiene directamente de la medida de la curvatura que adquieren las partículas cargadas bajo el campo magnético.

5. Cámara de muones:

Los muones juegan un importante papel en los análisis de física de partículas ya que se encuentran habitualmente entre los productos finales del decaimiento de numerosas partículas interesantes. Sin embargo su detección es complicada: pierden muy poca energía por unidad de distancia recorrida y atraviesan todas las capas previas. Su detección precisa requiere de un nuevo sistema: las cámaras de muones. Los muones son las únicas partículas que llegan a la capa mas externa del CMS (además de los neutrinos, que no se detectarán tampoco en las cámaras de muones y cuya presencia se infiere por el momento transversal total no encontrado) por lo que a diferencia del resto de capas la señal producida en esta se deberá únicamente a este tipo de partícula. La capa se articula de la siguiente manera: Se tienen cuatro estaciones de muones (MS1, MS2, MS3 y MS4) separadas por tres capas de hierro (que guían el campo magnético del solenoide mientras que funcionan también como filtro). En el barril central de la capa se localizan múltiples tubos de deriva mientras que en las tapas de cierre se tienen cámaras de tiras catódicas. También se realiza la identificación de la trayectoria simultáneamente con cámaras de gas multiplicadores de electrones (en la parte frontal de las cámaras de muones) y cámaras de láminas resistentes (en el barril central y las tapas). Los datos se combinan con los recolectados por los pixels y strips de silicio para la reconstrucción completa de la trayectoria de los muones desde su generación.

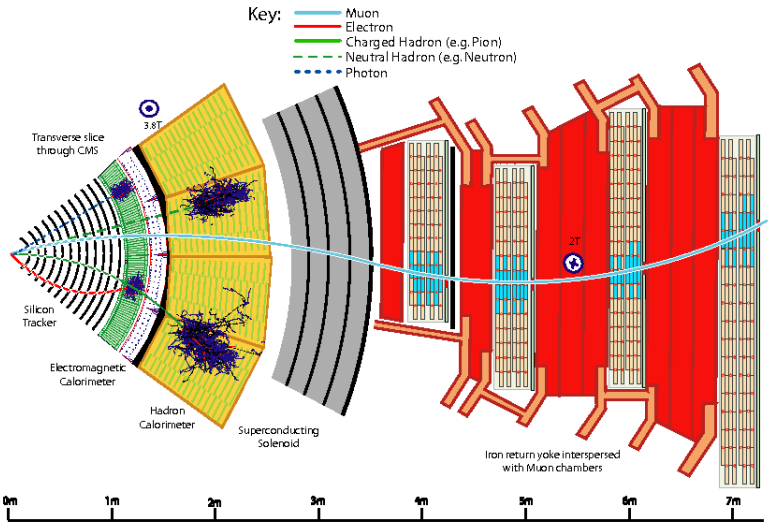

Figura 5: Corte transversal del CMS donde se indica las trayectorias típicas de varios tipos de partículas en sus diferentes capas.

2.2. Reconstrucción y filtrado de datos

El CMS tiene que lidiar en sus diferentes capas con un alto flujo de partículas y produce un tamaño considerable de datos en tiempos muy reducidos. Trabajar con semejantes datos iniciales y lograr unos datos tratables para los estudios a realizar y que sean de una calidad óptima es un problema complejo que se afronta en diferentes etapas.

El tratado de datos se realiza en pasos sucesivos, de los cuales muchos de ellos se hacen

de forma centralizada en las propias instalaciones del CMS. Se tiene un sistema electrónico complejo integrado en el CMS para hacer un filtrado inicial de los eventos interesantes detectados (trigger), dichos eventos son reconstruidos almacenando los objetos (partículas, jets de partículas, propiedades del detector) y sus variables (momento, energía... etc) y posteriormente se transforman los sets de datos eliminando propiedades de poco interés (de formato RECO a miniAOD) y son distribuidos.

2.2.1. Filtrado de eventos: triggers

El Trigger es un complejo sistema electrónico, de gran rapidez de procesamiento, integrado en el CMS que permite hacer un filtrado inicial de los eventos ocurridos en el detector. De esta forma eventos como una simple dispersión elástica de los protones son desechados. Está dividido en dos niveles: el trigger de nivel 1 (L1) y el High Level Trigger (HLT).

De los eventos producidos por segundo, alrededor de un billón, el trigger de nivel 1 (L1) selecciona unos 100.000 eventos atendiendo a que sean de la suficiente alta energía o de características inusuales [13]. Esta selección se realiza mediante las medidas recopiladas de dichos eventos en las capas ECal y HCal y las cámaras de muones sin necesidad de reconstruir el evento al completo en el detector.

La selección se envía al High Level Trigger (HLT) donde se sincroniza la información recolectada de las diferentes partes del detector y se hace una reconstrucción simple y rápida del evento. Se analiza de forma veloz las características de dichos procesos y se seleccionan alrededor de 100 eventos de los 100000 que hay en un segundo en el HLT [13].

No toda la sección de datos del trigger en su conjunto es guardada ya que sigue habiendo presente datos de escaso interés. Sin embargo los datos tienen un tamaño que permiten ser almacenados y tratados hasta una selección menor posterior.

2.2.2. Reconstrucción de eventos

Los datos seleccionados por el trigger, tras salir tanto de L1 como de HLT tienen el formato RAW. Tiene un tamaño de 0.70-0.75 Mb lo cual ocasiona un elevado peso en datos debido al amplio número de eventos seleccionados que se generan. Guarda los datos recopilados por el detector CMS y el evento no se encuentra aún reconstruido.

Posteriormente la información del evento se reconstruye. La información de los objetos del evento se almacena en el formato RECO. Los objetos son la identidad de las partículas detectadas, sus trayectorias, los jets de partículas en conjunto, los vértices reconstruidos... etc. Tiene un tamaño mayor, de 1.3-1.4 Mb por evento, fruto de las variables adicionales añadidas en la reconstrucción del evento. El formato RECO es distribuido y se usa en algunos pocos análisis de física por contener información necesaria que se elimina en formatos posteriores.

El AOD es un set de menor tamaño del formato RECO. Se eliminan objetos y variables asociadas del formato RECO, permitiendo una reducción de su tamaño a 0.05 Mb por evento. El formato AOD es distribuido y se usó mayoritariamente en el primer ciclo operacional

del LHC (entonces no existían formatos de menor tamaño). En este TFG se parte del formato miniAOD, un set aun menor del formato AOD, con un tamaño más pequeño que este. Es el formato usado por el grupo de b-tagging del CMS durante el Run 2 del LHC. El formato miniAOD a pesar de su reducido tamaño contiene la información necesaria para su uso en la mayor parte de análisis de física. El uso de un nuevo formato, el formato nanoAOD con un tamaño aun menor, esta empezando a ser utilizado por el grupo de b-tagging en los análisis derivados del Run 3 del LHC.

2.2.3. Datos simulados (MC)

En este TFG se hace un uso extensivo de datos provenientes de simulaciones. Son datos artificiales que parten de las bases teóricas de la teoría cuántica de campos y que sirven a varios propósitos. Su principal uso es la comparación con los datos reales ya que permiten testar los modelos y teorías en los que se basan los datos simulados y comprobar la validez de estos.

Los datos simulados se generan en dos pasos bien diferenciados: la generación de los propios eventos producidos en la colisión de los protones y la simulación de la respuesta del detector CMS de las partículas producidas en dichos eventos. Es un proceso muy costoso a nivel computacional, en especial la modelización de la respuesta del detector.

Los eventos generados en las colisiones p-p se modelizan con ayuda de softwares como el programa Pythia. Se utilizan tanto datos puramente teóricos derivados de la teoría cuántica de campos como datos experimentales obtenidos en otros experimentos de física de partículas. Se utilizan técnicas de MonteCarlo - basados en generación de números aleatorios - para modelizar los eventos (por ello de ahora en adelante a los datos simulados también se les referirá como datos MC). A nivel de generación se presentan problemas como el hecho de que ciertos procesos, en especial aquellos que involucran la interacción fuerte, no están entendidos a nivel teórico y por tanto bien modelados en los datos simulados. Estos son, entre otros, la fracción de decaimiento de los hadrones B, el decaimiento de gluones a pares b... etc. El modelado de la respuesta del detector requiere de otros programas (GEANT 4 principalmente) y simula la interacción de las partículas generadas en los eventos con los diferentes materiales del detector CMS y también la señal que producen.

De esta manera en los datos MC se tienen tanto las lecturas que los eventos simulados producirían en el detector CMS como los datos a nivel de generación de dichos eventos simulados. Es decir, por ejemplo, se tiene la trayectoria real que se ha simulado de una partícula al atravesar las capas del CMS y también se tiene la lectura simulada que se obtendría de esa partícula mediante el uso del CMS (al igual que otras variables como su momento, identidad de la partícula, energía... etc). Los datos MC pasan un procesado similar a los datos reales a partir de las señales simuladas obtenidas hasta llegar a los formatos AOD o miniAOD, aunque en todo momento se guarda la información adicional propia de los MC.

3. Algoritmos de b-tagging

Los algoritmos de b-tagging se aplican a los jets de partículas reconstruidos y permiten distinguir aquellos que son originados por un quark b y aquellos originados por sabores mas ligeros - en adelante denominados cl -. Dado que mucha física exótica o de interés general involucra a los quarks b su correcta y eficiente identificación es imprescindible para realizar análisis y estudios viables. Todos ellos se basan en las características típicas del decaimiento de un quark b: tiempo de vuelo largo debido a su larga semivida, decaimiento semileptónico de forma directa o indirecta casi siempre en muones... etc.

La eficiencia de un algoritmo de b-tagging para identificar un jet asociado a un quark b en una muestra de datos determinada se define como:

$$\epsilon = \frac{\textit{número de jets reconstruidos identificados como b-jet}}{\textit{número de b-jets reconstruidos}} \tag{1}$$

Si se obtiene la eficiencia de un algoritmo de b-tagging tanto para datos reales como para datos simulados (MC) se puede definir un factor de escala SF:

$$SF = \frac{\epsilon_{\textit{datos reales}}}{\epsilon_{\textit{datos simulados}}} \tag{2}$$

El factor de escala permite la calibración de los algoritmos de b-tagging. Es una muestra de la inexactitud de los datos simulados indicando la necesidad de ajustar parámetros u otras características de los MonteCarlo para que reproduzcan fielmente los datos reales.

La determinación de los factores de escala y eficiencias se realizan en este TFG a partir de la consideración de diferentes técnicas de etiquetado de b-jets poco correlacionadas en una muestra de datos enriquecida en b-jets los cuales contienen un muón. A partir de ello se construye un sistema de ecuaciones con ocho incógnitas y ocho parámetros (factores de correlación) cuya resolución permite obtener los SF y eficiencias deseadas. Se explicara en mas detalle en la sección 4.2. Otros cálculos de las eficiencias de los algoritmos de b-tagging dependen de selecciones de datos concretos como el de eventos de producción de pares $t\bar{t}$ que decaen en una amplia mayoría de los casos en b-jets y que no dependen de factores no bien modelados como el gluon splitting [14].

La eficiencia y SF de todos los algoritmos dependen fuertemente del momento transversal p_t considerado del jet y de su pseudorapidez, coordenada espacial que relaciona el ángulo θ del jet respecto al eje longitudinal de la colisión p-p y que es definida como:

$$\eta = -ln[tan(\theta/2)] = \frac{1}{2}\frac{|p| + p_l}{|p| - p_l} \tag{3}$$

Por ello los datos de la eficiencia y SF se muestran en diferentes rangos de p_t. En concreto los considerados en la realización de este TFG son cinco: 20-30, 30-50, 50-70, 70-100 y 100-140 GeV.

A continuación se resumen brevemente las características físicas relevantes de los quarks

b y su importancia en el b-tagging así como varios algoritmos de b-tagging: el algoritmo CSV y CSVv2, por su importancia histórica y papel en el desarrollo de algoritmos posteriores, y los algoritmos DeepCSV y DeepJet por ser los usados actualmente y objeto de estudio en este TFG.

3.1. Características del quark b

El quark b tiene una masa constituyente de $m \simeq 4.5$ GeV [15]. Sus dos modos de decaimiento principales son puramente hadrónico ($b \to q\bar{q}q_\alpha$) o semileptónico ($b \to l^-\bar{\nu}_l q_\alpha$), con $q_\alpha = u, c$. Es un decaimiento mediado por la fuerza débil y por el boson W^- (en el caso del antiquark $\bar{b}$, se tiene $q_\alpha = \bar{u}, \bar{c}$ y esta mediado por el bosón W^+) Los diagramas de Feynman de estos procesos se indican en la figura 6.

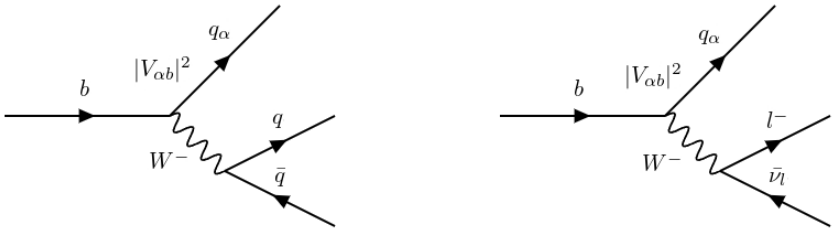

Figura 6: Diagramas de Feynman de los procesos $b \to q\bar{q}q_\alpha$ (izquierda) y $b \to l^-\bar{\nu}_l q_\alpha$ (derecha).

El acoplamiento al cuadrado del vértice bq_α es: $|g_{\alpha b}|^2 = |V_{\alpha b}|^2 g_W^2$. Los valores $|V_{\alpha b}|^2$ (indicados en la matriz CKM, que indica el grado de mezcla de los quarks d, s y b en la interacción débil) son muy reducidos - $|V_{ub}|^2 = (1{,}7 \pm 0{,}4)10^{-5}$ y $|V_{cb}|^2 = (1{,}7 \pm 0{,}1)10^{-5}$ [15] - y en consecuencia la semivida del quark b es considerablemente larga en comparación con el resto de quarks. La semivida de los hadrones B es de $\tau \simeq 10^{-12}s$ y debido a ello recorrerán distancias de r $\simeq c\tau \simeq 100\mu m$ antes de decaer, lo que facilita su detección (el hadrón en el que esta integrado deja un rastro detectable en el tracker del CMS).

Por culpa de la elevada masa del quark b y la considerablemente inferior masa de los productos de su decaimiento estos tienen un elevado momento transversal asociado a ellos el cual es habitualmente mayor al que tendrían si fuesen producto del decaimiento de otras partículas. La masa del b influye además en que los ángulos a los que se encuentran los productos de su decaimiento son mayores respecto al jet de partículas del quark b que si fuesen jets de quarks u, d, s y c. Por otro lado la masa invariante de las partículas asociadas al decaimiento del quark b es considerablemente mayor a la asociada al quark c y otros sabores mas ligeros. Por cada decaimiento de un quark b en un hadrón B se acaban produciendo alrededor de unas 5 partículas cargadas en su estado final, es decir, tienen una gran multiplicidad de partículas cargadas. El resto de sabores muestra una multiplicidad menor [16]. Además debido a su larga semivida las trazas de los b-jets tendrán un origen diferente y desplazado respecto a la mayor parte de trazas del resto de sucesos de los eventos.

