LIP-STUDENTS-22-32

# Produção central e exclusiva de pares $\tau^+\tau^-$ no LHC no canal de decaimento $e\tau_h$

Gonçalo Esteves[1], João Tavares[1], João Nunes[1], Francisco Dias[1]

[1] *Instituto Superior Técnico, Lisboa, Portugal*

Coordenador do Projeto: Matteo Pisano
Project developed within the course: LFUI, IST

*January 30, 2023*

**Abstract.** Os dois objetivos deste trabalho são: a observação da recolha de dados do Run 3 analisando a qualidade do fill com base nos dados retirados e a determinação da secção eficaz do processo $pp \to p + \tau^-\tau^+ + p$ (no canal de decaimento $\tau^-\tau^+ \to \tau_h + e + \bar{\nu}_e + \nu_\tau$).
Em relação ao primeiro objetivo, após analisar os dados concluímos que tudo correu de acordo com o previsto e os dados do Run 3 são viáveis para análise futura.
Quanto ao segundo objetivo, devido à grande quantidade de eventos a ocorrer em cada instante no CMS tivemos de recorrer a métodos computacionais de tratamento de dados para poder calcular a secção eficaz. Primeiramente, usámos filtros para diminuir a quantidade de dados e de seguida aplicámos classificadores que fazem uso de técnicas de análise multivariada para ficarmos apenas com os dados de interesse. Após toda a análise e controlo de qualidade deste processo de tratamento de dados pudémos então prosseguir com o cálculo da secção eficaz ($\sigma$). Infelizmente, a luminosidade integrada em 2018 não é suficiente para determinar o valor exato da secção eficaz. Por tanto, foi possivel apenas establecer um valor limite da secção eficaz, entre 20 fb e 180 fb. Valores de $\sigma$ inferiores ao limite não podem ser medidos.

Keywords: LHC, proton-proton collision, tau-tau exclusive, exclusive production, TMVA

## I Introdução Teórica

Ao longo do nosso projeto vamos estar interessados em estudar o processo $pp \to p + \tau^-\tau^+ + p$, colaborando com a experiência CMS (Compact Muon Solenoid) do LHC (CERN). No LHC, pacotes de protões são acelerados até alcançarem a energia de 6.5 TeV (2016-2018), que após a colisão interagem com os detetores de partículas. No nosso trabalho, vamos estudar as interações no IP5, no centro do CMS.

No nosso processo os protões interagem sem se dissociar (permanecem intactos), e a consequente perda de energia é utilizada para criar, no estado final, um par $\tau^-$ $\tau^+$.

O ângulo entre a direção final e inicial dos protões, "$\theta$", encontra-se relacionado com a fração de momento perdido pelos hadrões durante a interação, $\xi = \frac{p_i - p_f}{p_i}$, sendo cada protão caracterizado por um $\xi$. Como o ângulo $\theta$ é muito pequeno, o CMS não consegue detetar os protões. Para isso foi construído um outro detetor com o objetivo de detetar os protões, o PPS. Este é formado por dois braços simétricos, cada um formado por 3 "estações", designadas "Roman Pots" (RP), que ficam a uma distância $\approx 200\,m$ do IP5. Cada RP contém um conjunto de detetores de partículas, que podem ser de dois tipos:

- Estação de "tracking": Constituída por um conjunto de seis painéis de Si (2018). Quando um protão passa é detetado pelos pixéis, permitindo determinar a distância entre o protão e o feixe. Esta informação está relacionada com o ângulo $\theta$ e $\xi$.
- Estação de "timing": em 2018 a eficiência destes detetores era limitada não permitindo aproveitar a informação recolhida, útil para derivar o instante temporal do IP5.

De facto, o valor de $\xi$ dos dois protões está relacionado com a massa invariante ($M_X$) e com a rapididade ($Y_X$) do sistema central, formado pelo par $\tau^-$ $\tau^+$, e que vão ser fundamentais ao longo do trabalho:

$$M_X = \sqrt{s\xi_1\xi_2} \qquad Y_X = \frac{1}{2} log \frac{\xi_1}{\xi_2} \tag{1}$$

Quanto ao sistema central ($\tau^-$ $\tau^+$) é formado por um par de partículas instáveis que podem decair de várias formas:

- Decaímento eletrónico: $\tau^- \to e^- + \bar{\nu}_e + \nu_\tau$
- Decaímento muónico: $\tau^- \to \mu^- + \bar{\nu}_\mu + \nu_\tau$
- Decaímento hadrónico: $\tau^- \to \tau_h$, sendo $\tau_h$ um conjunto de hadrões.

O nosso papel é estudar o decaímento, $\tau\tau \to \tau_e\tau_h$, em que $\tau_e = e^- + \bar{\nu}_e + \nu_\tau$, de forma a simplificar a notação, e $\tau_h$ é um conjunto de hadrões, cujo objetivo primordial passa por medir a secção eficaz deste mesmo processo. Para isso, é necessário identificar os acontecimentos de interesse entre todos os acontecimentos gerados no IP5 (no LHC ocorrem cerca de 40 milhões de interações por segundo).

Mais, no ano de 2018, no CMS, ocorreram cerca de 30 interações ao mesmo tempo (embora estas aconteçam em instantes temporais distintos, o CMS não consegue distingui-las). O estado final é então caraterizado por várias partículas sobrepostas ao acontecimento de interesse. Este fenómeno é designado "Pile-Up", sendo este um dos desafios do nosso trabalho.

Porém, o desafio principal na determinação dos acontecimentos de interesse é identificar e eliminar acontecimentos semelhantes ao sinal, mas que provêm de interações distintas. Por exemplo, o processo $pp \rightarrow \tau^-\tau^+$ é parecido ao nosso sinal, mas não contém protões no estado final do processo tendo que ser descartado. A este tipo de processos dá-se o nome de "fundos" e é portanto necessário indentificá-los e consequentemente apagá-los:

- Fundo Drell Yan: $pp \rightarrow l^+l^-$, sendo $l$ um leptão carregado. Se $l$ for um $\tau$, o processo $pp \rightarrow \tau^-\tau^+$ é muito parecido ao sinal.
- Produção inclusiva de pares $t\bar{t}$: $pp \rightarrow t\bar{t}$, sendo $t$ o quark top. Se o top decair num leptão, o estado final é muito parecido ao sinal.
- Fundo QCD: é um processo com uma secção eficaz relevante. É caracterizado pela produção de quarks e gluões que podem ser confundidos com um $\tau_h$, apesar da não elevada probabilidade desta troca a secção eficaz é tão grande que é necessário ter este processo em conta.

Como podemos verificar nenhum dos fundos descritos contém protões no estado final, o que poderia facilitar a nossa análise. Porém, devido ao Pile-Up, o PPS deteta protões que estão associados a processos secundários que acontecem ao mesmo tempo da interação de interesse, o que pode levar à deteção de protões nos fundos anteriormente referidos, não podendo descartá-los de forma imediata.