Todas estas características fundamentales del decaimiento de un quark b se explotan en los algoritmos de b-tagging que se indican mas adelante.

3.2. Algoritmo CSV

El algoritmo de b-tagging CSV (Combined Secondary Vertex) se basa en la identificación del vértice primario del evento, en el cual se origina el quark b, y la posible identificación del vértice secundario, donde se produce el decaimiento del quark b (decaimiento del hadrón B) . Explotando las características del quark b (sección 3.1) se aplican unos criterios de selección que permiten discriminar los jets de partículas originados por un quark b respecto a otros. Fue usado durante el primer ciclo operacional del LHC (Run 1) [16].

Inicialmente se parte de la identificación de jets de partículas y de la localización de su vértice primario y se establece unos criterios de selección a dichos jets relacionados con su calidad de reconstrucción y sus variables dinámicas. De esta manera se hace un filtrado inicial de los eventos.

Posteriormente se trata de reconstruir el vértice secundario del evento. Para ello se utiliza otro algoritmo: Trimmed Kalman Vertex Finder [16]. De los vértices secundarios reconstruidos se hace una selección apropiada: que la distancia entre los vértices en el plano transversal este entre 100 μm y 2.5 cm, que dicha distancia dividida entre su error asociado sea mayor que 3, que la masa invariante de las partículas asociadas al vértice secundario no sea mayor a 6.5 GeV y que dicho vértice no sea compatible con el de un decaimiento de un kaón K_s^0 [16]. Si se ha encontrado un vértice secundario que cumple los criterios de selección mencionados dicho vértice se denomina RecoVertex. Si no se encuentra un candidato a vértice secundario se construye un vértice ficticio a partir de las trayectorias de partículas cargadas que no pertenecen al vértice primario, imponiendo un requisito en la significancia del parámetro de impacto transverso de la partícula respecto al jet. En caso de que ni un RecoVertex ni un PseudoVertex se encuentren se tiene un NoVertex.

En función de si se tiene un RecoVertex, un PseudoVertex o un NoVertex se utilizan diferentes variables para construir la función discriminatoria que permita distinguir entre b-jets y los debidos a otras partículas.

Para los RecoVertex se utilizan 6 variables:

a) La masa invariante de las partículas con carga no nula del vértice secundario.
b) La multiplicidad de partículas cargadas relacionadas con el RecoVertex.
c) La significancia del parámetro de impacto de aquellas trayectorias que cumplen el requisito de tener una masa invariante superior al quark c.
d) El cociente entre la energía de las partículas cargadas del vértice secundario y la de todas las partículas en el jet.
e) La rapidez de las partículas con carga del RecoVertex en relación a la dirección del jet que las contiene. Es definida para las partículas a partir de su energía y momento longitudinal respecto al jet como: $y = \frac{1}{2}\frac{E+p_{||}}{E-p_{||}}$

f) El cociente de la distancia entre los vértices primario y secundario en el plano transverso del CMS y su error. [16]

Los PseudoVertex utilizan solo las primeras cinco variables descritas y los NoVertex solo el parámetro de impacto de las trazas del jet.

El valor de las variables indicadas es significativamente diferente para jets asociados a quarks b, para los quarks c y para otros sabores mas ligeros debido a las características del quark b (sección 3.1) A cada variable x_i se le asocia una probabilidad mediante una función $f(x_i)$ de estar asociado a un b-jet, a un c-jet o a un jet de sabor mas ligero q y se multiplica por un factor dependiendo de si se tiene un RecoVertex, un PseudoVertex o un NoVertex. Se calcula el producto para todas las variables consideradas. Este valor es la función de probabilidad. Con esta función y definiendo la fracción de quarks c y quarks cl en jets no asociados a b se calcula la variable discriminante d para diferenciar los jets asociados a quarks b y quarks cl.

Los jets asociados a quarks b tienen valores de d próximos a la unidad mientras que los asociados a otros sabores tienen valores considerablemente inferiores. De esta manera se logra el objetivo de la identificación de b-jets.

Se tienen tres puntos operativos para el algoritmo (y los algoritmos sucesivos): tight T, medium M y loose L, definidos como los puntos en los que el algoritmo identifica erróneamente el 0.1, el 1 y el 10 % respectivamente de los jets asociados a sabores cl como un jet asociado a un quark b. El algoritmo se calibra en las diferentes campañas de datos para el ajuste a sus diferentes puntos operativos.

3.3. Algoritmo CSVv2

Es una versión mejorada del algoritmo CSV. Fue usado principalmente en los primeros años del Run 2 del LHC y ofrece una mejor eficiencia respecto al algoritmo CSV. Los principales cambios introducidos son los siguientes: 1) Se sustituye el algoritmo de reconstrucción de vértice secundario usado por el algoritmo CSV por el algoritmo Inclusive Vertex Finder por su mejor eficacia y rapidez en el procesamiento [17]. 2) Se introducen nuevas variables asociadas a las partículas y vértices detectados. 3) Se prescinde de la función de probabilidad (4) y en su lugar se usan redes neuronales artificiales para combinar las variables y construir la función discriminatoria [17].

3.4. Algoritmo DeepCSV

Su uso comenzó en la segunda mitad del segundo ciclo operacional del LHC y se extiende hasta la actualidad. Se construye en base al algoritmo original CSV y las mejoras introducidas por el algoritmo CSVv2 y otros algoritmos como JP y CMVAv2. Utiliza variables similares a las del algoritmo CSVv2 y un mayor número de trazas en su cálculo [18]. Al igual que el algoritmo CSVv2 tiene tres puntos operativos: t, l y m. En este caso se incluye el uso de técnicas de aprendizaje profundo en vez de la función de probabilidad de CSV y las redes neuronales simples de CSVv2 debido a su idoneidad para afrontar problemas complejos de clasificación como este caso. Los vértices primarios y secundarios se reconstruyen directa-

mente con el uso de las técnicas de aprendizaje profundo [19]. La capacidad de aprendizaje de estos algoritmos tiene menos limitaciones que las redes neuronales simples. El algoritmo esta entrenado dependiendo de su versión con entre 40 y 100 millones de datos de jets [19].

3.5. Algoritmo DeepJet

Al igual que el algoritmo DeepCSV se usan técnicas de aprendizaje profundo. Se tienen en este caso mas variables en cuenta, hasta 650 variables de entrada en total de partículas cargadas y neutras [20], en contraste con DeepCSV. Requiere mas capas para el tratamiento por separado de los datos asociados a los vértices y las partículas cargadas y neutras.

Es el algoritmo que ofrece una mejor eficiencia para todos los rangos de p_t y η [20] y su uso se prevee que sea extenso en el tercer ciclo operacional del CERN.

4. Análisis

La determinación de las eficiencias de los algoritmos de b-tagging se ha realizado con el framework de la colaboración CMS: CMSSW, el cual hace uso de los servidores del CERN, el sistema operativo LINUX, el lenguaje de programación C++ y el framework ROOT, desarrollado por el CERN y adecuado para el análisis de datos científicos a gran escala. Una explicación esquemática y con detalle del framework sobre el que se ha trabajado y su complejidad puede hallarse en la sección 7.1 Anexo.

A continuación se describen los análisis realizados: El tratamiento adicional de datos, una explicación detallada de la técnica usada en el cálculo de las eficiencias de los algoritmos de b-tagging - el System8 - los diferentes métodos de resolución del System8 empleados y un estudio de la incertidumbre generada debido a sus factores de correlación.

4.1. Tratamiento adicional de datos

En la sección 2.1 se ha indicado el tratamiento general que sufren los datos recolectados en el CMS hasta que están preparados para su distribución general. Los set de datos de los que se parte en este TFG son los pertenecientes al segundo ciclo operacional del LHC, año 2018 Ultra Legacy (2018UL) - su nombre se refiere a que los datos corresponden a la última reconstrucción de datos de dicho año realizada con la mejor precisión en las calibraciones del detector - en formato miniAOD para datos reales (12Nov2019_UL2018) y simulados (RunIISummer19UL18MiniAOD-106X_upgrade2018_realistic_v11_L1v1).

Estos datos distribuidos son de nuevo tratados, en esta ocasión a nivel local, para su transformación en sets de datos con formatos y estructura que faciliten su uso para la realización del estudio de las eficiencias de los algoritmos de b-tagging DeepCSV y DeepJet. Los pasos son los siguientes:

a) Generación de NTuplas.
El primer paso es la generación de NTuplas a partir de los archivos miniAOD. Se seleccionan eventos y objetos de manera que las NTuplas generadas estén enriquecidas en b-jets con muones. Se seleccionan eventos con dijets (con triggers de dijets) y los jets seleccionados tienen $p_t > 20$ GeV, $|\eta| < 2.4$ y al menos un muón entre sus partículas [21]. Las NTuplas contienen solo datos básicos de información y no complejos. Para su generación se ejecuta el módulo RecoBTag-Performance, previamente instalado en CMSSW, tanto para los archivos miniAOD de datos reales y de simulaciones.

b) Generación de Templates.
Una vez se han generado las NTuplas se procede a la generación de los Templates. Para ello se utiliza el modulo TemplateProduce, instalado e integrado previamente en el módulo RecoBTag-Performance de CMSSW. El código de TemplateProducer lee las NTuplas y aplica los criterios de selección - como el uso de un algoritmo de etiquetado de b-jets concreto - a los datos y guarda la información generada relevante para el cálculo de las eficiencias de los algoritmos de b-tagging.

4.2. System8

El método System8 se basa en que el uso de N criterios de selección independientes entre si da lugar a 2^N observables [22]. De esta forma se puede generar un sistema de 8 ecuaciones con 8 incógnitas que se puede resolver para conocer las variables que inicialmente eran desconocidas.

Aplicándolo a la determinación de la eficiencia de un algoritmo de b-tagging: se escoge como criterio de selección uno de los algoritmos de etiquetado de b-jets (DeepCSV o DeepJet) cuya eficiencia para determinar si es un quark b o un quark de sabor mas ligero es desconocida (variables a determinar) y se escogen otros dos criterios poco relacionados entre sí. Como dichos criterios de selección no son totalmente independientes hay que introducir ocho parámetros adicionales en el sistema de ecuaciones que den cuenta del grado de correlación entre los criterios: los factores de correlación. Si los criterios están poco correlacionados serán todos cercanos a la unidad.

Los tres criterios de selección usados para construir el System8 son los siguientes:

1. El algoritmo de etiquetado de b-jets a estudiar: DeepCSV o DeepJet (tag). Su introducción como criterio de selección para construir el System8 es lo que permite la determinación de su eficiencia.

2. Se utiliza como criterio de referencia (ref) un corte en la selección de jets requiriendo que el momento transversal del muón del jet de partículas sea al menos de 0.8 GeV. A dicho criterio se le nombrara en adelante con el sufijo p_T^{rel}.

3. Como criterio adicional de selección se requiere que el segundo jet seleccionado del suceso se halla etiquetado como b-jet y que se encuentre en un región aproximadamente opuesta al jet que se esta considerando. El uso o no uso de dicho criterio se refiere con la letra p y n respectivamente.

El sistema de ecuaciones del System8 entonces se escribe como:

$$
\begin{cases}
n = n_b + n_{cl} \\
p = p_b + p_{cl} \\
n^{tag} = \epsilon_b^{tag} n_b + \epsilon_{cl}^{tag} n_{cl} \\
p^{tag} = \beta \epsilon_b^{tag} p_b + \alpha \epsilon_{cl}^{tag} p_{cl} \\
n^{p_T^{rel}} = \epsilon_b^{p_T^{rel}} n_b + \epsilon_{cl}^{p_T^{rel}} n_{cl} \\
p^{p_T^{rel}} = \delta \epsilon_b^{p_T^{rel}} p_b + \gamma \epsilon_{cl}^{p_T^{rel}} p_{cl} \\
n^{tag,p_T^{rel}} = \kappa_b \epsilon_b^{tag} \epsilon_b^{p_T^{rel}} n_b + \kappa_{cl} \epsilon_{cl}^{tag} \epsilon_{cl}^{p_T^{rel}} n_{cl} \\
p^{tag,p_T^{rel}} = \kappa_b^{123} \beta\delta \epsilon_b^{tag} \epsilon_b^{p_T^{rel}} p_b + \kappa_{cl}^{123} \alpha\gamma \epsilon_{cl}^{tag} \epsilon_{cl}^{p_T^{rel}} p_{cl}
\end{cases}
\tag{4}
$$

El módulo TemplateProduce indicado en la anterior subsección proporciona el nº de jets que pasan los anteriores criterios de selección para todos los algoritmos y puntos operativos considerados, es decir, proporciona en un formato que se puede leer con facilidad los valores de las distintas regiones n, p, n^{tag}, p^{tag}, $n^{p_T^{rel}}$, $p^{p_T^{rel}}$, $n^{tag,p_T^{rel}}$ y $p^{tag,p_T^{rel}}$ (lado izquierdo de la ecuación).

Las incógnitas del sistema son: ϵ_b^{tag} y ϵ_{cl}^{tag}, las eficiencias del algoritmo de b-tagging aplicado a b-quarks (principal incógnita de interés) y cl-quarks respectivamente. $\epsilon_b^{p_T^{rel}}$ y $\epsilon_{cl}^{p_T^{rel}}$, las eficiencias del criterio de referencia. n_b y n_{cl}, el numero de b-jets y cl-jets respectivamente y p_b y p_{cl}, el numero de b-jets y cl-jets respectivamente que cumplen el criterio de selección tercero.

Por último los factores de correlación son: α, β, γ, δ, κ_{cl}, κ_b, κ_{cl}^{123} y κ_b^{123}. El método del System8 se basa en el hecho de que los factores de correlación son conocidos, siendo determinados a partir de los datos MC los cuales son capaces de reproducir los factores de correlación de los criterios usados.

4.2.1. Cálculo de los factores de correlación

Los factores de correlación del System8 - ecuación (4) - dan cuenta del grado de correlación entre los criterios seleccionados. Debido a que están poco relacionados son cercanos a la unidad. Están definidos como:

Dos parámetros, α y β, que muestran la correlación entre el algoritmo de estudio (tag) y el requisito (o falta de él) del hallazgo de un b-jet opuesto al estudiado para cl-jets y b-jets respectivamente:

$$\alpha = \frac{(\epsilon_{cl}^{tag})_p}{(\epsilon_{cl}^{tag})_n} \tag{5}$$

$$\beta = \frac{(\epsilon_{b}^{tag})_p}{(\epsilon_{b}^{tag})_n} \tag{6}$$

Dos parámetros, γ y δ, que muestran la correlación entre el criterio de referencia (p_T^{rel}) en las muestras n y p para cl-jets y b-jets respectivamente:

$$\gamma = \frac{(\epsilon_{cl}^{p_T^{rel}})_p}{(\epsilon_{cl}^{p_T^{rel}})_n} \tag{7}$$

$$\delta = \frac{(\epsilon_{b}^{p_T^{rel}})_p}{(\epsilon_{b}^{p_T^{rel}})_n} \tag{8}$$

Dos parámetros, κ_{cl} y κ_b, que muestran la correlación entre el algoritmo a estudiar (tag) y el criterio de referencia (p_T^{rel}) aplicados a la muestra n para cl-jets y b-jets respectivamente:

$$\kappa_{cl} = \frac{(\epsilon_{cl}^{tag,p_T^{rel}})_n}{(\epsilon_{cl}^{tag})_n(\epsilon_{cl}^{p_T^{rel}})_n} \tag{9}$$

$$\kappa_{b} = \frac{(\epsilon_{b}^{tag,p_T^{rel}})_n}{(\epsilon_{b}^{tag})_n(\epsilon_{b}^{p_T^{rel}})_n} \tag{10}$$

Dos parámetros, κ_{cl}^{123} y κ_b^{123}, definidos de forma análoga al caso anterior pero en este caso aplicados a la muestra p:

$$\kappa_{cl}^{123} = \frac{(\epsilon_{cl}^{tag,p_T^{rel}})_p}{(\epsilon_{cl}^{tag})_p(\epsilon_{cl}^{p_T^{rel}})_p} \tag{11}$$

$$\kappa_{b}^{123} = \frac{(\epsilon_{b}^{tag,p_T^{rel}})_p}{(\epsilon_{b}^{tag})_p(\epsilon_{b}^{p_T^{rel}})_p} \tag{12}$$

No es posible determinar de forma exacta los factores de correlación del System8 correspondientes a un set de datos reales dados ya que requiere el conocimiento previo de las eficiencias indicadas aplicadas a las muestras consideradas, tal y como se indican en las ecuaciones (5)-(12). Dichas eficiencias son incógnitas del propio System8.