De forma a facilitar a compreensão da secção da aquisição de dados é necessário introduzir uma quantidade designada "luminosidade" (L). A luminosidade fornecida por um acelerador de partículas é uma quantidade que relaciona o número de protões que interagem durante uma recolha de dados. Considerando um processo físico P, cuja secção eficaz é $\sigma$, e N o número de eventos observados num intervalo de tempo t, a luminosidade fornecida pelo acelerador relaciona-se por: $L = \frac{N}{\sigma}$. É de referir que esta quantidade não depende do processo físico considerado, mas apenas das características do acelerador. A luminosidade pode ser calculada a partir da seguinte fórmula:

$$L = \frac{k n_1 n_2 f}{4\pi\sigma_x\sigma_y} \tag{2}$$

Em que: $f$ é a frequência de interação entre pacotes de protões, $k$ o número de pacotes em cada feixe, $n_{1,2}$ o número toral de partículas em cada feixe e $\sigma_{x,y}$ a "espessura" em duas direções perpendiculares (o perfil dos feixes é uma Gaussiana, em que tanto $\sigma_x$ como $\sigma_y$ representam a dispersão do feixe na direção dos eixos x e y).

Um dos maiores desafios dos cientistas e engenheiros é aumentar a luminosidade fornecida pelos aceleradores de partículas, pois estamos interessados em estudar eventos com baixos valores de $\sigma$.

É importante referir que, embora a análise seja efetuada baseando-se nos dados de 2018, o grupo participou na recolha de dados de 2022, cuja explicação da parte experimental será feita na secção II.

Numa secção posterior será descrito o processo de filtragem destes fundos e a consequente determinação da secção eficaz do nosso processo, objetivo primordial desta experiência.

## II Setup - Funcionamento do LHC e Aquisição de Dados

Antes de explicar o funcionamento do LHC (Large Hadron Collider), convém descrever um pouco a sua história. Até à data, este funcionou durante 3 épocas distintas (Runs):

- Run 1 - 2010-2013
- Run 2 - 2016-2018
- Run 3 - 2022-

O período entre *Runs* é denominado por *Shutdown*, e serve não só para os cientistas do CERN resolverem problemas técnicos dos detetores e dos aceleradores, mas também para melhoramentos da tecnologia existente e implementações novas. Podemos enumerar algumas:

- aumentar a luminosidade do acelerador - quanto maior, mais provavél é observar fenómenos cuja secção eficaz é muito reduzida;
- aumentar a eficiência dos detetores - não se pode só aumentar a luminosidade fornecida, porque sabemos que nem toda chega aos detetores, devido ao *pile-up*. É então necesário maximizar este rácio, e isso só é atingido fazendo alterações nos diferentes detetores, que serão descritos em detalhe mais à frente;
- aumentar a energia de colisão dos feixes de protões - isto pode ser atingido aumentando a intensidade do campo magnético fornecida. Deste modo, poderá ser possível estudar fenómenos que envolvem partículas mais pesadas, ou fenómenos que se dão a altas energias.

O nosso papel passará por participar na recolha de dados do *Run* 3, e posteriormente analisar alguns desses dados. Posteriormente, iremos recorrer a uma grande quantidade de dados do *Run* 2, disponibilizada ao nosso coordenador pelo CERN, de forma a fazer um estudo mais profundo, dado que teremos acesso a quantidades como a energia, momento, rapididade, etc.

### II.1 CMS

Passamos agora à descrição do CMS - na seguinte figura está contemplado um esquema deste detetor.

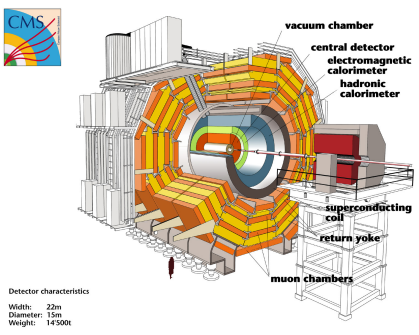


**Figure 1:** Esquema simplificado do CMS

Como vemos na figura, O CMS é constituído por uma série de subdetetores concêntricos:

***Central detector*, ou *Tracker***

Este subdetetor consiste em layers concêntricos (cilíndricos) de cerca de 1800 módulos. Cada módulo contém 66000 pixels. O *Tracker* é o detetor mais próximo do ponto de colisão dos feixes de protões, devido a isto permite-nos reconstruir as trajetórias de eletrões altamente energéticos, muões e hadrões carregados, e ainda de partículas com tempos de vida mais curtos. O tamanho de cada pixel (100x150 $\mu m^2$) permite-nos detetar a trajetória das partículas com uma precisão de 10 $\mu m$. Como as partículas carregadas são submetidas ao campo magnético do LHC, as suas trajetórias são espirais, e a curvatura dessas trajetórias permite ao *Tracker* determinar o momento das mesmas.

***Eletromagnetic Calorimeter***

Este subdetetor é responsável por medir as energias dos eletrões e fotões com alta precisão. Como estas partículas interagem eletromagneticamente, é designado por calorímetro eletromagnético.

***Hadronic Calorimeter***

Partículas que interagem pela força forte, como os hadrões, depositam a maioria da sua energia na camada seguinte, o calorímetro hadrónico. As únicas partículas que não são detetadas por estes 3 detetores são os muões e os neutrinos. Os muões são detetados ainda mais longe do ponto de colisão, nos *Muon Chamber Detectors*.

Um exemplo de um fenómeno que o calorímetro hadrónico deteta é o fenómeno dos *jets*.

Finalmente, podemos derivar quantidades cinemáticas sobre os neutrinos: a secção eficaz neutrino-matéria é muito reduzida e por tanto não podemos observá-los diretamente. Devido ao principio de conservação da energia e do momento linear, o momento total final tem de ser zero (o momento total inicial dos protões é zero). Por tanto, adicionando o momento de todas as partículas no estado final e trocando o sentido do vetor obtido, é possível calcular o momento associado aos neutrinos - como é óbvio esta quantidade é caraterizada por uma grande incerteza experimental.

## II.2 **Aquisição de dados**

### II.2.1 *Fill*

Os protões do LHC são acelerados por um conjunto de aceleradores (lineares e circulares). A recolha de dados no LHC é composta por *Fills*. Cada *Fill* é caraterizado por fases bem definidas:

- Fase 1: os protões são enviados no acelerador. Durante cada *Fill* cerca de $10^{14}$ protões entram no acelerador.
- Fase 2: os protões são acelerados até alcançarem uma energia de 13.6 $TeV$ (no Run 3). Os protões permanecem no LHC até atingirem a energia desejada.
- Fase 3: os protões interagem em pontos de interação bem definidos (zonas em que existem detetores, como o que nós estudamos - CMS)

Cada Fill tem uma duração diferente (entre uma e 35 horas). Obviamente, se ocorrer algum problema técnico durante a aquisição, o *Fill* é cancelado.

### II.2.2 *Triggers e Data Flow*

Após os fenómenos se darem nos pontos de colisão e serem detetados pelas diferentes componentes do CMS, é necessário guardar toda a informação proveniente dos detetores. Como sabemos, o número de interações é muito elevado, logo é impossível guardar toda esta informação de uma maneira simples. Para isso, no CERN é implementado um sistema de aquisição de dados (DAQ) para reconhecer, guardar e reconstruir os eventos de interesse.

O fucionamento do DAQ é o seguinte:

- **Definição dos eventos de interesse** - é necessário, antes de proceder a filtragens de informação, saber bem quais são os processos que pretendemos estudar, e que conjunto de informações é necessário ter para analisar esse processo. Esta fase é essencial, porque todos os dados que não são guardados pelo DAQ não poderão ser objeto de estudo.
- **LVL1 *Trigger*** - Este *Trigger* é o mecanismo de seleção responsável pela filtragem inicial de acontecimentos. Como há milhões de canais provenientes dos detetores, o *Trigger* tem de ser um mecanismo rápido e eficiente.