Sin embargo si es posible determinar de forma directa mediante las ecuaciones (5)-(12) los factores de correlación correspondiente a un set de datos simulados. En ese caso las eficiencias se pueden determinar de forma directa mediante la aplicación de la ecuación (1) al ser conocido en los datos MC el nº de b-jets reconstruidos en las muestras consideradas.

Para la resolución del System8 con un set de datos reales se toman como factores de correlación los determinados a partir de datos simulados. Es de importancia recalcar que esta

elección es solo una aproximación. Los factores de correlación de ambos casos no son exactamente iguales aunque se estima que los factores de correlación de los datos reales son muy cercanos a los factores de correlación de los datos simulados. Tradicionalmente, en el caso de algoritmos más sencillos que no usaban tanta información, como el algoritmo CSV, se había verificado que esta aproximación era razonable. Convendría en el futuro estudiarlo para los nuevos algoritmos. Seria un estudio muy diferente al considerado en este TFG que involucraría encontrar zonas de datos puras... etc.

4.3. Método numérico

El método indicado bajo el nombre *numeric* es el único método plenamente integrado y funcional en el framework del System8 al inicio de la realización de este TFG. Es el método usado por defecto en la determinación de las eficiencias de los algoritmos de b-tagging.

El funcionamiento del método numérico es el siguiente:

Inicialmente se transforma los datos conocidos del System8 (n, p, n^{tag}, p^{tag}, n^{Ref}, p^{Ref}, $n^{tag,p_T^{rel}}$ y $p^{tag,p_T^{rel}}$) a un conjunto de datos ortogonales (Sección 4.5.1 Figura 14, Derecha). Es decir, a muestras de datos que no solapen y que sean independientes entre sí.

El sistema de ocho ecuaciones tiene ocho variables. Dicho sistema se recombina en una única ecuación de ocho incógnitas y posteriormente siete de las ocho variables se escriben en función de la restante. Su resolución se realiza mediante el método numérico de la bisección: se toman distintos valores de la incógnita y se calcula el valor de la función hasta que se obtiene un par de casos en los que el valor de la función tiene signos diferentes. Entonces se calcula el punto medio entre los valores de la variable y se obtiene el valor de la función. Se considera un nuevo intervalo correspondiente al último valor de la variable considerada y aquel anterior a este cuyo signo de la función es contrario a esté ultimo y se repite el proceso. Tras numerosas iteraciones y tras alcanzar la precisión considerada se toma el valor de la ultima iteración como la raiz de la ecuación: la solución de la variable tratada. El resto de las 7 variables se calcula de forma directa a partir de esta. Dicho proceso se repite en diferentes intervalos para localizar las diferentes soluciones del sistema.

Una vez localizadas las soluciones se comprueba que cumplan unos requisitos básicos, concretamente que la fracción de b-jets (n_b/n) no sea menor a 0.20 y que ninguna de las variables sea negativa. Si ninguna solución lo cumple el método finaliza sin indicar solución hallada. Si varias de ellas lo cumplen se elige habitualmente como solución correcta la que tiene una mayor fracción de b-jets (n_b/n) aunque en ese supuesto se requiere un análisis detallado de las posibles soluciones. Una vez determinada una solución correcta se procede a hacer un smearing gaussiano respecto al set de datos ortogonales introducido.

El smearing gaussiano se realiza para tener en cuenta las fluctuaciones estadísticas en el número de sucesos en las diferentes regiones consideradas. Para ello se varia el valor de cada uno de las ocho regiones de datos disconjuntos (de x a x') mediante:

$$x' = x + \sqrt{x} * RndmGauss(0, 1) \tag{13}$$

Donde RndmGauss(0,1) es un número generado de forma aleatoria con una probabilidad acorde a una función gaussiana estándar centrada en 0 y en un intervalo de sigma = 1. El valor generado aleatoriamente puede ser positivo o negativo.

A continuación se vuelve a repetir el proceso anterior y se vuelve a hallar las soluciones correspondiente al nuevo set de datos. El proceso se repite 1000 veces y posteriormente los resultados se ajustan a función gaussiana diferente para cada una de las incógnitas. La solución que proporciona el método numérico para las variables se corresponde al valor central de la gaussiana y el error considerado a su desviación estándar.

4.3.1. Recreación de resultados de la campaña de datos 2018UL

A continuación se presentan los resultados obtenidos por el método numérico con sus opciones predeterminadas - criterios de selección, factores de correlación... etc. - y sin realizar cambios en el código para la campaña de datos 2018UL.

Los resultados obtenidos, de forma gráfica, se pueden ver a continuación en la Figura 7 y 8. Los resultados numéricos se pueden ver en las tablas (2) a (7) en la sección 4.5.2. Son la recreación de los resultados presentados por el IFCA el día 28 de octubre de 2021 en la colaboración CMS. Los resultados se dibujan con su correspondiente error calculado a partir del smearing gaussiano realizado pero su valor es tan pequeño que no se aprecia en las gráficas.

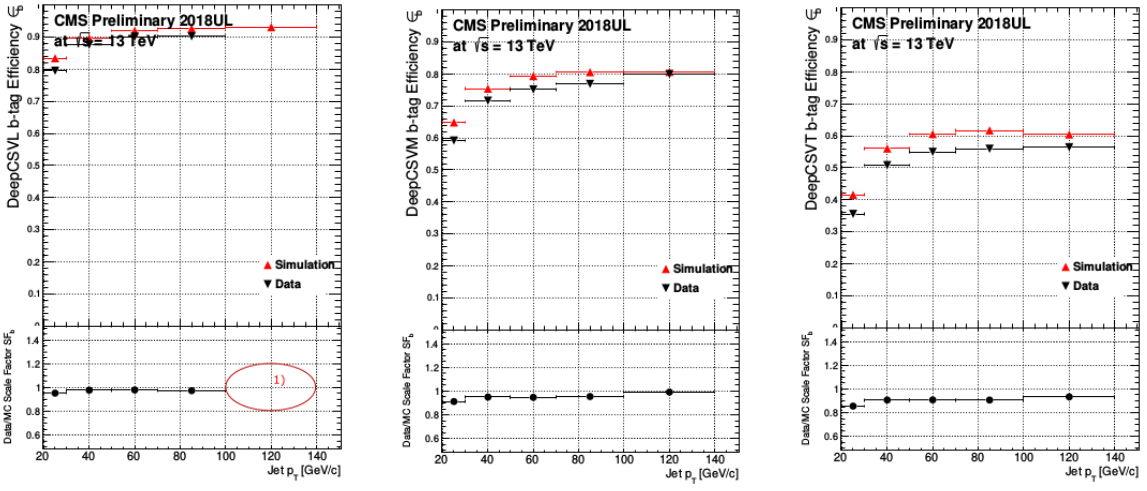

Figura 7: Eficiencias de los algoritmos de b-tagging (arriba) y scale factors (abajo) determinados mediante el método numérico. DeepCSVL (izquierda), DeepCSVM (centro) y DeepCSVT (derecha).

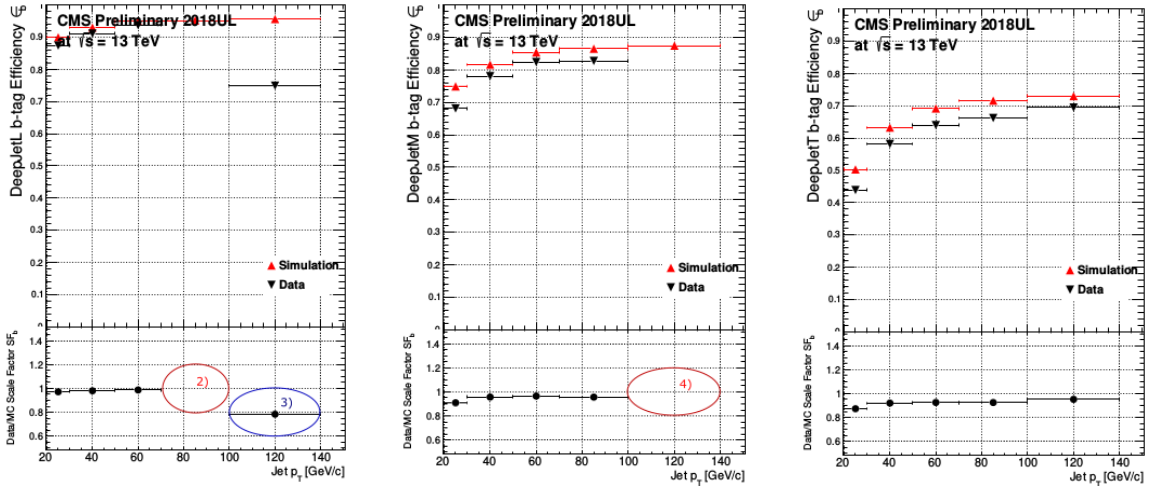

Figura 8: Eficiencias de los algoritmos de b-tagging (arriba) y scale factors (abajo) determinados mediante el método numérico. DeepJetL (izquierda), DeepJetM (centro) y DeepJetT (derecha).

Un análisis de los resultados permite observar que el método numérico no encuentra solución compatible con los criterios de selección (que las incognitas sean todas positivas y la fracción de b-jets superior a 0.20) para ciertos rangos de p_T de los jets en algunos puntos operativos de los algoritmos (Figuras 7 y 8, casos numerados 1., 2. y 4.) Es posible recuperar algunas de sus soluciones mediante la relajación del criterio de selección de la fracción de b-jets a 0.15:

Caso	**Algoritmo**	**Bin** (GeV)	**Eficiencia algoritmo**	**Fracción b-jets**
2)	DeepJetL	70-100	0.931 ± 0.004	0.20 ± 0.01
4)	DeepJetM	100-140	0.763 ± 0.008	0.181 ± 0.005

Tabla 1 Resultados recuperados del método numérico, campaña de datos 2018UL.

La solución de DeepJetL, bin: 70-100 GeV (caso 2), se recupera y da una fracción de b-jets mayor a 0.20. Inicialmente no se recuperaba porque la solución inicial era menor a 0.20 y por tanto no se hacia el smearing gaussiano. Al relajar el criterio de selección si se realiza y su resultado final es ligeramente mayor a 0.20. La solución de DeepJetM, bin: 100-140 GeV (caso 4), se recupera y da una fracción de b-jets menor a 0.20. La solución correspondiente a DeepCSVL, bin: 100-140 GeV (caso 1), no se puede recuperar tampoco con el nuevo criterio de selección. Las soluciones recuperadas son cuestionables, una fracción tan baja de b-jets comparada con la predicción MC (alrededor de 0.30) puede no ser del todo aceptable. Por otro lado la solución de DeepJetL, bin: 100-140 GeV (caso 3) a pesar de cumplir los criterios de selección iniciales de una fracción de b-jets mayor a 0.20 muestra problemas de consistencia: el SF es considerablemente mas bajo de lo previsto.

Por otro lado de la revisión de las soluciones halladas se observa, tal y como se esperaba, que los algoritmos DeepJet y DeepCSV muestran su mejor eficiencia en su punto operativo L - aunque ello es a expensas de tener un 10 % de jets cl que se asocian erróneamente a un b-jet

- que los puntos operativos M y T disminuyen considerablemente su eficiencia pero se evita que los cl-jets pasen el filtro de b-jets y que DeepJet muestra una ligera mejor eficiencia que DeepCSV.

4.3.2. Errores sistemáticos

Existen diversos errores sistemáticos en la determinación de las eficiencias de b-tagging y sus scale factors inherentes a cualquier método de resolución del System8. Dichos errores provienen de la modelización de los datos simulados, concretamente de factores que no se conocen con precisión y que involucran procesos donde interviene la fuerza fuerte, y de las características del detector CMS y su reconstrucción de datos. En concreto en el System8 se consideran, entre otros, los siguientes: la incertidumbre en la modelización del decaimiento de los gluones en pares de b-quarks (gluon splitting), el modelo de fragmentación del quark b y del quark c, la posibilidad de identificar erróneamente los productos del decaimiento de los quark b con los de K_s^0 y Λ, la incertidumbre en el momento transversal del muón del jet y de su ubicación relativa en él, el error en la escala de energía de los jets, el pileup en el detector CMS debido a las colisiones simultaneas p-p, la calibración en la resolución de los parámetros de impacto del jet... etc.

Para el calculo de dichos errores sistemáticos se dispone de datos MC generados con las variaciones sistemáticas consideradas. Su variación altera las eficiencias obtenidas para los datos MC y los factores de correlación determinados a partir de ellos y ya que en la resolución del System8 para datos reales se toman como factores de correlación los correspondientes a las simulaciones también alteran las eficiencias obtenidas para datos reales.

En la realización de este TFG se ha optado por no realizar un estudio de estos errores sistemáticos, ya documentados en los resultados oficiales presentados por el IFCA en la colaboración CMS para la campaña de datos 2018UL, y en su lugar realizar por primera vez en el IFCA un estudio preliminar de las potenciales incertidumbres en los factores de correlación y sus efectos no cubiertos por los errores sistemáticos mencionados en esta sección.

4.3.3. Dependencia de las soluciones con los factores de correlación (numérico)

Tal y como se ha indicado en la sección 4.2.1. los factores de correlación se obtienen a partir de los datos simulados MC mediante la ecuación (1) y las ecuaciones (7)-(14). Se procede al estudio de la dependencia de las soluciones del System8 mediante el método numérico con los factores de correlación. Para ello se hace uso de diferentes opciones no habilitadas inicialmente en el código. Por defecto se utilizan los factores de correlación obtenidos de la forma mencionada anteriormente, sin embargo en el propio código a partir de los valores de los factores de correlación de cada bin realiza internamente un ajuste para seis de los factores de correlación (α, β, δ, γ, κ_b y κ_{cl}) a una constante (fit0), a una regresión lineal (fit1) y a un polinomio de grado 2 (fit2) y existe la posibilidad de tomar como factores de correlación en cada bin para su resolución los valores correspondiente a la función ajustada en el centro de cada bin. La opción fit2 muestra una muy leve variación de los factores de correlación considerados respecto a su valor exacto MC y la opción fit0 una mayor diferencia. En la Figura 9 se puede observar el valor de los seis factores de correlación considerados y

sus ajustes a fit2, fit1 y fit0 para el algoritmo DeepJetL.

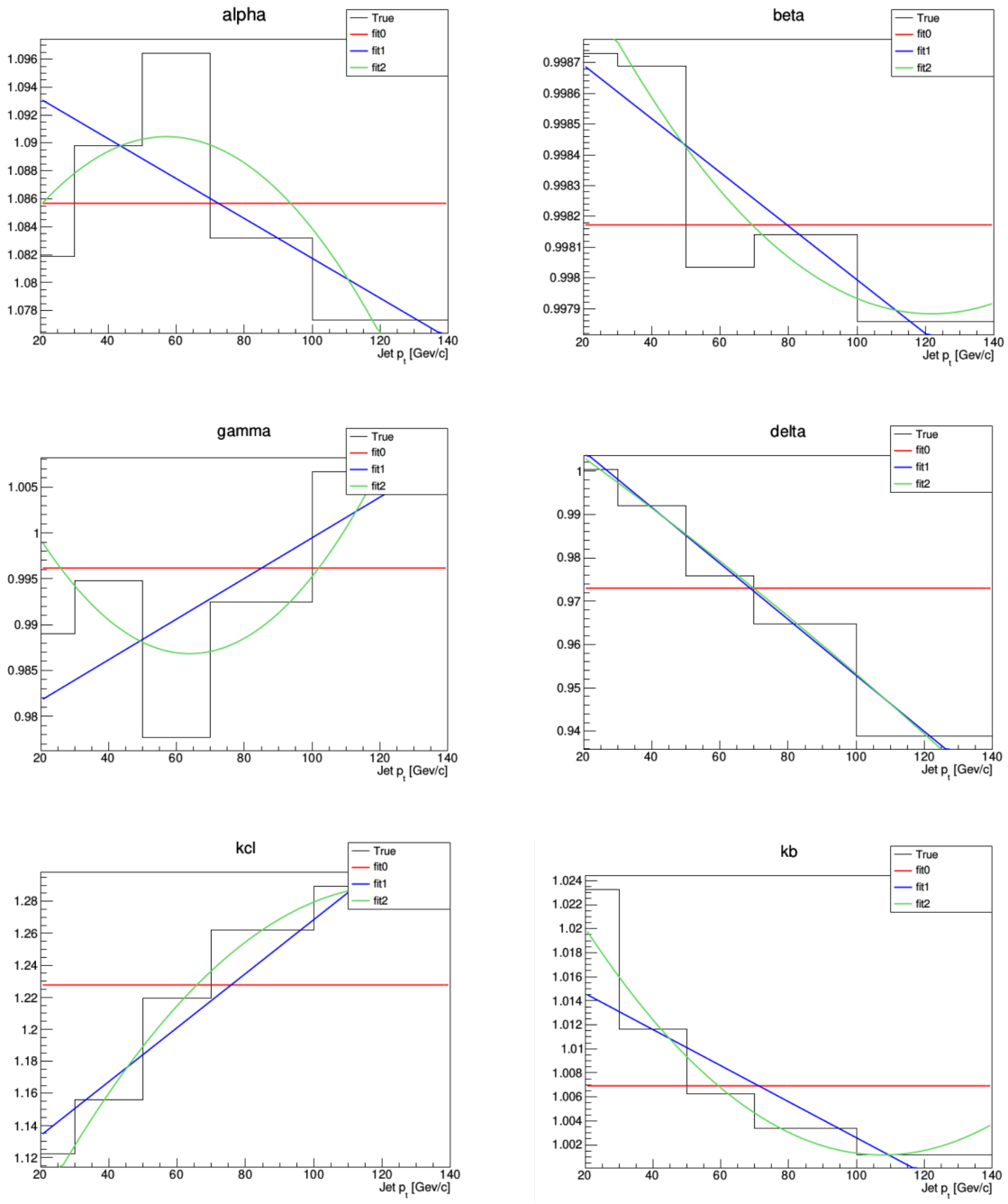

Figura 9: Factores de correlación para el algoritmo DeepJetL, ordenados de izquierda a derecha y de arriba a bajo: α, β, γ, δ, κ_{cl} y κ_b, y sus ajustes a fit2, fit1 y fit0.

Para algunos factores de correlación en ocasiones parece observarse tendencias con p_T del jet considerado, pudiendo resultar adecuada la toma del ajuste a fit1 o fit2, mientras que en otros casos parecen solo observarse fluctuaciones estadísticas con el ajuste a fit0. En cual-

quier caso, en ciertas situaciones no está muy claro cuál de los valores es más correcto y por tanto se puede decir que tenemos una indeterminación en los valores óptimos del MC en los factores de correlación. Revisando los valores numéricos y comparando las diferencias entre los valores true y los fits se observa que las variaciones típicas son de alrededor de un 1 %. Por ejemplo para el factor kb en el algoritmo DeepJetL la diferencia en el bin 20-30 GeV es de $\sim$ 1.6 %. A continuación se muestran las eficiencias y *scale factors* obtenidos para el algoritmo DeepJetL para los casos en los que el método numérico se utiliza con los valores exactos obtenidos por los datos simulados para cada bin de p_t (true), y con los valores obtenidos tras el ajuste para todos los bines a fit0, fit1 y fit2 (Figura 10). El algoritmo DeepJetL es uno de los casos más extremos de cambio en SF al usar true, pol0, pol1 o pol2.

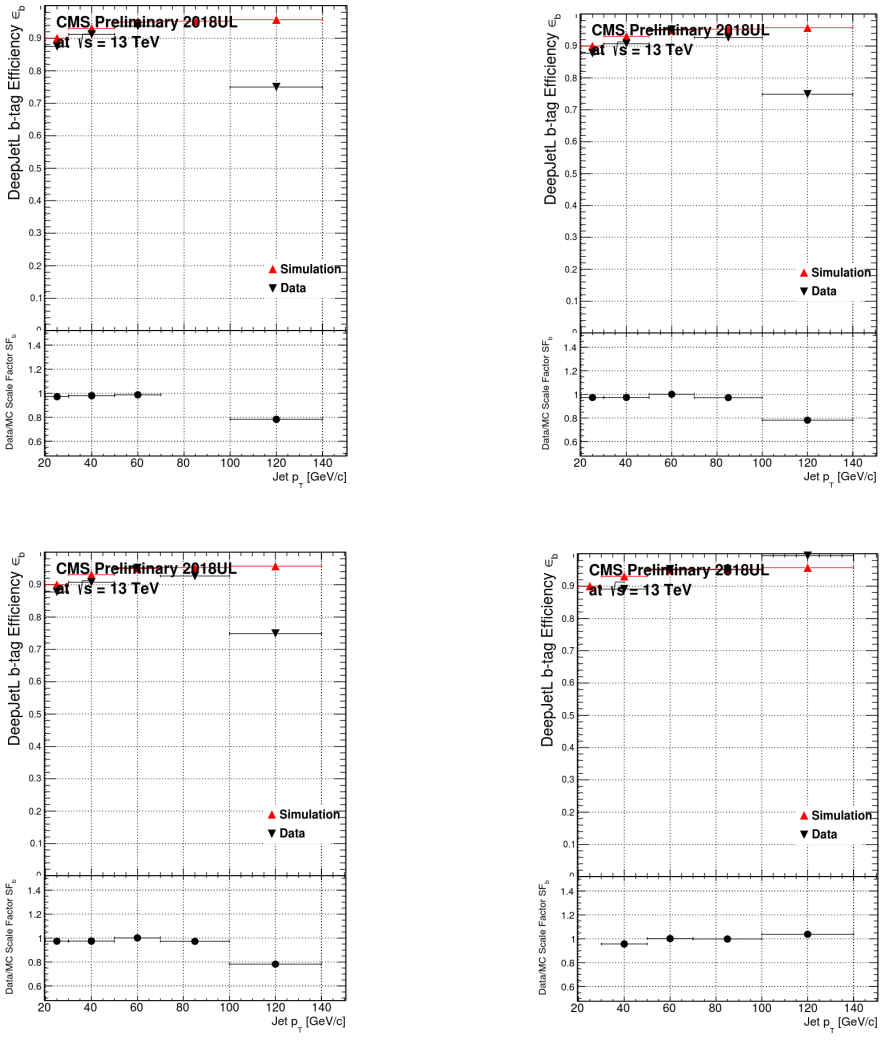

Figura 10: Eficiencias (arriba) y scale factors (abajo) de DeepJetL determinados mediante el método numérico. Ordenados de izquierda a derecha y de arriba a abajo: true, fit2, fit1 y fit0.

Se evidencia que las soluciones halladas muestran una fuerte dependencia con unas pequeñas

variaciones de los factores de correlación. Con fit2, fit1 y fit0 aparecen soluciones que cumplen los requisitos del selector de solución del método numérico y desaparecen otras soluciones que encontraba el método numérico con la opción true. En casos que todas las opciones encuentran solución algunas difieren considerablemente. En concreto para el algoritmo DeepJetL el uso de fit2, fit1 y fit0 permite recuperar la solución del bin de 70-100 GeV que no se encuentra usando los factores de correlación MC exactos mientras que por otro lado la opción fit0 pierde la solución del bin de 20-30 GeV que encuentra el resto de opciones y además da un resultado considerablemente diferente para el bin de 120-140 GeV.

La dependencia de las soluciones con los factores de correlación se evidencia y justifica mas claramente estudiando alternativamente la resolución del System8 a partir de los datos simulados MC. En ese caso, que usa los factores de correlación exactos MC, se obtiene que las soluciones del método numérico son exactamente las mismas a las predichas por MC por la ecuación (1) hasta una elevada precisión. El motivo es que él System8 que se resuelve es exactamente el correspondiente a los datos simulados y además el método es adecuado por lo que las soluciones obtenidas deben ser exactamente iguales a las predichas. Si se utiliza el método numérico a partir de los datos simulados MC pero a diferencia del caso anterior no se utilizan exactamente los valores de los factores de correlación predichos por MC si no unos cercanos no se estará resolviendo el System8 exacto que corresponde a los valores MC si no un aproximadamente parecido. La lejanía de las soluciones halladas con respecto de los valores dados por MC (o alternativamente la lejanía con respecto de un factor de escala unidad) da una mejor idea de la dependencia de las soluciones con los factores de correlación. Mediante el uso de esta opción se constata de nuevo la importante dependencia de las soluciones con los factores de correlación. Una comparativa de los resultados obtenidos de las eficiencias a partir de los datos MC si los factores de correlación usados son exactamente los correspondientes a MC (true) o si están ajustados a fit2, fit1 y fit0 respectivamente para el algoritmo DeepJetL se puede ver en la Figura 11 al final de esta subsección.

En el caso que se este trabajando con datos reales se tiene el problema de que los factores de correlación reales no se pueden hallar de forma directa. El tomar los valores MC, o el ajuste a un polinomio de grado 2, 1 o 0 es en cualquier caso una aproximación y no se tiene porque corresponder con los valores exactos de los factores de correlación del sistema. Los valores reales de los factores de correlación deberían ser cercanos a MC y fit2 si los Monte-Carlo están modelando bien los procesos físicos considerados, pero su desconocimiento con precisión hace que al considerar los datos reales no haya motivación adecuada asegurar como soluciones correctas las obtenidas con los factores de correlación de MC, de fit2, de fit1 o de fit0. En la sección 4.5.3, motivado por los hallazgos de esta sección, se realiza un nuevo estudio de la dependencia de las soluciones con mayor detalle mediante el método fit una vez esta calibrado y variando los factores de correlación entorno al $\sim 1\,\%$.

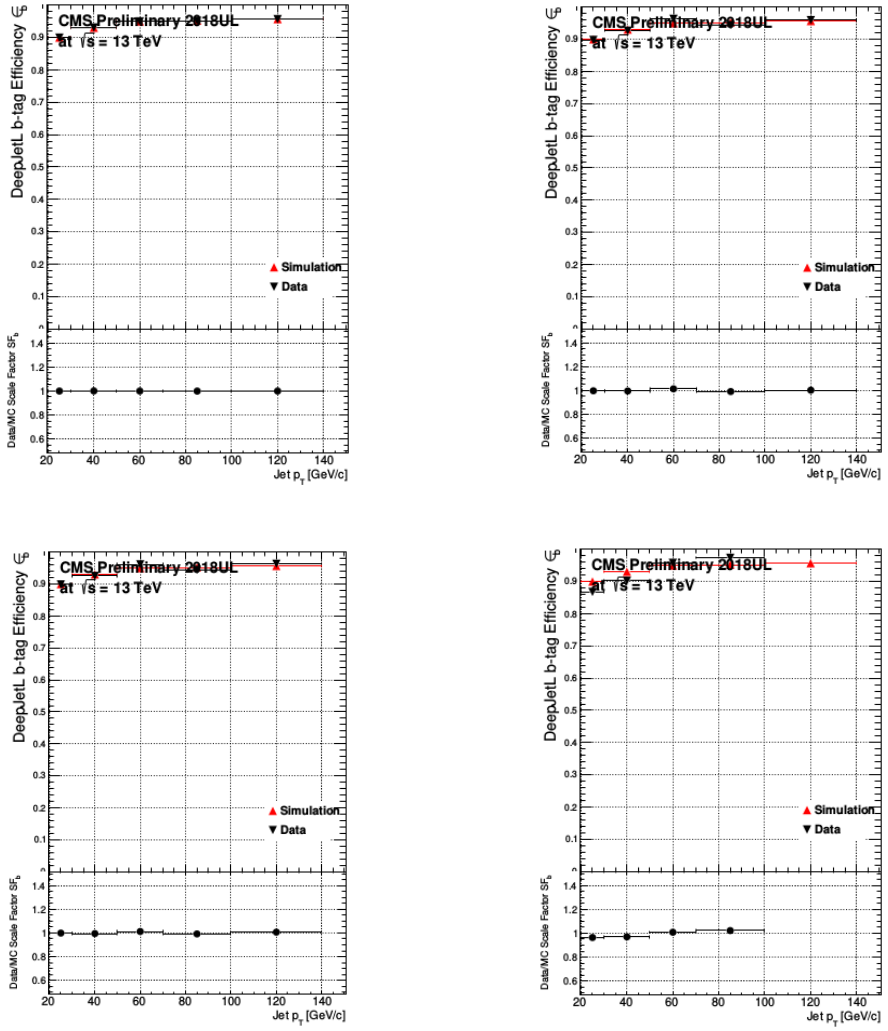

Figura 11: Eficiencias (arriba) y scale factors (abajo) de DeepJetL determinados mediante el método numérico. Ordenados de izquierda a derecha y de arriba a abajo: true, fit2, fit1 y fit0.

4.4. Método analítico

El método analítico no se encontraba plenamente integrado en el framework en el que se trabaja el System8. La opción que permitía su uso no exportaba ni gráficamente ni en formato texto las soluciones halladas. El estudio de este método ha requerido de una modificación del código donde se realiza el método nativamente para obtener las gráficas de las eficiencias de b-tagging y las soluciones numéricas de las ocho variables del System8.

El sistema de ecuaciones del System8 puede tener solución analítica. Sin embargo la solución analítica integrada en el método se corresponde al caso en el que se ha realizado la aproximación de que todos los factores de correlación son igual a la unidad. Como se muestra en la Figura 9 esta aproximación no es correcta (en los casos más extremos los factores de correlación llegan a alcanzar valores superiores de ~ 1.3 e inferiores de ~ 0.94). Además, tal

y como se muestra en las Figuras 10 y 11 incluso pequeñas desviaciones en el valor de los factores de correlación pueden tener un impacto a considerar en la solución del sistema.