  No LHC os protões interagem com uma frequência de 40 $MHz$, e o L1 *Trigger* guarda apenas 0.25% da informação total, reduzindo a frequência para 100 $kHz$. O L1 *Trigger* tem acesso a um número muito restrito de variáveis (não esquecer que tem de ser eficiente) proveniente dos calorímetros e espectómetro de muões, e tem de determinar se o evento é de interesse ou não num intervalo de 3.8 $\mu s$, já que aparece um evento a cada 10 $\mu s$.
- **Fluxo de Dados** - Após o LVL1 *Trigger* reduzir o número de acontecimentos por um fator de 400, por cada acontecimento guardado o output é enviado para uma *Read-out unit* (RU). Os dados ficam na RU até o *event manager* consentir a sua transladação para a *Building unit*, onde o evento será reconstruído - isto é, já terá todas as variáveis necessárias (momento transversal, energia, ângulos, rapididade, entre outros).
- ***High Level Trigger* (HLT)** - Finalmente, os dados passam por este filtro que tem acesso a todas as variáveis e faz testes físicos muito específicos, para ter a certeza que o *output* final faz parte da lista de acontecimentos de interesse definida no passo 1. É de referir que este *Trigger* reduz o número de acontecimentos por um fator de 100, sendo o rate de acontecimentos final cerca de 2-3 $kHz$

### II.2.3 *Informações relevantes para a recolha de dados*

O objetivo principal da recolha de dados é verificar que o *Fill* está a decorrer sem problemas, e para isso é necessário controlar um certo número de variáveis.

A seguinte figura contém um dos painéis de informação mais importante que controlámos no dia da recolha de dados:

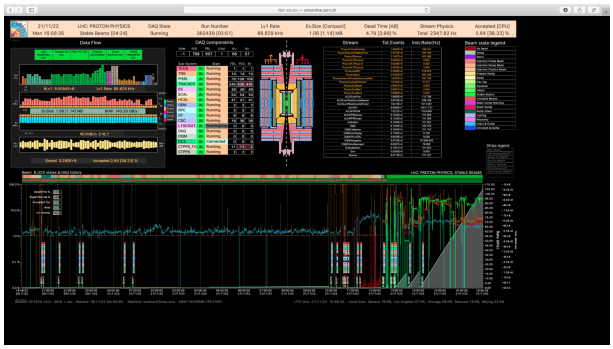

**Figure 2:** CMS online - Ecrã principal

Primeiramente, é necessário controlar o estado do feixe, que é o conjunto de gráficos na parte inferior da imagem. Entre estes gráficos destacam-se os rates dos *Triggers* ao longo do tempo, assim como diversos tempos mortos do sistema e a percentagem de eventos que passa os *Triggers*.

Foi também preciso controlar os valores numéricos das quantidades mencionadas no parágrafo anterior, que se situam na parte superior do ecrã. Mais à frente no trabalho estarão contemplados gráficos elaborados a partir da evolução temporal dessas quantidades.

Finalmente, é necessário verificar que todas os detetores do CMS estão a funcionar corretamente e verificar que a luminosidade fornecida pelo acelerador é constante ao longo do tempo, assim como a luminosidade atingida.

### II.2.4 *Sessão 1 - 21/11/2022*

Esta sessão corresponde à parte laboratorial do nosso trabalho; neste dia reunímos com o coordenador no LIP, e a aquisição de dados coincidiu propositadamente com o decorrer de um *Fill* do LHC.

A recolha de dados consistiu na aquisição das quantidades referidas na subsecção anterior (a cada 5 min fez-se uma captura e apontaram-se as quantidades numa folha de Excel). Este processo decorreu durante cerca de 2 horas.

### II.2.5 *Dados adquiridos*

Nesta subsecção será feita uma descrição dos dados adquiridos, será dada uma explicação das incertezas e será introduzida notação, relevantes para a posterior análise dos dados. As variáveis recolhidas foram:

- L1 - *LVL1 Trigger* Rate [kHz]
- RU - *Read-out Unit* Rate [Gb/s]
- BU - *Bulding Unit* Rate [Gb/s]
- ES - Event size [Mb/s]
- P - Percentagem de eventos guardados [%]
- TMT - Tempo Morto Total [%]
- TMF - Tempo Morto Feixe [%]

**Incertezas:** Para encontrar um valor para as incertezas foi feito o seguinte raciocínio: os dados recolhidos apareciam no painel com um número fixo de casas decimais; o que nós verificámos foi que regra geral a última casa decimal era bastante volátil enquanto que a penúltima não; assim escolhemos como incerteza um valor correspondente à penúltima casa decimal (exemplo: se o valor retirado é 2,56 ou 95,183 então as suas incertezas serão, respetivamente, 0,1 e 0,01). Esta regra aplica-se se a variável considerada não tiver uma flutuação estatística intrinseca maior da flutuação temporal. De facto, a única exceção a esta regra é a variável L1 cuja incerteza é 6% do seu valor: $\epsilon_{L1} \approx \epsilon_N = \frac{1}{\sqrt{N}} = \frac{1}{\sqrt{L1 \times \Delta t}} \approx 6\%$. Para calcular o erro é preciso lembrar que o rate do LV1 trigger é definido como $L1 = N/t$, sendo $N$ o número de acontecimentos que passam o trigger e $t$ o período de atualização dos valores no painel. Por tanto, $\epsilon_{L1} = \epsilon_N + \epsilon_t$, sendo $\epsilon_t$ desprezável. Além disso é de referir que foi usada uma incerteza temporal de 3 segundos visto que era este o período de atualização dos valores no painel.

Através das variáveis recolhidas foram calculadas outras variáveis de interesse, que nos ajudam a estudar melhor o funcionamento dos detetores e a verificar que os dados não apresentam anomalias:

- HLT - HLT Rate [kHz]
- TMD - Tempo morto detetor [%]

as quais são calculadas pelas equações:

$$HLT = \frac{L1 \times P}{100} \tag{3}$$

$$TMD = TMT - TMF \tag{4}$$

e cujas incertezas são:

$$\sigma_{HLT} = \frac{1}{100}\sqrt{P^2\sigma_{L1}^2 + (L1)^2\sigma_P^2} \tag{5}$$

$$\sigma_{TMD} = \sqrt{\sigma_{TMT}^2 + \sigma_{TMF}^2} \tag{6}$$

Em seguida apresenta-se uma pequena porção dos dados:

| Tempo (min) | 50 | 54 |
|---|---|---|
| L1 (kHz) | 94,91 ± 0,01 | 95,00 ± 0,01 |
| RU (Gb/s) | 148,6 ± 0,1 | 151,2 ± 0,1 |
| BU (Gb/s) | 148,6 ± 0,1 | 151,2 ± 0,1 |
| ES (Mb/s) | 1,6 ± 0,1 | 1,6 ± 0,1 |
| P (%) | 2,5 ± 0,1 | 2,4 ± 0,1 |
| TMT (%) | 4,1 ± 0,1 | 4,5 ± 0,1 |
| TMF (%) | 3,3 ± 0,1 | 3,7 ± 0,1 |
| HLT (kHz) | 2,35 ± 0,09 | 2,32 ± 0,09 |
| TMD (%) | 0,8 ± 0,1 | 0,8 ± 0,1 |

**Table I:** Amostra dos dados adquiridos e derivados

Caso seja de interesse os dados completos encontram-se em IV.6 (referências).

### II.2.6 *Análise dos dados adquiridos*

Agora sim estamos em condições de analisar os dados recolhidos. Serão apresentados os seguintes gráficos:

- L1 e HLT em função do tempo: permitir-nos-á avaliar se os valores obtidos para estes triggers é coerente com o esperado e desejado;
- P em função do tempo: averiguaremos se a percentagem de eventos guardados é constante no tempo;
- Informação perdida, $IP$, em função do tempo: a informação flui entre várias unidades até ser efetivamente guardada o que pode dar origem a perdas de informação;
- Tempo morto Total.

Começando portanto pelos triggers obteve-se o seguinte ajuste:

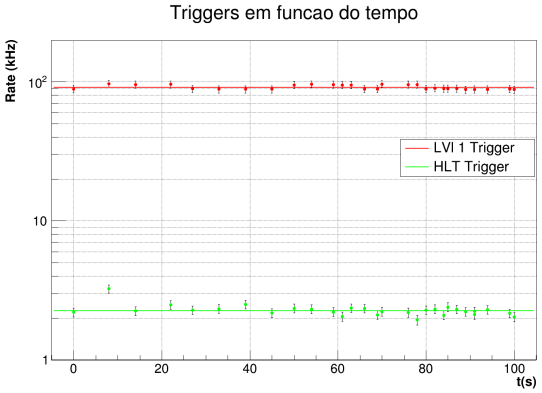


**Figure 3:** Triggers L1 e HLT em função do tempo

Verifica-se, como seria de esperar, que os rates são constantes no tempo sendo os valores médios encontrados (ajuste):

| L1 (kHz) | HLT (kHz) |
|---|---|
| 91 ± 1 | 2,25 ± 0,03 |

**Table II:** Valores médios Triggers (L1 e HLT)

Ambos os valores médios são coerentes com o que seria de esperar: sendo que a saída do LVL1 trigger é ligada às RU através de um cabo ethernet que pode transportar um volume de dados menor ou igual a 200 Gb/s, o valor máximo para o rate do LVL1 Trigger é de 100 kHz ($\frac{200Gb/s}{ES} > 100$ (kHz) já que por norma $1,5 < ES(Mb/s) < 2$); quanto ao HLT o valor estipulado pelo CERN para este trigger está no range $\leq$ 3-5 kHz o que também está de acordo como os nossos dados experimentais.

Quanto à percentagem de eventos guardados, P, obtivemos o seguinte ajuste:

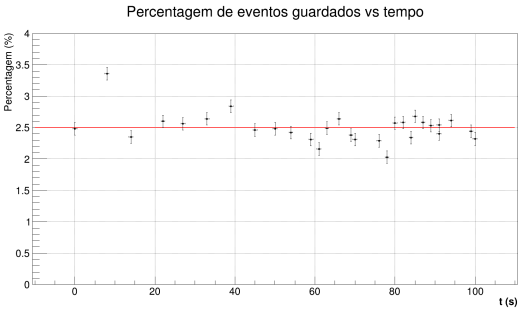


**Figure 4:** P em funcao do tempo

Tal como para os triggers, P é constante no tempo, o que faz sentido visto que a probabilidade dos processos físicos não depende do tempo. O seu valor médio encontrado através do ajuste é: $2,50 \pm 0,02$ (%), sendo este valor coerente com o rácio $\frac{HLT}{L1} \times 100$.

No caso da IP (Informação perdida) podemos avaliá-la em diferentes etapas no fluxo de informação:

- LVL1 $\rightarrow$ RU: esta é a fase mais crítica na perda de informação e é esta que vamos estudar;
- RU $\rightarrow$ BU: nesta fase a perda de informação é pequena ou mesmo inexistente, o que se pode verificar pelos nossos dados visto que nestes os rates RU e BU eram exatamente iguais.

Assim definimos a perda de informação como:

$$IP = \frac{\phi_{L1} - \phi_{RU}}{\phi_{L1}} \times 100 \tag{7}$$

Obtemos então o seguinte gráfico:

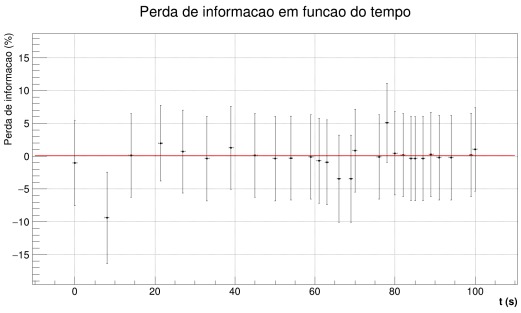


**Figure 5:** Perda de informação em função do tempo

O valor médio encontrado para a perda de informação é $0 \pm 1$ (%). Significa isto que a perda de informação é de 0 (o desejável). De notar que o segundo ponto representa um "ganho" de informação extra que está em concordância com o segundo ponto da figura 4. Uma possível explicação seria um atraso no fluxo de informação deixando alguma informção do acontecimento precendente na RU (e portanto registando-se um excesso de informação momentâneo).

Por fim analisaremos as variações do Tempo morto do detetor (TMD):

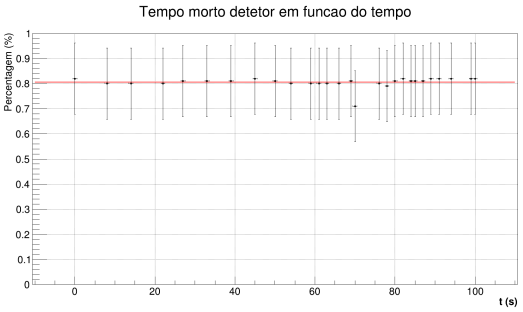


**Figure 6:** Tempo morto detetor em função do tempo

De acordo com o artigo IV.6 durante a aquisição de dados o tempo morto do CMS tem de ser inferior a 1%, coisa que de facto se verifica para todos os nossos pontos. O valor médio encontrado é $0,81 \pm 0,03$ (%), o que corresponde a um valor razoável tendo em conta os objetivos do CERN.

# III Tratamento e Análise dos dados do Run2

O conjunto de dados recolhidos pelo CMS não discrimina sobre nenhum processo específico e portanto corresponde a uma sobreposição de todos os processos possíveis. Isto faz com que seja necessário realizar uma fase de filtragem que nos selecione apenas os eventos correspondentes ao nosso decaímento. Estes filtros que usamos consistem em impor limites aos valores aceitáveis das variáveis que são guardadas pelo CMS III. Estes limites são escolhidos com base no modelo teórico e nas limitações dos próprios detetores.

Variáveis guardadas pelo CMS:

- $p_T$ o momento tranversal que é definido como o momento da partícula projetado no plano xy;
- $\Phi$ o ângulo entre o vetor $p_T$ e o eixo x;
- $\eta$ a pseudo-rapidez definida como $\eta = -\ln[\tan(\frac{\theta}{2})]$ em que $\theta$ é o ângulo entre o momento e o eixo z;
- E a energia das partículas;
- A carga das partículas.