Para cada una de las ochos variables se obtienen dos expresiones analíticas diferentes. Solo una de ellas se asocia a un resultado que tiene cierto sentido físico. Las soluciones descartadas indican valores de las muestras de b-jets anómalamente superiores y de las eficiencias de b-tagging considerablemente inferiores a los valores predichos por MC. Los resultados obtenidos tomando como datos para la resolución del System8 mediante el método analítico correspondientes a la campaña de datos 2018UL son: (Figura 12 y 13)

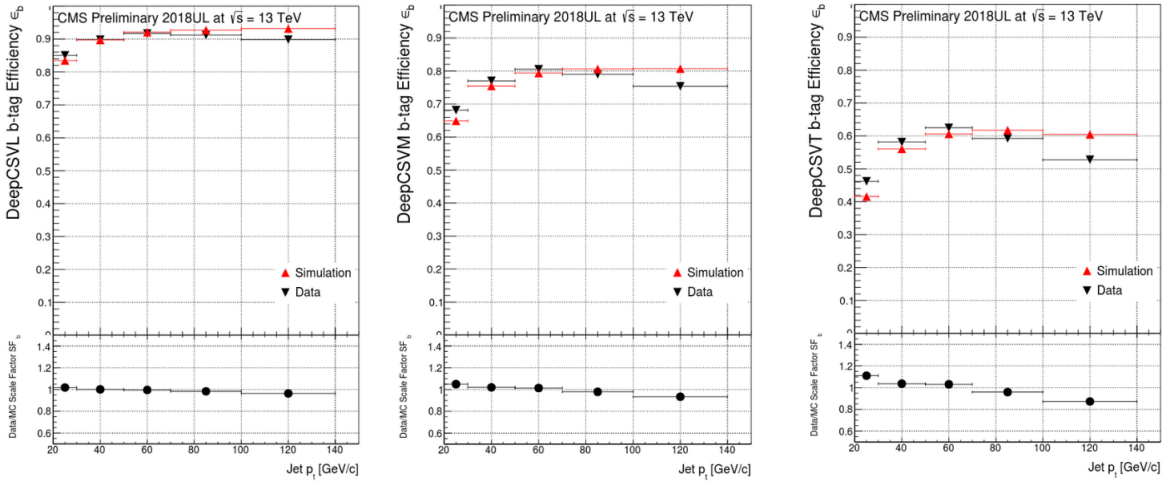

Figura 12: Eficiencias (arriba) y *Scale Factors* (Abajo) del algoritmo DeepCSV para la campaña de datos 2018 UL obtenidas mediante el método analítico en sus tres puntos operativos *loose* L (izquierda), *medium* M (centro) y *tight* T (derecha).

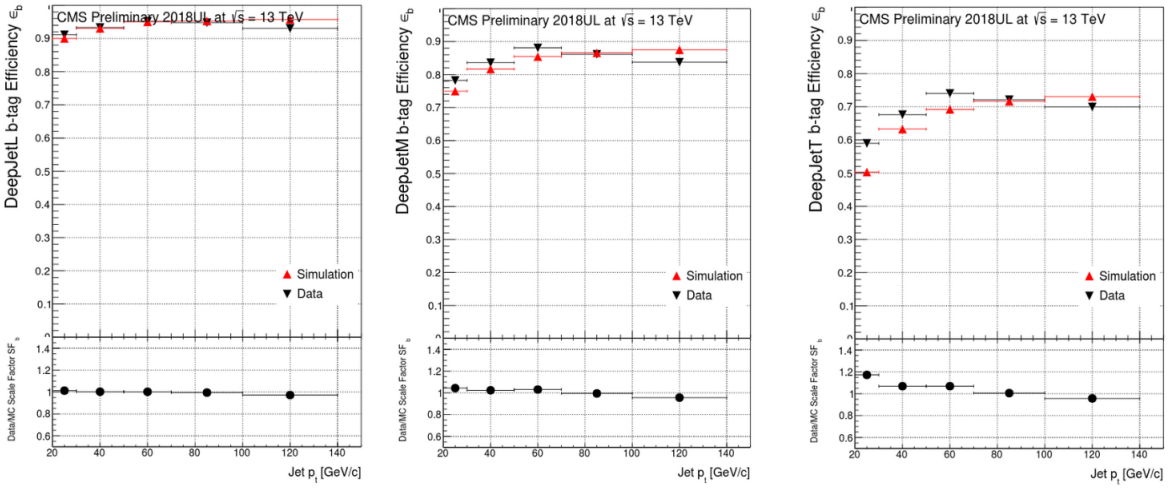

Figura 13: Eficiencias (arriba) y *Scale Factors* (Abajo) del algoritmo DeepJet para la campaña de datos 2018 UL obtenidas mediante el método analítico en sus tres puntos operativos *loose* L (izquierda), *medium* M (centro) y *tight* T (derecha).

Se encuentra solución para ambos algoritmos de b-tagging considerados y todos sus puntos operativos. Las soluciones halladas para las eficiencias son cercanas a los valores predichos por los datos MC. El análisis de la fracción de b-jets muestra un aumento con el aumento del bin de p_t considerado llegándose a obtenerse en el ultimo bin (p_t = 100-140 GeV) fracciones de b-jets de $\sim 50\,\%$ en vez del $\sim 30\,\%$ indicado por MC. Este hecho discordante con la predicción de los datos simulados y que además que para ciertos bines se encuentra que el valor de varias incógnitas es negativo son otros indicadores de que la aproximación de que los factores de correlación son igual a la unidad para la resolución del System8 no es adecuada.

La no viabilidad de la aproximación empleada se confirma por otra vía utilizando el método analítico a partir de los datos simulados MC en vez de a partir de los datos reales. Si la aproximación fuese adecuada se obtendrían resultados iguales o suficientemente aproximados a los obtenidos por MC, tal y como ha ocurrido en el método numérico al usar los datos MC. Sin embargo los resultados obtenidos de esta forma son algo distantes a los valores predichos MC - se tiene la paradójica situación de que se tienen SF mas lejanos de la unidad que utilizando el método analítico con los datos reales - y en cualquier caso se sigue observando la importante discrepancia en la fracción de b-jets hallada. La aproximación en la que se indican todos los factores de correlación a la unidad por lo tanto no es correcta y muestra de nuevo que las soluciones encontradas del System8 son muy sensibles al valor de dichos factores.

Esta discusión sobre los factores de correlación motivada por la introducción del método analítico no nos aporta información sobre la validez del método en sí. Para estudiarla se comprueba que las soluciones se ha comprobado que las soluciones que encuentra el método analítico son correctas en el caso de que los factores de correlación sean todos iguales a la unidad mediante la resolución del System8 con el método numérico y el método fit (una vez se ha integrado correctamente en el framework del System8. Sección 4.5) forzando que sus parámetros de correlación sean también iguales a la unidad. En dicho caso las soluciones con los tres métodos coinciden. Muestras además la buena implementación de los tres métodos. En base a ello se propone en un futuro obtener una resolución analítica del System8 en función de los parámetros de los factores de correlación. Si se lograse la integración se obtendrían importantes beneficios respecto al resto de métodos numéricos.

Las soluciones obtenidas por el método analítico juegan un papel importante en el método que se describirá a continuación, el método fit. Dichas soluciones son usadas en el método fit original como valores iniciales a la hora de realizar una minimización mediante la cual se encuentran las soluciones del System8.

4.5. Método fit

El método fit busca la resolución del System8 mediante la recombinación de sus ecuaciones en una unica función con sentido estadístico de 8 variables que posteriormente se minimiza en función de dichas variables permitiendo encontrar los valores correspondientes a la solución del System8. Previo a la realización de la minimización se transforman los datos (n, p...etc) en un nuevo set de datos que, originalmente, no es disconjunto pero muestra un menor solapamiento que sin su transformación.

La función que se minimiza es un versión modificada adecuadamente de la función de verosimilitud logarítmica (Log-likelihood):

$$lnL = \sum_{i=1} (\frac{\mu_i^{x_i} e^{-\mu_i}}{x_i!}) \tag{14}$$

Siendo x_i los valores de las variables i obtenidos de los datos reales y μ_i los valores obtenidos de las expresiones analíticas de las ecuaciones del System8. La función de verosimilitud es adecuada si la distribución de datos se ajusta a una estadística de Poisson, como es habitual en física de partículas. En el caso de que predicciones teóricas y datos reales coincidan su valor es máximo. Realizando una transformación adecuada se llega a:

$$f = -2 * \sum_{i=1}^{8} x_i * ln(\mu_i - x_i) \tag{15}$$

El termino -2 esta indicado por términos estadísticos, en concreto por la relación de la función con una función del tipo Xi Squared (X^2) si el numero de eventos es suficientemente grande. En este caso el valor de la función f sera menor cuanto mas se asemejen ambos valores: datos reales y predicción teórica.

La minimización se realiza mediante el paquete MINUIT implementado en ROOT. MINUIT ofrece diversos algoritmos para realizar la minimización y en este caso se parte del algoritmo MIGRAD, que usa un método de variación de métrica de las variables [23]. Se parte de unos valores razonables de las incógnitas (originalmente las soluciones del método analítico), se calcula la función (15) y posteriormente se varia el valor de las incógnitas de acuerdo al algoritmo para obtener un valor de f inferior. Se realizan numerosas iteraciones de este proceso hasta que se converge a un valor mínimo de (15). Originalmente además se establece unos limites máximos y mínimos que las variables pueden tomar durante la minimización. El valor de las incógnitas utilizadas en la ultima iteración son las soluciones a las incógnitas del System8.

Se presenta el problema que al usar el código el método es incapaz de sacar soluciones al problema. Se ha requerido un extenso estudio del código original a nivel técnico para su solución.

4.5.1. Integración y solución de errores

El uso del método fit tal y como se encontraba al inicio de la realización de este TFG y como se ha descrito en la sección anterior no es capaz de encontrar ninguna solución del System8 o exportar gráficas. Se han realizado sustanciales arreglos de errores, mejoras y modificaciones en general para lograr una integración plena y la convergencia a soluciones correctas. A continuación se explican algunas de los mas relevantes:

a) Obtención correcta de factores de correlación:

Se localiza un grave error por el cual en la toma de datos en el método fit el código asignaba erróneamente a los factores de correlación α, β, γ, δ, κ_b y κ_{cl} siempre el valor 0. Era el causante de que el método no exportase ningún resultado: había un conflicto con el termino logarítmico de (15) y el valor 0. Las soluciones obtenidas tras solo este cambio son erróneas.

b) Forma correcta del System8:

En el código del archivo S8fcn.cc la última ecuación del System8 viene escrita de la siguiente forma:

$$p^{tag,p_T^{rel}} = \kappa_b \beta\delta\epsilon_b^{tag}\epsilon_b^{p_T^{rel}} p_b + \kappa_{cl}\alpha\gamma\epsilon_{cl}^{tag}\epsilon_{cl}^{p_T^{rel}} p_{cl} \tag{16}$$

Dicha ecuación es incorrecta. Existen discrepancias en su forma entre el código, la nota interna de CMS donde se detalla la implementación del método por parte de la colaboración [24] y la nota pública de la colaboración Atlas donde se describe el System8 [22] . Su forma correcta es la mostrada por la nota publica del Atlas - y que se ha indicado en (4) - y es:

$$p^{tag,p_T^{rel}} = \kappa_b^{123} \beta\delta\epsilon_b^{tag}\epsilon_b^{p_T^{rel}} p_b + \kappa_{cl}^{123}\alpha\gamma\epsilon_{cl}^{tag}\epsilon_{cl}^{p_T^{rel}} p_{cl} \tag{17}$$

En efecto, calculando los productos $\kappa_b^{123}\beta\delta$ y $\kappa_{cl}^{123}\alpha\gamma\epsilon$:

$$\kappa_b^{123}\beta\delta = \frac{(\epsilon_b^{tag,p_T^{rel}})_p}{(\epsilon_b^{tag})_p(\epsilon_b^{p_T^{rel}})_p} * \frac{(\epsilon_b^{tag})_p}{(\epsilon_b^{tag})_n} * \frac{(\epsilon_b^{p_T^{rel}})_p}{(\epsilon_b^{p_T^{rel}})_n} = \frac{(\epsilon_b^{tag,p_T^{rel}})_p}{(\epsilon_b^{tag})_n(\epsilon_b^{p_T^{rel}})_n} \tag{18}$$

$$\kappa_{cl}^{123}\alpha\gamma = \frac{(\epsilon_{cl}^{tag,p_T^{rel}})_p}{(\epsilon_{cl}^{tag})_p(\epsilon_{cl}^{p_T^{rel}})_p} * \frac{(\epsilon_{cl}^{tag})_p}{(\epsilon_{cl}^{tag})_n} * \frac{(\epsilon_{cl}^{p_T^{rel}})_p}{(\epsilon_{cl}^{p_T^{rel}})_n} = \frac{(\epsilon_{cl}^{tag,p_T^{rel}})_p}{(\epsilon_{cl}^{tag})_n(\epsilon_{cl}^{p_T^{rel}})_n} \tag{19}$$

El producto indica las correlaciones correctas para la ecuación (17) que deben tener en cuenta la muestra n en ellas.

Una vez determinada la forma correcta de la última ecuación del System8 se ha modificado en el código. Los factores de correlación κ_b^{123} κ_{cl}^{123} no se encontraban definidos en el código para su utilización en el método fit. Para poder usarlos se han calculado explícitamente a partir de las muestras MC y se han preparado los archivos para su introducción y correcto uso.

c) Límites en la minimización:

En el código la minimización realizada con MINUIT (clase definida en ROOT) establece unos límites mínimos y máximos a los valores que puede tomar las incógnitas en el proceso. En concreto se establece para los valores n_b, n_{cl}, p_{cl} y p_b una variación máxima y mínima de la suma/resta de la raíz cuadrada del valor inicial tomado. Como los valores iniciales son en un principio los resultados del método analítico y se muestra para los últimos rangos de p_t considerados unos resultados para estas variables erróneos dicho error es arrastrado al

método fit. Para las eficiencias se indicaba unos limites mínimos y máximos que eran ambos mayores que el valor inicial tomado, forzando a que los resultados de las eficiencias fueran mucho mayores de lo esperado. Además el uso de límites en la minimización mediante MINUIT a nivel interno de código transforma un problema de minimización en un problema no lineal que ocasiona una menor precisión y un aumento de tiempo en la computación del problema [23]. Se han eliminado todos los límites en la minimización: si está bien implementado el método se tiene que llegar a soluciones correctas sin necesidad de forzar las soluciones.

d) Cambio de valores iniciales:

Originalmente se toman como valores iniciales para realizar la minimización del método fit las soluciones proporcionadas por el método analítico. Se han realizado extensas modificaciones en el código para que se tomen como punto inicial de la minimización los valores predichos por MonteCarlo y dejando como opción que se tomen en su lugar las soluciones del analítico, de tal forma que se realice el ajuste sin sesgos adicionales del MC. La toma de valores iniciales MC viene motivada inicialmente por la búsqueda de la solución mas próxima a ellos, aunque comprobaciones posteriores, una vez se han solventado todos los problemas del código (incluido la eliminación de los limites en la minimización del punto 3) han mostrado que se llega a las mismas soluciones independientemente de cualquier de los dos puntos de partida. La toma de valores iniciales MC presenta la ventaja de que habitualmente se converge a la solución en un menor número de iteraciones.

e) Set de datos disconjuntos:

La transformación de n, p, n^{tag}, p^{tag}, $n_T^{p\,rel}$, $p_T^{p\,rel}$, $n^{tag,p_T^{rel}}$ y $p^{tag,p_T^{rel}}$ a un set de datos disconjuntos no es correcta en el método fit original. Existe un solapamiento en varios pares de regiones como se muestra en la Figura 14 (Izquierda). El hecho de que no formen un set de datos ortogonales puede ocasionar problemas en la exactitud de la minimización realizada en el método fit. Se ha recalculado las transformaciones originales e insertado en el código para tener un set totalmente disconjunto como se observa en la Figura 14 (Derecha).