### III.0.1 *Filtros*

1. O primeiro filtro a que os dados são submetidos é a verificação da presença de um eletrão e um $\tau_h$ isolados, isto é, que apresentem uma distância angular $\Delta R = \sqrt{\Delta\Phi^2 + \Delta\eta^2} > 0,4$ e cuja pseudo-rapidade ($\eta$) seja, em módulo, inferior a 2,4 porque para valores fora deste intervalo as partículas não interagem com os detetores.

2. A segunda condição que os dados têm de satisfazer para passar a fase de filtragem deve-se aos limites do detetor PPS. Este detetor só deteta protões que satisfaçam a condição $M_X = \sqrt{s\xi_1\xi_2} > 300\ GeV$.

   Tendo isto, uma boa aproximação é considerar que $p_T(\tau_h) \approx p_T(\tau_e)$ ou seja, $p_T(\tau) > \frac{M_X}{2} = 150\ GeV$. Embora sendo uma estimativa teórica não deixa de ser um valor aproximado e a energia pode não estar distribuída de maneira exatamente igual entre os dois $\tau$. Para além disso o CMS e o PPS não conseguem captar toda a energia das partículas portanto os valores reais medidos vão estar abaixo da previsão teórica. Por estas duas razões vamos minorar esta estimativa e usar como condição de seleção $p_T(\tau_h) > 100\ GeV$ com a finalidade de não excluir dados correspondentes ao decaímento em que estamos interessados.

   Aplicando a mesma lógica ao decaímento $\tau_e = e{+}2\nu$ temos $p_T(e) > \frac{p_T(\tau_e)}{3} = 50\ GeV$. E, novamente, de maneira a não excluir dados de interesse minoramos o limite e usamos a condição mais modesta $p_T(e) > 35\ GeV$.

3. Finalmente, o terceiro destes filtros que aplicamos nesta fase deve-se ao princípio da conservação de carga. Uma vez que o eletrão e o $\tau_h$ são produzidos a partir de dois $\tau$ de sinais opostos o produto das cargas das nossas partículas selecionadas tem de ser negativo.

## III.1 Análise de dados

De forma a procedermos com o objetivo primordial do trabalho, cálculo da secção eficaz do nosso decaímento, é necessário verificar que após a aplicação dos filtros referidos na introdução anterior os resultados satisfazem o que se espera teoricamente.

É então importante nesta parte verificarmos que grandezas como a Massa invariante e a acoplanaridade estão de acordo com o esperado.

Para uma melhor compreensão desta secção é necessário adicionar informação importante quanto aos fundos possíveis presentes, cuja breve explicação foi dada na secção (I). No fundo Drell Yan é necessário distinguirmos dois processos distintos que podem ocorrer:

$$pp \rightarrow Z \rightarrow l^+l^- \tag{8}$$

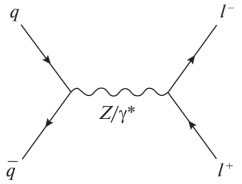


**Figure 7:** 1º Processo do fundo DY

Neste processo o ângulo $\theta$ entre os dois leptões é tão pequeno que é $\approx 0$.

O segundo processo que pode ocorrer neste fundo é simplesmente:

$$pp \rightarrow l^+l^- \tag{9}$$

Neste caso o processo é *back to back* e o ângulo existente entre os dois eletrões é de 180º.

### Massa Invariante

Os conceitos acima referidos vão ser fundamentais para uma análise mais pormenorizada e explicíta acerca do gráfico da Massa Invariante, $M_X = \sqrt{2P_1P_2(1-cos(\theta))}$, dos vários fundos, que se encontra na figura abaixo:

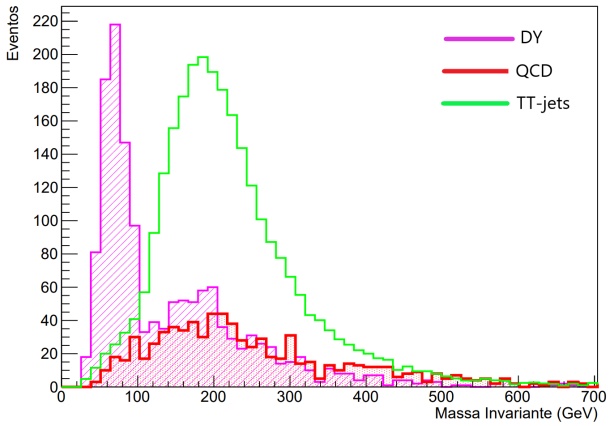


**Figure 8:** Valores da Massa Invariante nos vários fundos

Visualizando o gráfico acima é percetível a existência de dois tipos distintos de picos. Um primeiro, muito próximo de um valor de 70 $GeV$, que apenas ocorre na curva a rosa, pertencente ao gráfico do Drell Yan, e um segundo em torno dos 180 $GeV$ tanto no fundo Drell Yan, como no TT-jets.

O primeiro é explicado pelo facto de o fundo Drell Yan decair num bosão Z que tem uma massa invariante de 90 $GeV$, porém como o bosão decai noutras partículas que não são detetadas pelos vários detetores o pico tem uma energia inferior aos 90 $GeV$.

É de notar que o fundo QCD não tem nenhum pico definido, isto porque sendo um fundo aleatório pode ter todos os valores de $\theta$ possíveis. Porém, como foi explicado na subsecção anterior existem valores mínimos de energia que o $P_1$ e o $P_2$ podem tomar, 35 $GeV$ e 100 $GeV$ respetivamente. Daí resulta que o valor mínimo da Massa Invariante teria que ser, em teoria, superior a 120 GeV, isto porque o valor mais provável do ângulo $\theta$ é de 180º. Como podemos verificar pelo gráfico a maior parte dos acontecimentos acontecem para valores de energia superiores a este, o que está de acordo com o esperado. Os poucos dados que têm um valor de Massa invariante inferior a 120 $GeV$ têm um valor de $\theta$ inferior a 180º.

Quanto ao 2º pico é preciso recorrermos a um gráfico adicional para uma conclusão correta da posição do pico. O gráfico que apresentamos de seguida diz respeito aos valores do momento transversal do eletrão e do *tau*, $p_{e^-}$ e $p_\tau$ respetivamente, no fundo TT-jets:

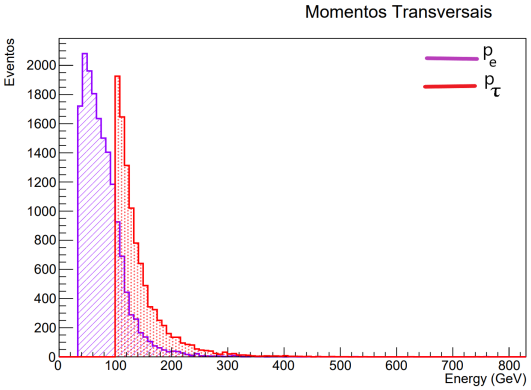


**Figure 9:** Valores do Momento Transversal no fundo TT-jets

Aquilo que podemos retirar de forma imediata é que o valor mais provável do momento transversal do eletrão é de 60$GeV$ e do $\tau$ é de 120 $GeV$. Tendo em conta que o processo do fundo TT-jets é semelhante ao 2º do Drell Yan, 9, o valor da Massa invariante calcula-se a partir da equação:

$$M_x = 2\sqrt{P_e P_\tau} \tag{10}$$

Colocando os valores acima referidos na fórmula chegamos a um resultado de 170 $GeV$, estando próximo do valor do pico a que nos estamos a referir.

### Acoplanaridade

Outra quantidade que demonstra se os nossos filtros são adequados e fazem sentido é a acoplanaridade, cuja equação fundamental é a seguinte:

$$a = \frac{|\Delta\phi|}{\pi} \tag{11}$$

De acordo com a explicação fornecida na introdução desta secção, caso um evento seja do tipo *back to back* o valor de $\Delta\phi$ será de $\pi$, ou seja $a = 1$, fundamental na análise do gráfico que se encontra de seguida.

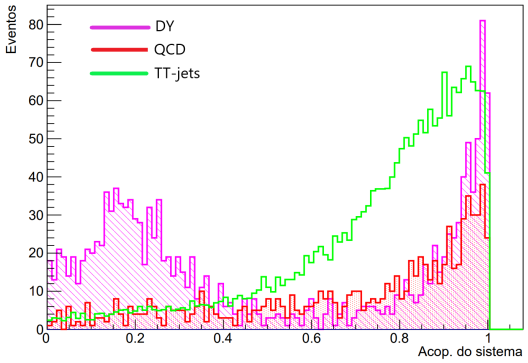


**Figure 10:** Valores da acoplanaridade nos vários fundos

Como podemos observar, no fundo Drell Yan distinguimos perfeitamente dois picos distintos, o que faz sentido visto que neste fundo podem ocorrer dois processos distintos.

No processo ilustrado pela figura (III.1), os valores dos ângulos são muito pequenos o que leva a um menor valor da acoplanaridade, situando-se este perto dos valores de 0.2.

O segundo pico deste fundo, assim como o primeiro do TT-jets tem valores de acoplanaridade muito próximos de 1. Como já foi anteriormente referido, a cinemática do fundo TT-jets é semelhante ao segundo processo do fundo Drell Yan (9), caracterizando-se por ser um processo *back to back* e daí o valor de $a = 1$.

Por último, no fundo QCD, os valores da acoplanaridade encontram-se mais espalhados numa gama mais extensa de valores, devido à sua aleatoriedade, embora consigamos distinguir um pico em $a = 1$, pois o valor do ângulo mais provável neste fundo é respetivamente 180º.

**Resolução do detetor**

Após falar com nosso coordenador, decidimos que seria boa ideia tratar da resolução dos detetores do CMS. Para isso, resolvemos estudar o Calorímetro Hadrónico, dado que se trata do detetor com mais incerteza de medição, e desempenha um papel muito importante no LVL1 *Trigger*. Logo, que existe uma incerteza considerável na medição de energia deste calorímetro, que vai corresponder à incerteza sistemática do detetor.

Como está explicado atrás, quando se dão colisões de protões, dá-se a produção de quarks, que por sua vez produzem *jets*. Estes *jets*, que corresepondem a sucessivos decaimentos de quarks, propagam-se pelos detetores, e deixam grande parte da sua energia no Calorímetro Hadrónico. Então, para estimar a resolução de energia do mesmo, consideraram-se ficheiros de simulação Monte-Carlo que continham dados sobre o momento linear transversal inicial dos *jets*, que vamos denominar por $p^t_{gen\_jet}$, e que também continham informações sobre o momento após interação com o Calorímetro Hadrónico. A esse momento vamos chamar $p^t_{jet}$. Para calcular a resolução em energia, fez-se a distribuição da variável $\Delta p = p^t_{jet} - p^t_{gen\_jet}$ para milhares de eventos simulados.

O gráfico da distribuição encontra-se na seguinte figura:

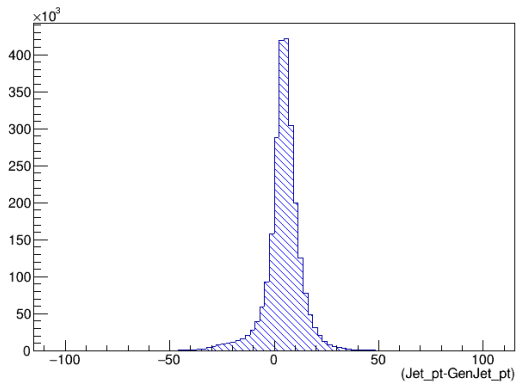


**Figure 11:** Distribuição da diferença entre momento inicial e momento detetado pelo Calorímetro Hadrónico

Podemos notar, por análise ao gráfico, que obtemos uma distribuição em torno do 0, o que quer dizer que na maioria das vezes o momento detetado corresponde ao momento inicial. No entanto, podem acontecer dois casos distintos:

- $\Delta p < 0$ - Situação expectável, porque o momento inicial é, em princípio maior do que o detetado, devido à eficiência limitada do calorímetro.
- $\Delta p > 0$ - Corresponde a casos em que o momento detetado é superior ao inicial, e isto acontece porque *a priori* o calorímetro hadrónico pode identificar outros momentos de outras partículas como sendo *jets*.

Então para estimar a resolução em energia tomamos o valor do desvio padrão da distribuição obtido, que foi $\sigma = 9GeV$

Agora podemos questionar-nos se este é um bom ou mau valor - a resposta é comparar $\Delta p$ com o momento do *jet* detetado ($p^t_{jet}$), ou seja fazer a distribuição $\frac{\Delta p}{p^t_{jet}}$. Na seguinte figura está contemplada essa distribuição:

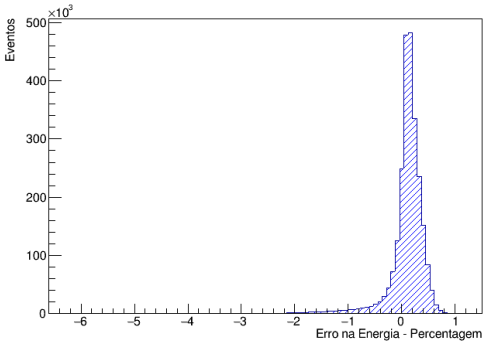


**Figure 12:** Distribuição Relativa

Tomando o desvio padrão desta distribuição obtemos que a resolução, em percentagem, é cerca de 30 %.

# IV Análise Multivariada

## IV.1 Introdução

Como foi mencionado na introdução teórica (I) existem três processos, a que chamamos fundos, que têm decaímentos muito semelhantes ao processo que queremos estudar: o processo de Drell Yan, o processo de Produção inclusiva de pares $t\bar{t}$ e o Fundo QCD.

Estes três fundos, devido à grande semelhança com o nosso sinal, não são totalmente removidos pelos três filtros mencionados na secção (III.0.1) o que fez surgir a necessidade de recorrer a métodos adicionais de tratamento de dados.

Os métodos que escolhemos baseiam-se no uso de técnicas de análise multivariada e fazem uso da biblioteca de ROOT TMVA.

## IV.2 Algoritmos de Classificação

Como mencionado na secção anteior, para desenvolver estes classificadores mais anvançados recorremos à biblioteca TMVA. Esta biblioteca disponibiliza vários algoritmos de classificação no entanto nós escolhemos usar os métodos conhecidos por MLP (Multilayer perceptron) e BDT (Boosted Decision Trees). Estes dois algoritmos são muito semelhantes a um nível fundamental. Ambos têm a mesma finalidade: classificar um conjunto de dados desconhecido em classes (No nosso caso queremos dividir os dados captados pelos detetores em fundo, que inclui os fundos de DY, QCD e do processo inclusivo de produção de pares $t\bar{t}$, e em decaímento elêtrónico). Para além disso ambos usam métodos de classificação baseados na mesma ideia elementar: Dado um conjunto de variávieis de input o algoritmo vai dividir o espaço formado por estas variáveis em regiões que maximizem a correspondência das classes com um conjunto de dados pré-classificados, a que geralmente se chama conjunto de treino (No nosso caso são os ficheiros MC após terem passado pelo processo de filtração).