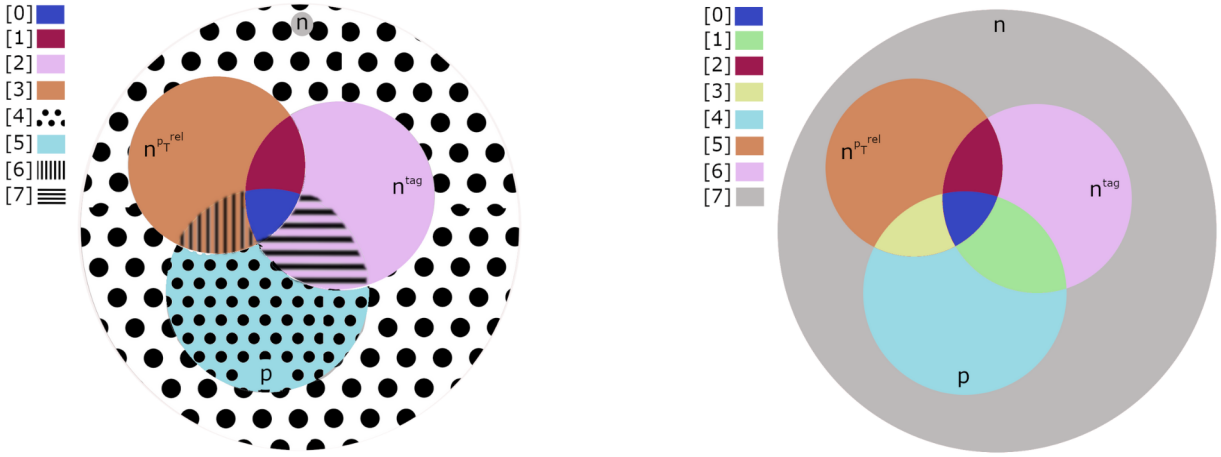

Figura 14: Diagramas de Venn mostrando el conjunto de datos usados ([0]-[8]) en el método fit. Izquierda: original, set de datos no ortogonales. Derecha: modificados: set de datos ortogonales.

f) Exportación de gráficas y depuración del código:

Una vez se han llegado a resultados correctos finalmente se ha realizado una implementación en el código para la realización automática de las gráficas, con la máxima similitud a las del método numérico, de las eficiencias de los algoritmos de b-tagging y la reproducción de sus scale factors.

Además se han realizado mejoras generales destinadas a una mayor limpieza del código e implementado opciones para la activación y desactivación de las diferentes opciones implementadas para su facilidad en futuros estudios.

4.5.2. Resultados y discusión

Una vez realizado el arreglo de errores e implementación de mejoras descritas en la anterior subsección se procede a la obtención y presentación de los resultados pertenecientes al set de datos 2018UL con el método fit (Figura 15). Sus errores son tomados a partir de la precisión de la minimización. Son dibujados en las gráficas pero son tan reducidos que no son apreciables.

A continuación se indican en formato numérico los resultados obtenidos para las eficiencias de b-tagging con el método fit y su comparativa con los valores obtenidos por el método numérico (con el criterio de selección de b-jetes establecido en 0.15) y los valores predichos MC (Tablas 2-7).

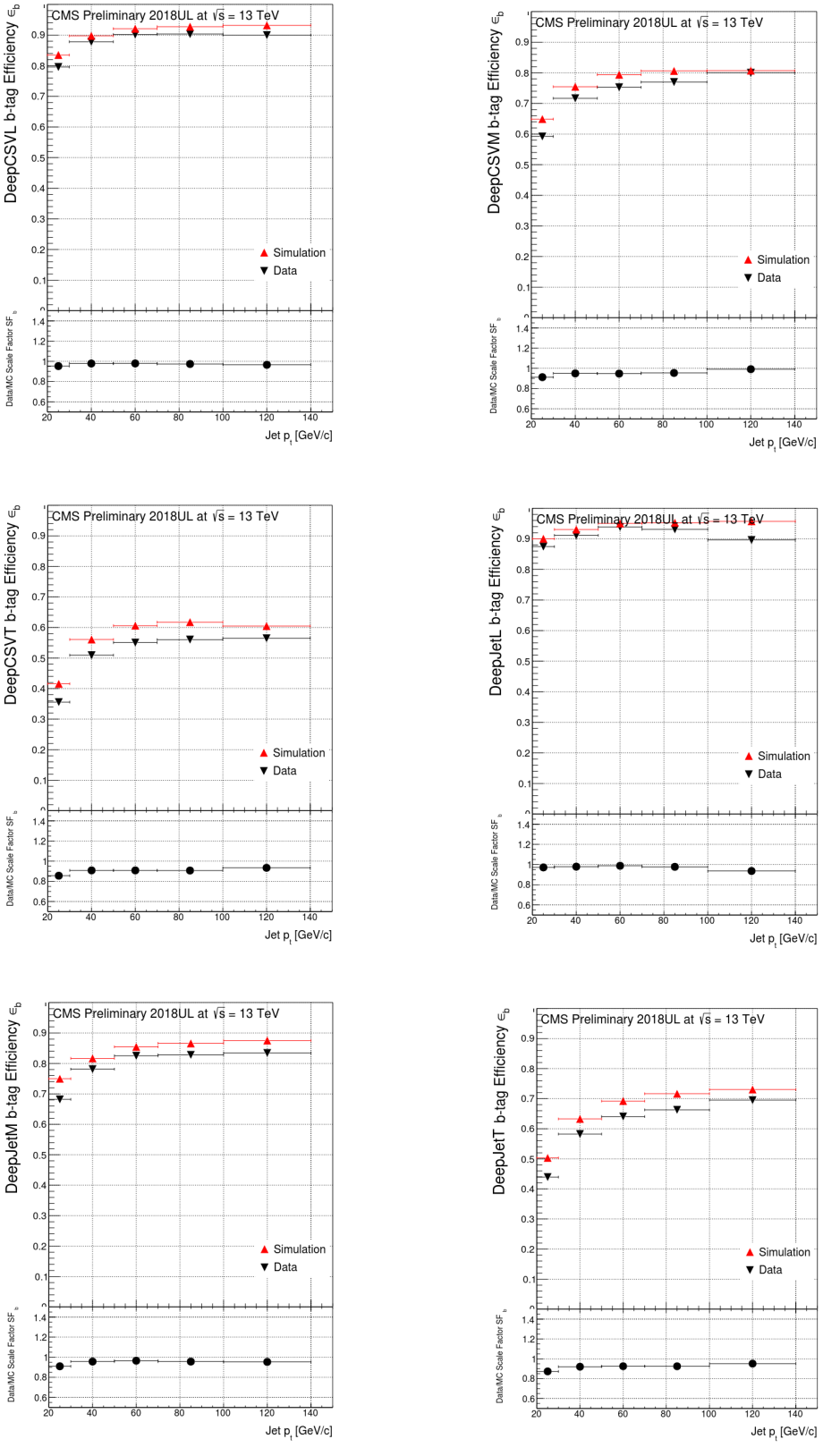

Figura 15: Eficiencias de los algoritmos de b-tagging (arriba) y scale factors (abajo) determinados mediante el método fit . Ordenados de izquierda a derecha y de arriba a abajo: DeepCSVL, DeepCSVM, DeepCSVT, DeepJetL, DeepJetM y DeepJetT.

Bin (GeV)	Método numérico	Método fit	MC
20-30	0.796 ± 0.006	0.796 ± 0.004	0.835 ± 0.001
30-50	0.878 ± 0.005	0.878 ± 0.002	0.8974 ± 0.0003
50-70	0.902 ± 0.005	0.902 ± 0.003	0.9208 ± 0.0004
70-100	0.903 ± 0.004	0.903 ± 0.003	0.9274 ± 0.0003
100-140	No	0.900 ± 0.005	0.9318 ± 0.0003

Tabla 2 Eficiencias de b-tagging del algoritmo DeepCSVL.

Bin (GeV)	Método numérico	Método fit	MC
20-30	0.592 ± 0.007	0.592 ± 0.005	0.649 ± 0.001
30-50	0.717 ± 0.004	0.717 ± 0.002	0.7546 ± 0.0005
50-70	0.752 ± 0.004	0.753 ± 0.003	0.7940 ± 0.0006
70-100	0.770 ± 0.005	0.770 ± 0.003	0.8063 ± 0.0004
100-140	0.801 ± 0.008	0.800 ± 0.006	0.8070 ± 0.0005

Tabla 3 Eficiencias de b-tagging del algoritmo DeepCSVM.

Bin (GeV)	Método numérico	Método fit	MC
20-30	0.356 ± 0.006	0.356 ± 0.004	0.416 ± 0.001
30-50	0.509 ± 0.005	0.509 ± 0.002	0.5606 ± 0.0005
50-70	0.550 ± 0.006	0.551 ± 0.002	0.6060 ± 0.0007
70-100	0.560 ± 0.004	0.560 ± 0.003	0.5537 ± 0.0006
100-140	0.565 ± 0.008	0.565 ± 0.004	0.6046 ± 0.0006

Tabla 4 Eficiencias de b-tagging del algoritmo DeepCSVT.

Bin (GeV)	Método numérico	Método fit	MC
20-30	0.875 ± 0.006	0.875 ± 0.005	0.900 ± 0.001
30-50	0.912 ± 0.005	0.911 ± 0.002	0.9307 ± 0.0003
50-70	0.938 ± 0.005	0.939 ± 0.002	0.9503 ± 0.0003
70-100	0.931 ± 0.004	0.931 ± 0.003	0.9526 ± 0.0002
100-140	0.750 ± 0.004	0.896 ± 0.005	0.9573 ± 0.0002

Tabla 5 Eficiencias de b-tagging del algoritmo DeepJetL.

Bin (GeV)	Método numérico	Método fit	MC
20-30	0.682 ± 0.007	0.682 ± 0.005	0.750 ± 0.001
30-50	0.782 ± 0.005	0.781 ± 0.003	0.8167 ± 0.0004
50-70	0.825 ± 0.004	0.825 ± 0.003	0.8546 ± 0.0005
70-100	0.828 ± 0.005	0.828 ± 0.004	0.8659 ± 0.0004
100-140	0.763 ± 0.008	0.834 ± 0.009	0.8753 ± 0.0004

Tabla 6 Eficiencias de b-tagging del algoritmo DeepJetM.

Bin (GeV)	Método numérico	Método fit	MC
20-30	0.439 ± 0.006	0.439 ± 0.005	0.503 ± 0.001
30-50	0.582 ± 0.004	0.582 ± 0.002	0.6327 ± 0.0005
50-70	0.640 ± 0.006	0.641 ± 0.003	0.6921 ± 0.0007
70-100	0.662 ± 0.005	0.663 ± 0.003	0.7164 ± 0.0005
100-140	0.695 ± 0.008	0.695 ± 0.006	0.7304 ± 0.0005

Tabla 7 Eficiencias de b-tagging del algoritmo DeepJetT.

Los resultados en la mayor parte de bines, algoritmos y puntos operativos coinciden con una elevada precisión con los pertenecientes al método numérico. Muestra la correcta implementación del método fit. Existen solo unas pocas excepciones, las cuales se discuten a continuación:

Caso 1: DeepCSVL, bin: 100-140 GeV.
El método fit halla una eficiencia de 0.900 ± 0.005. El porcentaje de b-jets esta en el límite de lo que se puede considerar una solución aceptable. El método numérico por otra parte es incapaz de encontrar solución incluso relajando los criterios de selección de b-jets a 0.15.

Caso 2: DeepJetL, bin: 100-140 GeV.
Tanto el método numérico como el método fit llegan a una solución pero en este caso difieren considerablemente. A priori, por su cercanía a MC y la general similitud de las eficiencias entre el último y penúltimo bin la solución del método fit de una eficiencia de 0.896 ± 0.005 y una fracción de b-jets de 0.312 ± 0.002 resulta mas correcta que la del numérico, de una eficiencia de 0.750 ± 0.004 y una fracción de b-jets 0.2058 ± 0.0004 que se sitúa en el límite del criterio de selección aplicado en el método numérico.

Caso 3: DeepJetM, bin 100-140 GeV.
El método numérico no localiza solución inicialmente y posteriormente a la relajación del criterio de selección a un 0.15 en la fracción de b-jets localiza una solución con una eficiencia de 0.763 ± 0.008 y una fracción de b-jets de 0.181 ± 0.005. Esta solución contrasta con la hallada por el método fit de una eficiencia de 0.834 ± 0.009 y una fracción de b-jets de 0.247 ± 0.004. En este caso se puede tomar la solución del método fit como correcta y descartar la solución del método numérico.

4.5.3. Dependencia de las soluciones con los factores de correlación (fit)

Motivado por los resultados de la sección 4.3.3 (Ver Figura 10) se procede a un estudio mas formal de la dependencia de las soluciones con los factores de correlación mediante el uso del método fit.

Para ello se implementa un smearing gaussiano similar al utilizado por defecto en el método numérico y al implementado en el método fit con los valores de n, p... etc, pero con los factores de correlación. Se varían todos ellos de acuerdo a la siguiente ecuación, teniendo una variación de cada uno en un intervalo alrededor de un 1 %.

$$x' = x + x * RndmGauss(0, 0.01) \tag{20}$$

Se realizan un número elevado de 10000 iteraciones y se obtienen los resultados. Para cada iteración si el algoritmo empleado por MINUIT en la minimización no encuentra solución se emplean dos algoritmos adicionales de MINUIT y si ninguno de ellos converge a un mínimo se salta a la siguiente iteración. Se presenta un histograma de las soluciones de las eficiencias para el ultimo bin (100-140 GeV) de DeepJetT en la Figura 16 (Izquierda). No se localiza solución para el total de iteraciones. Dependiendo del algoritmo y bin el porcentaje de soluciones halladas varia. Los casos en los que no se encuentra solución pueden deberse a la eficacia del método fit o al hecho de que el propio sistema puede no tener solución. Una revisión numérica de las soluciones halladas muestra que no todas tienen sentido físico. Algunas muestran fracciones de b-jets muy superiores o inferiores a lo aceptable, otras tienen valores negativos para alguna de las variables consideradas... etc. Se repite el proceso pero con una selección de soluciones que cumplan los siguientes requisitos:

1. La fracción de b-jets no puede ser inferior a 0.24 y superior a 0.36.
2. El valor de todas las variables debe ser positivo.
3. El valor de las eficiencias no puede ser superior a la unidad.
4. Las eficiencias de los algoritmos y del criterio de selección p_T^{rel} debe ser mayor para b-jets que para cl-jets.

Las soluciones halladas con dichos criterios para el mismo bin de DeepJetT se presentan en la Figura 16 (Derecha). Se reduce considerablemente el número de soluciones halladas respecto al caso anterior. Se observa que aquellas soluciones que se encuentran y que además tienen sentido físico se encuentran en un intervalo estrecho: la desviación estándar es mas pequeña que lo que se podía esperar a priori de la discusión de la dependencia de los factores de correlación en la sección 4.3.2.

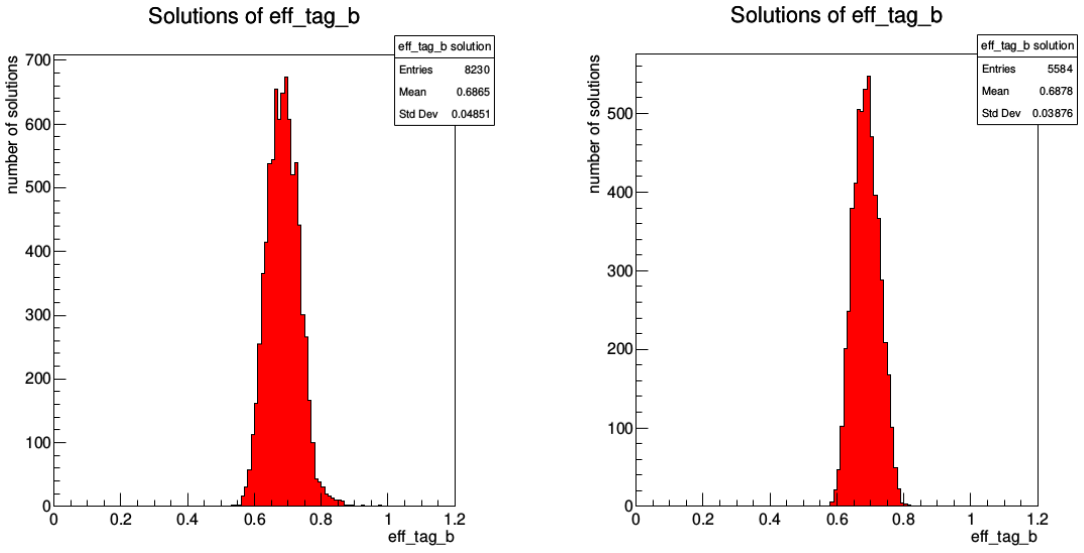

Figura 16: Smearing Gaussiano de los factores de correlación para el último bin de DeepJetT. Izquierda: eficiencia sin selección de soluciones. Derecha: eficiencia con selección de soluciones.

Se podría realizar un estudio variando los factores de correlación de forma individual un máximo y un mínimo de un 1 %, revisando las soluciones y asignando como error sistemático a cada factor de correlación la diferencia máxima de las eficiencias entre el valor central y el obtenido a través de la variación del 1 %. Posteriormente se sumaría en cuadratura los errores sistemáticos de cada factor de correlación y se asignaría el resultado como incertidumbre total debido al conjunto de los factores de correlación. Esto presenta un problema, una variación exacta del 1 % de cada factor de correlación (o cualquier otro porcentaje) no localiza para numerosos bines de los algoritmos un resultado.

En su lugar se procede a un estudio individualizado de la incertidumbre sistemática individual de cada factor de correlación y en conjunto de la siguiente forma: Se realiza un smearing gaussiano con 10000 iteraciones variándolos de la forma indicada en (21) para cada factor de correlación de forma individual y no en conjunto. Se toma el valor central del ajuste a una gaussiana del histograma. Se calcula la incertidumbre asociada al factor de correlación mediante la siguiente ecuación a partir del valor obtenido (Tablas (8) a (13)) y el valor sin realizar el smearing gaussiano (Tablas (2) a (7)):

$$Error = \frac{|valor\ central\ smearing\ gaussiano\ -\ valor\ sin\ semaring\ gaussiano|}{valor\ sin\ semaring\ gaussiano} \tag{21}$$

Los errores asociados a los factores de correlación se suman en cuadratura y se asigna el resultado como el error sistemático asociado a los factores de correlación en su conjunto.

Es necesario recalcar que uno estudio riguroso de la propagación de las incertidumbres de los factores de correlación requeriría de tener en cuenta las correlaciones entre los mismos. Dicho estudio va más allá de los objetivos de este TFG pero debería considerarse en estudios posteriores. Los resultados que se muestran a continuación nos proporcionan sin embargo una idea preliminar de la importancia de estas incertidumbres.

A continuación se presenta para el último bin considerado de DeepJetT los histogramas del smearing gaussiano con cada factor de correlación (Figuras 17 y 18):

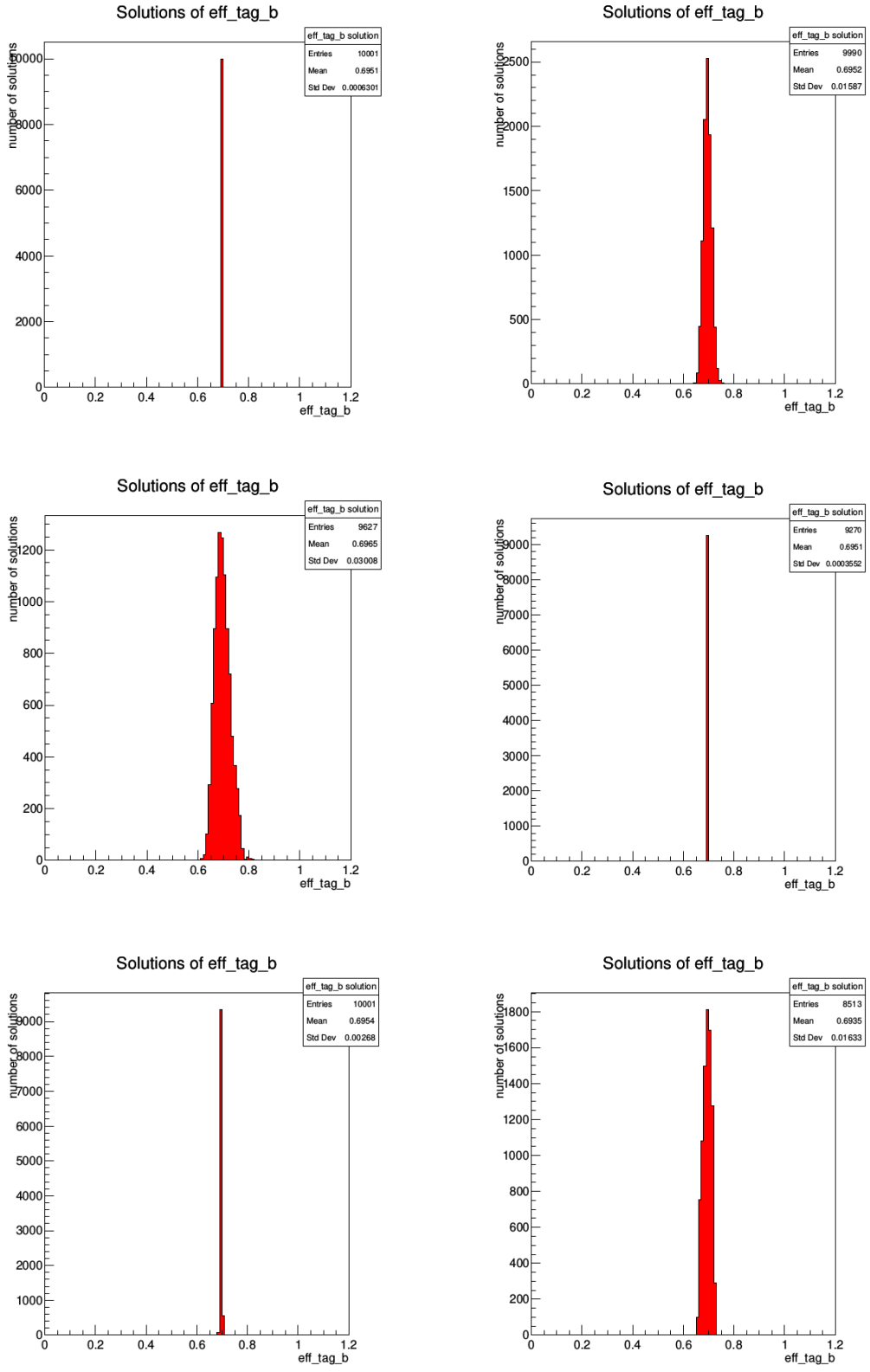

Figura 17: Eficiencias con smearing Gaussiano para el ultimo bin de DeepJetT individual para el factor de correlación, de izquierda a derecha y de arriba a abajo, gamma, delta, kb123 y kb.

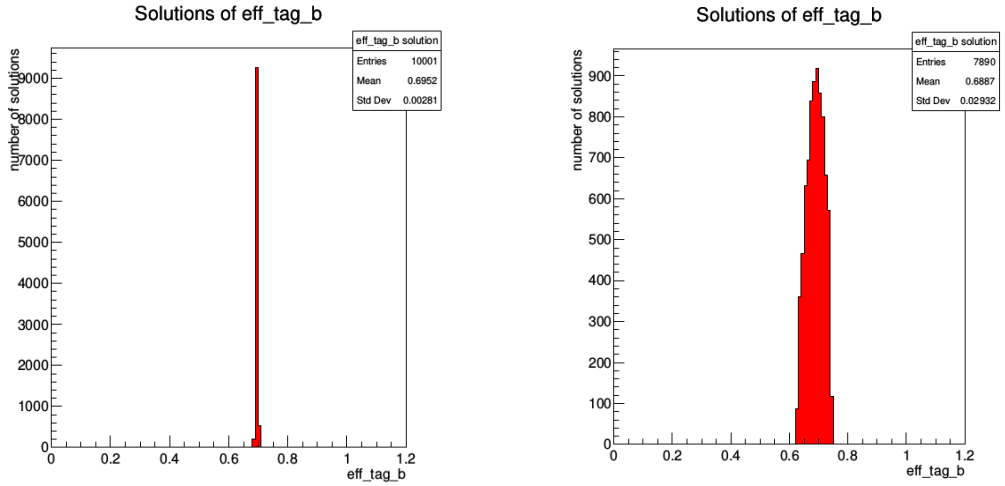

Figura 18: Eficiencias con smearing Gaussiano para el ultimo bin de DeepJetT individual para el factor de correlación: kcl123 (Izquierda) y kb123 (Derecha).

A continuación se resume para todos los algoritmos, puntos operativos y bines los errores sistemáticos de los factores de correlación en formato numérico (Tablas 8-13). Si la incertidumbre indicada para un factor de correlación indicada es 0 es debido a que se ha obtenido que la solución central encontrada del Smearing Gaussiano coincide con el valor sin él. Alternativamente se ha asignado el símbolo - para ciertos casos en los que no se ha encontrado solución alguna con el Smearing Gaussiano compatible con los criterios de selección a pesar de haberse realizado 10000 iteraciones.

Bin (GeV)	α	β	γ	δ	k_{cl}	k_b	k_{cl}^{123}	k_b^{123}	Total
20-30	0.0001	0.001	0	0.0001	0	0.001	0.0006	0.003	0.003
30-50	0	0.0009	0.0001	0	0	0.0007	0.0003	0.0008	0.001
50-70	0	0.0009	0.0001	0	0.0001	0.0002	0.0006	0.003	0.003
70-100	0.002	0.0007	0.0007	0.0001	0.002	0.004	0.0009	0.006	0.008
100-140	-	-	-	-	0.02	0.002	0.01	0.03	0.04

Tabla 8 Errores sistematicos debido a los factores de correlación en las eficiencias de b-tagging para el algoritmo DeepCSVL.

Bin (GeV)	α	β	γ	δ	k_{cl}	k_b	k_{cl}^{123}	k_b^{123}	Total
20-30	0.0002	0.001	0.0002	0.0008	0.0005	0.003	0.0002	0.01	0.01
30-50	0	0.001	0.0001	0	0	0.0008	0.0001	0.0006	0.001
50-70	0	0.0009	0.0003	0.0001	0	0.0003	0.0001	0.006	0.006
70-100	0	0.001	0	0.0001	0.0001	0.003	0.0005	0.001	0.003
100-140	-	0.04	0.06	0.003	0.02	0.03	0.02	0.05	0.09

Tabla 9 Errores sistematicos debido a los factores de correlación en las eficiencias de b-tagging para el algoritmo DeepCSVM.

Bin (GeV)	α	β	γ	δ	k_{cl}	k_b	k_{cl}^{123}	k_b^{123}	Total
20-30	0.0003	0.001	0.0003	0.003	0	0.003	0	0.002	0.005
30-50	-	0.002	0.003	0.0002	0	0.008	-	0.02	0.02
50-70	0	0.0009	0.0004	0.0006	0	0.003	0	0.01	0.01
70-100	0.0002	0.002	0.0009	0.0009	0.0002	0.001	0.0002	0.009	0.009
100-140	0.0002	0.001	0.003	0.0002	0.0002	0.004	0.0005	0.003	0.006

Tabla 10 Errores sistematicos debido a los factores de correlación en las eficiencias de b-tagging para el algoritmo DeepCSVT.

Bin (GeV)	α	β	γ	δ	k_{cl}	k_b	k_{cl}^{123}	k_b^{123}	Total
20-30	0.0001	0.0007	0	0.0001	0.0009	0.003	0.0005	0.008	0.008
30-50	0.0001	0.0009	0.0001	0	0.0006	0.002	0.0004	0.004	0.004
50-70	0	0.0006	0.0001	0.0001	0.001	0.002	0.0004	0.002	0.003
70-100	-	0.02	-	0.006	0.007	0.005	0.01	0.02	0.03
100-140	0.0005	0.0009	0.005	0.0001	0.001	0.0009	0.001	0.004	0.007

Tabla 11 Errores sistematicos debido a los factores de correlación en las eficiencias de b-tagging para el algoritmo DeepJetL.

Bin (GeV)	α	β	γ	δ	k_{cl}	k_b	k_{cl}^{123}	k_b^{123}	Total
20-30	0	0.0009	0.0003	0.0002	0.0002	0.001	0.0002	0.002	0.003
30-50	0	0.0009	0.0003	0.0001	0.0001	0.001	0.0001	0.0001	0.001
50-70	0	0.0009	0.0002	0.0001	0	0.002	0.0002	0.002	0.003
70-100	0	0.0006	0.0001	0.0005	0.002	0.006	0.0004	0.01	0.01
100-140	0.0004	0.0002	0.02	0.003	0.002	0.004	0.002	0.01	0.02

Tabla 12 Errores sistematicos debido a los factores de correlación en las eficiencias de b-tagging para el algoritmo DeepJetM.

Bin (GeV)	α	β	γ	δ	k_{cl}	k_b	k_{cl}^{123}	k_b^{123}	Total
20-30	0	0.0002	0.0002	0.007	0	0.003	0	0.004	0.009
30-50	0	0.002	0.0002	0.0002	0	0.002	0.0002	0.009	0.009
50-70	0	0.0009	0.0003	0.0002	0	0.003	0	0.009	0.01
70-100	0	0.001	0	0.0006	0	0.0005	0.0002	0.002	0.002
100-140	0	0.001	0.002	0	0.0001	0.004	0.0004	0.006	0.008

Tabla 13 Errores sistemáticos debido a los factores de correlación en las eficiencias de b-tagging para el algoritmo DeepJetT.