## IV.3 Possíveis limitações dos algoritmos

A principal limitação destes dois algoritmos é o facto de, como foi referido previamente, a divisão do espaço das variáveis de input que nos dá o nosso critério de seleção ser proveniente de um conjunto de dados pré-classificados. Isto pode ser problemático porque, uma vez que este conjunto corresponde apenas a uma amostra do universo total de dados pode, por azar, existir "bias" na amostra que escolhemos usar como conjunto de treino e, consequentemente, os limites que o algoritmo vai fixar vão estar desviados.

Para combater este problema o que decidimos fazer foi uma subdivisão dos nossos dados de treino em dois grupos. Após esta divisão treinamos dois modelos, um em cada um destes subconjuntos de dados, e comparamos os resultados. Se a nossa amostra não for biased então os resultados obtidos têm de ser independentes do sample que escolhemos utilizar.

## IV.4 Variáveis de Input

Tendo definido anteriormente que o objetivo do *Machine Learning* é distinguir o sinal do conjunto de fundos, é necessário selecionar um conjunto de variáveis discriminantes do sinal e do fundo. Para isso escolhemos as seguintes variáveis:

- Variáveis globais do sistema:
  - massa invariante
  - acoplanaridade
  - momento transversal
- momentos transversais do eletrão e do $\tau$
- acoplanaridade do sistema central (CMS)
- *missing energy*

Também selecionámos duas variáveis muito específicas que permitem distinguir o sinal dos fundos:

- *Matching Mass*

$$f_m = M_X - \sqrt{s\xi_1\xi_2} \tag{12}$$

  Onde s é a energia total dos protões, e $\xi_i$ é a fração de momento perdido pelos mesmos durante a interação. $M_X$ é calculado a partir do CMS, enquanto o $\sqrt{s\xi_1\xi_2}$ é calculado a partir do PPS

- *Matching Rapidity*

$$f_r = Y_X - \frac{1}{2}\log(\frac{\xi_1}{\xi_2}) \tag{13}$$

  onde $Y_X$ é a rapidade do sistema central (CMS) e $\frac{1}{2}\log(\frac{\xi_1}{\xi_2})$ é a rapidade medida pelo PPS

A partir deste momento as amostras de fundo foram unificadas numa só amostra, logo não é de interesse distinguir os vários fundos, portanto, o algoritmo de machine learning foi treinado para separar o sinal da soma dos fundos.

**Amostras de treino para algumas variáveis de input**

Apresentamos agora gráficos com as distribuições das variáveis de input que considerámos mais relevantes e que de facto se revelaram ser as mais importantes no treino dos classificadores na secção seguinte, e o porquê das diferenças apresentadas entre sinal e fundo. Isto é importante porque se houverem diferenças significativas entre o sinal e o fundo para cada variável quer dizer que o classificador vai conseguir encontrar diferenças entre os mesmos, e consequentemente haverá progresso na separação sinal-fundo.

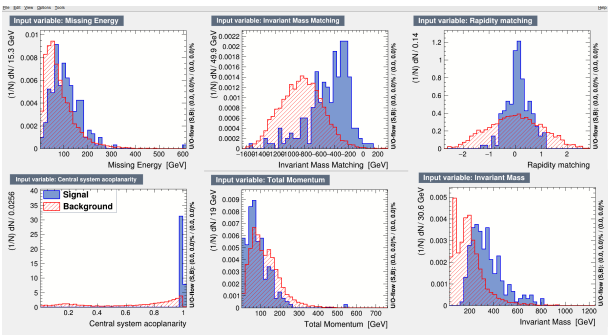


**Figure 13:** Amostras de treino do classificador para algumas variáveis de input

Começando pelo primeiro gráfico, (*missing* energy) notamos que a energia que falta é maior no sinal, o que faz sentido, dado que indicia a presença de energia não detetada, e confirma a existência e não deteção dos neutrinos. Nos fundos a *missing energy* é menor mas não é desprezável porque no fundo *ttjets* o bosão W pode decair num leptão + neutrino, o que indica perda de energia. No QCD também ocorre perda de energia, já que os decaimentos que podem ocorrer *jets* podem gerar neutrinos.

No gráfico da *mass matching* do sinal podemos verificar que a maior parte dos valores são negativos - isto explica-se pelo facto de que é perdida energia no CMS e

que o PPS é mais preciso na reconstrução de certos eventos. Vimos também que no CMS existem erros no cálculo dos momentos devido à resolução limitada do calorímetro hadrónico, e o erro associado ao cálculo de $P_1P_2$ é superior ao erro de $\xi_1\xi_2$ o que faz com que o valor de $f_m$ (eq. 12) seja negativo. O que é relevante é que para o sinal a maior parte dos valores estão próximos de 0, logo podemos aferir um matching entre o sistema central e o PPS, apesar de se perder energia no CMS. Para o fundo a história é diferente, dado que os protões que entram para o cálculo de $f_m$ são protões de *pile-up*, o que origina uma distribuição aleatória. O treino do classificador para esta variável é muito relevante para a separação sinal-fundo, porque como se pode ver no gráfico a distribuição varia muito do sinal para o fundo.

No gráfico seguinte (*Rapidity Matching*) é importante frizar que o erro cometido na medição das direções é reduzido tanto no CMS como no PPS, pelo que se obtem uma distribuição simétrica em torno de 0 para o caso do sinal. No caso do fundo, a distribuição é mais uma vez aleatória, porque os protões que entram nas simulações Monte-Carlo são de *pile-up*.

Analisando agora o gráfico da Acoplanaridade, vemos claramente que a distribuição obtida para o fundo corresponde à sobreposiçaõ dos vários fundos (fig.[10]), como analisado na secção (III.1). Para o sinal espera-se obter uma acoplanaridade igual a 1, porque o processo é *back-to-back*. Como podemos ver no gráfico, obtem-se exatamente isso com uma precisão elevadíssima (já vimos que a medição de ângulos no sistema central ou no PPS é muito precisa)

Para o gráfico do momento total podemos notar que no fundo existem dois picos, um mais próximo de 0 e um na zona dos 150 $GeV$ sabemos que esta distribuição faz sentido, porque o processo 8 do fundo Drell Yan tem momento total positivo, e o processo 9 tem momento total 0. O pico próximo de 0 é mais significativo porque todos os processos dos outros fundos têm a mesma cinemática do que o processo (8) do Drell Yan. O sinal, como é de esperar, terá de ter momento total próximo de 0, já que o nosso processo é $\tau^+\tau^-$ (*back-to-back*). Este momento não é exatamente 0 devido à resoluçaõ limitada dos calorímetros, como já vimos.

Finalmente, para a massa invariante, para os fundos obtemos a sobreposição das distribuições do gráfico da Fig.[8], como era de esperar, e para o sinal obtemos um pico mais disperso que tem como valor mais provável cerca de 240 $GeV$. Este valor corresponde à massa invariante do nosso processo *back-to-back* $\tau^+\tau^-$ ($M_X = \sqrt{2P_1P_2(1-cos(\theta))}$), se considerarmos que o momento do $\tau$ é cerca de 120 $GeV$, e o ângulo cerca de 180º.