Se observa que no todos los factores de correlación insertan la misma incertidumbre sistemática. Solo los factores β, γ, κ_b y κ_{cl} introducen un error no despreciable y considerablemente mayor al estadístico del propio método fit. Los errores sistemáticos debido a los factores de correlación en conjunto dependen del algoritmo y rango de p_T considerado (no observándose tendencias claras para los algoritmos estudiados) y oscilan entre valores despreciables de 0.0001 y valores elevados de 0.09.

5. Conclusiones

Se ha realizado un estudio de las eficiencias y los diferentes métodos de resolución del System8 considerados (método numérico, método analítico y método fit) para los dos algoritmos de b-tagging DeepCSV y DeepJet en sus puntos operativos, *loose* L, *medium* M y *tight* T.

Se ha logrado entender el método numérico. Es el método que presenta una mayor fiabilidad a la hora de la determinación de las soluciones del System8. Posteriormente se ha logrado recrear los resultados oficiales presentados por el IFCA en la colaboración CMS para la campaña de datos 2018UL.

Se ha integrado el método analítico como método de resolución del System8 para el caso en que los factores de correlación son igual a la unidad. Para dicho caso arrojan una solución correcta. Una resolución analítica para diferentes factores de correlación debe ser explorada en un futuro.

Se ha logrado modificar el método fit para integrarlo plenamente en el entorno usado para resolver el System8 logrando que alcance soluciones correctas. Su uso es una alternativa válida y complementaria al método numérico para la determinación de las soluciones del System8 capaz de encontrar soluciones razonables que el método numérico no consigue (Bin 100-140 GeV de los algoritmos DeepCSVL, DeepJetL y DeepJetM).

Se ha determinado que los resultados obtenidos para los tres métodos diferentes considerados en la determinación de las eficiencias de los algoritmos de b-tagging son sensibles a pequeñas variaciones de los factores de correlación. Los factores de correlación usados por defecto en el System8 se determinan a partir de los datos simulados MC y no tienen porque corresponderse exactamente a aquellos de los datos reales. La imposibilidad de calcular los factores de correlación exactos a partir de los datos reales impone una incertidumbre en el valor de los mismos que se propaga a la determinación mediante el System8 de los SF y eficiencias de los algoritmos de b-tagging independientemente del método usado.

Se ha logrado realizar un estudio preliminar que analiza el impacto de la incertidumbre que generan los factores de correlación en las eficiencias obtenidas a través del método fit mediante un smearing gaussiano y la selección de soluciones que cumplan unos criterios atendiendo a su sentido físico y sin considerar la relación entre los propios factores de correlación. Su valor dependiendo del algoritmo y rango de p_T considerado oscila entre 0.0001 y 0.09.

6. Bibliografía

[1] Peter W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". En: *Phys. Rev. Lett.* 13 (16 oct. de 1964), págs. 508-509. DOI: 10.1103/PhysRevLett.13.508. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.508.

[2] G. Aad et al. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". En: *Physics Letters B* 716.1 (2012), págs. 1-29. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2012.08.020. URL: https://www.sciencedirect.com/science/article/pii/S037026931200857X.

[3] S. Chatrchyan et al. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". En: *Phys. Rev. D* 86.1 (7 oct. de 2012), pág. 072010. ISSN: 0370-2693. DOI: 10.1103/PhysRevD.86.072010. URL: https://link.aps.org/doi/10.1103/PhysRevD.86.072010.

[4] *The Large Electron-Positron Collider*. Consultado: Noviembre 2022. URL: https://home.cern/science/accelerators/large-electron-positron-collider.

[5] *The Large Hadron Collider*. Consultado: Noviembre 2022. URL: https://www.home.cern/science/accelerators/large-hadron-collider.

[6] B Salvachua. "Overview of Proton-Proton Physics during Run 2". En: (2019), págs. 7-14. URL: https://cds.cern.ch/record/2750272.

[7] David Wardrope et al. "Non-resonant Higgs-pair production in the $b\bar{b}$ $b\bar{b}$ final state at the LHC". En: *Eur. Phys. J. C* 75.5 (2015), pág. 219. DOI: 10.1140/epjc/s10052-015-3439-0. arXiv: 1410.2794 [hep-ph].

[8] Qing-Hong Cao, Shao-Long Chen y Yandong Liu. "Probing Higgs width and top quark Yukawa coupling from $t\bar{t}H$ and $t\bar{t}t\bar{t}$ productions". En: *Phys. Rev. D* 95 (5 mar. de 2017), pág. 053004. DOI: 10.1103/PhysRevD.95.053004. URL: https://link.aps.org/doi/10.1103/PhysRevD.95.053004.

[9] Chatrchyan S. et al. "Search for supersymmetry in events with b-quark jets and missing transverse energy in pp collisions at 7 TeV". En: *Phys. Rev. D* 86 (7 oct. de 2012), pág. 072010. DOI: 10.1103/PhysRevD.86.072010. URL: https://link.aps.org/doi/10.1103/PhysRevD.86.072010.

[10] *CMS*. Consultado: Noviembre 2022. URL: https://home.cern/science/experiments/cms.

[11] L. Cremaldi. "CMS pixel detector—overview". En: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 511.1 (2003). Proceedings of the 11th International Workshop on Vertex Detectors, págs. 64-67. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(03)01752-2. URL: https://www.sciencedirect.com/science/article/pii/S0168900203017522.

[12] *Bending Particles*. Consultado: Noviembre 2022. URL: https://cms.cern/index.php/detector/bending-particles.

[13] *Triggering and data acquisition*. Consultado: Diciembre 2022. URL: https://cms.cern/index.php/detector/triggering-and-data-acquisition.

[14] Luca Scodellaro. *b tagging in ATLAS and CMS*. 2017. DOI: 10.48550/ARXIV.1709.01290. URL: https://arxiv.org/abs/1709.01290.

[15] B.R. Martin y G. Shaw. *Particle Physics*. The Manchester Physics Series. John Wiley Sons Ltd., 2017. ISBN: 9781118912164.

[16] Christian Weiser. *A Combined Secondary Vertex Based B-Tagging Algorithm in CMS*. Inf. téc. Geneva: CERN, 2006. URL: http://cds.cern.ch/record/927399.

[17] Barbara Chazin Quero. *Machine learning techniques for heavy flavour identification*. Inf. téc. Geneva: CERN, 2018. DOI: 10.22323/1.321.0066. URL: https://cds.cern.ch/record/2638064.

[18] Andrea Perrotta. "CMS event reconstruction status in Run 2". En: *EPJ Web Conf.* 214 (2019), pág. 02015. DOI: 10.1051/epjconf/201921402015. URL: https://cds.cern.ch/record/2728524.

[19] Markus Stoye y on behalf of the CMS collaboration. "Deep learning in jet reconstruction at CMS". En: *Journal of Physics: Conference Series* 1085.4 (sep. de 2018), pág. 042029. DOI: 10.1088/1742-6596/1085/4/042029. URL: https://dx.doi.org/10.1088/1742-6596/1085/4/042029.

[20] E. Bols et al. "Jet flavour classification using DeepJet". En: *Journal of Instrumentation* 15.12 (dic. de 2020), P12012-P12012. DOI: 10.1088/1748-0221/15/12/p12012. URL: https://doi.org/10.1088%2F1748-0221%2F15%2F12%2Fp12012.

[21] A.M. Sirunyan et al. "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV". En: *Journal of Instrumentation* 13.05 (mayo de 2018), P05011-P05011. DOI: 10.1088/1748-0221/13/05/p05011. URL: https://doi.org/10.1088%2F1748-0221%2F13%2F05%2Fp05011.

[22] *b-Jet Tagging Efficiency Calibration using the System8 Method*. Inf. téc. Geneva: CERN, 2011. URL: https://cds.cern.ch/record/1386703.

[23] *ROOT: TMinuit Class Reference*. Consultado: Diciembre 2022. URL: https://root.cern.ch/download/minuit.pdf.

[24] Saptaparna Bhattacharya et al. *Measurement of the b-tagging efficiency in pp collisions at 8 TeV using the System8 method*. Diciembre 2012.

7. Anexo

7.1. Descripción del framework del System8

Se procede a una breve descripción del framework y del flujo de trabajo del System8. Un resumen esquemático puede encontrarse en la Figura 19. Esta integrado en el sistema operativo LINUX, su acceso se realiza mediante los servidores del CERN y los archivos involucrados hacen uso del lenguaje de programación C++ y del propio lenguaje de LINUX. La descripción se corresponde a la existente tras todas las modificaciones introducidas en el TFG y se han obviado en el relato numerosos archivos auxiliares sin especial interés.

Inicialmente se parte de los archivos miniAOD distribuidos por el CERN. Dichos datos son tratados de la manera descrita en la sección 4.1 (Tratamiento adicional de datos). El proceso se inicia ejecutando el archivo contenido en la carpeta s8Solver: runOOPs.csh. En el archivo mencionado se puede indicar los templates que se toman para realizar la resolución del System8. El archivo run_s8.C dentro de la carpeta BASE de s8Solve puede modificarse para ajustar ciertas opciones dentro de los archivos internos de la carpeta S8Solver sin necesidad de modificarlos individualmente y recompilar posteriormente. Una vez iniciado el proceso se realiza un tratamiento más de los datos en el archivo S8NumericInput.cc. Allí se preparan los parámetros constantes del System8, es decir los inputs n, p... etc y los factores de correlación calculados a partir de MC para su introducción en los siguientes archivos. Los inputs n, p... etc producidos pasan al archivo general S8Solver.cc y los factores de correlación al archivo System8Solver.cc, donde se realiza el método numérico.

S8Solver.cc es el archivo central del proceso. Se ajusta el tamaño de las muestras a bines de tamaño fijo (los rangos de p_t considerados) y se realiza un ajuste de los datos, se recalculan los factores de correlación y sus ajustes a un polinomio de grado 0, 1 y 2 a partir de las muestras MC - necesarios para el método Fit ya que este no coge los factores de correlación del archivo S8NumerInput.cc - y en función de las opciones activadas el S8Solver.cc da paso al método analítico, al método fit o al método numérico.

Si el método analítico esta activo se calculan para cada bin considerado las incógnitas del System8 en el archivo S8AnalyticSolver.cc. Las soluciones se exportan de nuevo al S8Solver.cc, allí se realizan las gráficas deseadas y posteriormente se guardan los resultados en la carpeta s8Solver dentro de la carpeta correspondiente al algoritmo considerado (en formato numérico en solverOut.log y en formato gráfico en Analytichisto.pdf). El recorrido para el método fit es análogo al del método analítico pero en el archivo S8FitSolver.cc. Existe un paso adicional debido a que el cálculo del set de datos disconjuntos y la función que se minimiza (15) están definidos externamente en el archivo S8fcn.cc. Los resultados numéricos vuelven a guardarse en el archivo solverOut.log y los gráficos en Fithisto.pdf. El método por defecto, el numérico, toma los inputs n, p... etc del archivo S8Solver.cc y los factores de correlación de S8NumericInput.cc. Se ejecuta el método y se exportan las soluciones a S8Solver.cc. Después las soluciones se guardan en un formato adecuado en S8Solution.h y se preparan para su representación gráfica en S8SolverInput.cc. De allí se realizan las gráficas con los archivos S8GraphGroup.cc, superimpose_eff.C y eff_sf.c. Finalmente los resultados se exportan numéricamente a solverOut.log y gráficamente a s8_eff_sf_Deep*.pdf.

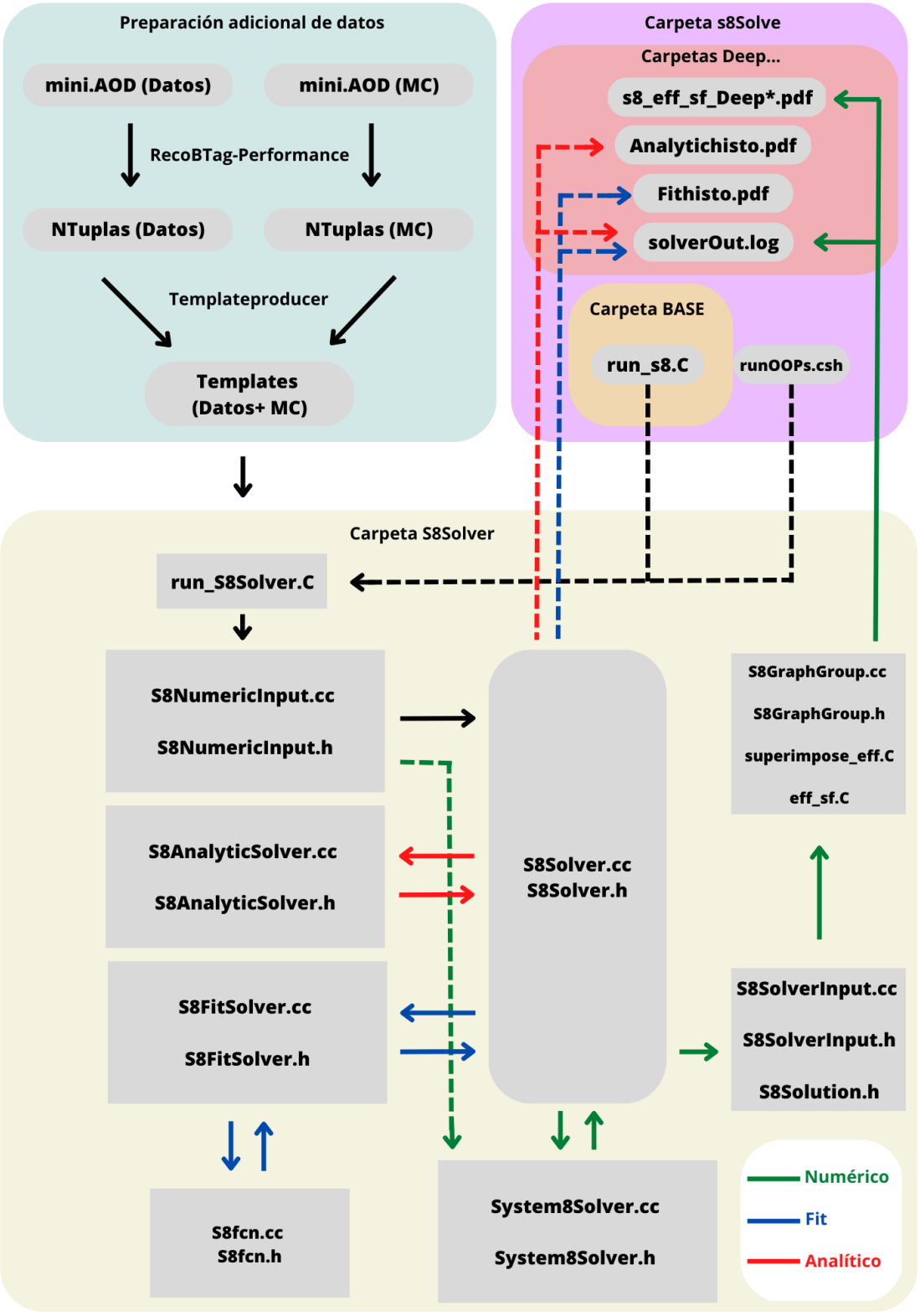

Figura 19: Resumen esquemático del análisis del workflow del System8.