Tendo selecionado as variáveis de input, treinámos dois classificadores - MLP e BDT.

## IV.5 Output do treino de classificadores

Nesta subsecção apresentamos os resultados da análise multivariada: como o resultado da separação entre sinal e fundo (objetivo principal), correlação entre variáveis de input, eficiência do classificador ultilizado.

**Eficiência do classificador**

No seguinte gráfico está contemplada a eficiência do de cada classificador:

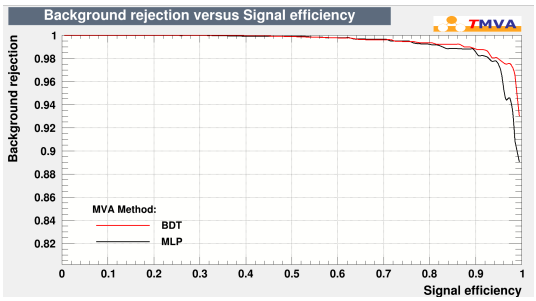


**Figure 14:** Rejeição do fundo vs eficiência do sinal para cada classificador

Por análise ao gráfico podemos ver que a rejeição do fundo foi sempre superior para o classificador BDT. Quanto maior a rejeição do fundo à medida que se aumenta a eficiência do sinal melhor, porque o nosso objetivo é separar o sinal do fundo. Então, após retirar esta conclusão decidimos apenas usar os resultados do classificador BDT.

**Matrizes de correlação**

Nas seguintes figuras está contemplada a correlação entre as variáveis de *input* utilizadas:

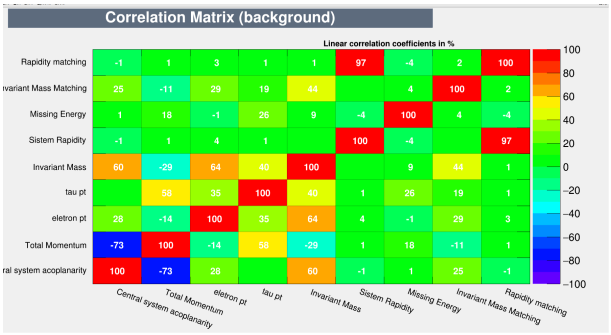


**Figure 15:** Correlação das variáveis para o fundo

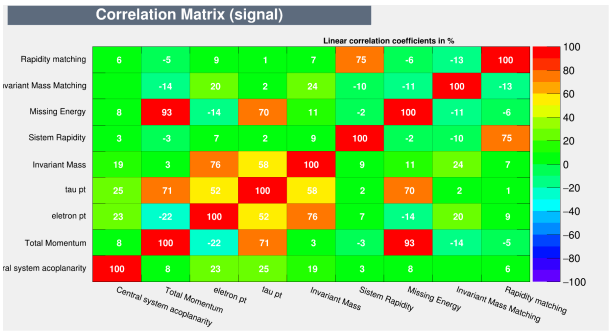


**Figure 16:** Correlação das variáveis para o sinal

Podemos ver que nos dois casos a correlação das variáveis mais importantes (as 6 usadas anteriormente) é no geral baixa, o que indica que estas têm um alto grau de independência, sendo todas elas muito relevantes na separação sinal-fundo. As variáveis que apresentam maior

correlação (consequentemente mais dependentes umas das outras) são por exemplo a rapididade do sistema com a *rapidity matching*, o que quer dizer que a utilização de apenas uma delas seria suficiente para o funcionamento igual do algoritmo.

**Separação Sinal-Fundo**

Finalmente, antes de embarcar na obtenção de um limite para a secção eficaz, apresentamos um gráfico que evidencia o grau de separação final do sinal e do fundo, e confirmamos o sucesso e relevância do *machine learning* do classificador BDT neste processo.

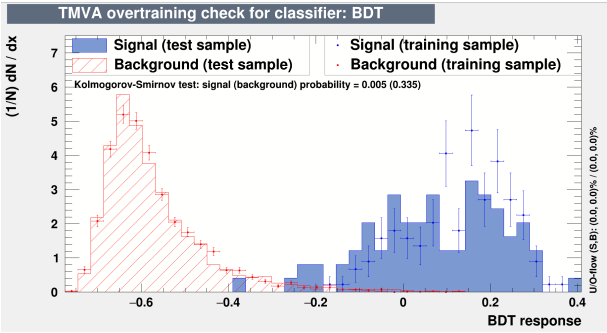


**Figure 17:** Separação sinal-fundo após treino e execução do classificador BDT e probabilidades obtidas

## IV.6 Obtenção do limite para a secção eficaz

Para concluir o nosso trabalho procedemos à última parte, ou seja, o cálculo do limite para obter a secção eficaz. De seguida, será apresentada uma imagem ilustrativa deste processo:

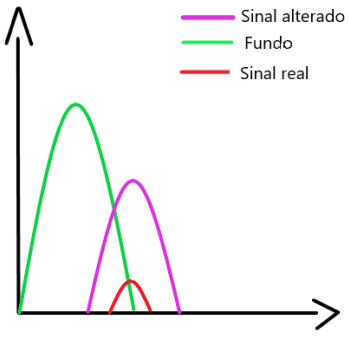


**Figure 18:** Imagem ilustrativa para o cálculo do limite da secção eficaz

A figura (17) ilustra as distribuições de output do algoritmo BDT no caso em que o número de acontecimentos do sinal e do fundo são os mesmos. Infelizmente, o número de acontecimentos do sinal é muito inferior (quatro ordens de magnitude) do número de acontecimentos do fundo. Por tanto, se graficarmos as distribuições normalizadas à secção eficaz dos processos, obteríamos um gráfico parecido à imagem (18), onde a curva a verde representa a distribuição do fundo e a curva a vermelho a distribuição do sinal. De facto, o sinal não é distinguível do fundo, sendo a incerteza da distribuição a verde dominante na região do sinal. De seguida, as distribuições mencionadas foram utilizadas como input duma ferramenta que permite calcular o limite teórico da secção eficaz do sinal. Para isso, a ferramenta multiplica a distribuição do sinal por um coeficiente, designado "signal strength" (R). O resultado final é a distribuição a roxo, que tem a mesma forma da distribuição vermelha mas, desta vez, é distinguível da distribuição verde, sendo a diferença entre o número de acontecimentos contidos no pico da distribuição roxa ($N_s$) menos o número de acontecimentos do fundo na mesma região ($N_f$) maior que $3\sqrt{N_f}$. Por tanto, o valor de R permite calcular o limite teórico da secção eficaz do sinal: sendo $\sigma_{teo} = 1fb$, a quantidade $\sigma^* = R\sigma_{teo}$ é o mínimo valor da secção eficaz que permite distinguir o sinal do fundo.

Foi possível determinar que os valores do parâmetro R variam entre 20 e 180.

## References

[1] https://docs.google.com/spreadsheets/d/1IArkTu1RNGOfTK-WlZ84AdHeWMnDUDOkEnLs0QgVsgo/edit?usp=sharing

[2] https://iopscience.iop.org/article/10.1088/1748-0221/16/12/P12014

[3] https://root.cern

[4] https://root.cern/manual/tmva/

[5] https://arxiv.org/abs/1005.2841v1

[6] https://pdg.lbl.gov/2022/html/authors_2022.html