Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciências
Instituto Politécnico

Sarita de Miranda Rimes

Filtragem inversa não-linear para estimação de sinais em calorímetros operando a alta taxa de eventos

Nova Friburgo
2021

Sarita de Miranda Rimes

Filtragem inversa não-linear para estimação de sinais em calorímetros operando a alta taxa de eventos

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Modelagem Computacional, da Universidade do Estado do Rio de Janeiro.

Orientador: Prof. Dr. Bernardo Sotto-Maior Peralva

Nova Friburgo

2021

CATALOGAÇÃO NA FONTE
UERJ/REDE SIRIUS/BIBLIOTECA CTC/E

R575 Rimes, Sarita de Miranda.
Filtragem inversa não-linear para estimação de sinais em calorímetros operando a alta taxa de eventos / Sarita de Miranda Rimes. – 2021.
86 f. : il.

Orientador: Bernardo Sotto-Maior Peralva.
Dissertação (mestrado) - Universidade do Estado do Rio de Janeiro, Instituto Politécnico.

1. Processamento de sinais – Métodos de simulação - Teses. 2. Estimativa de parâmetros - Teses. 3. Ruído – Teses. 4. Calorimetria – Teses. I. Peralva, Bernardo Sotto-Maior. II. Universidade do Estado do Rio de Janeiro. Instituto Politécnico. III. Título.

CDU 621.391

Bibliotecária Cleide Sancho CRB7/5843

Sarita de M. Rimes
Assinatura

03/03/2021
Data

Sarita de Miranda Rimes

Filtragem Inversa Não-Linear para Estimação de Sinais em Calorímetros Operando a Alta Taxa de Eventos

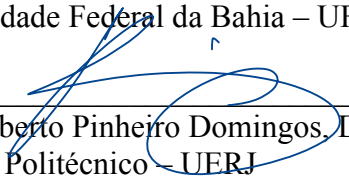
Dissertação apresentada como requisito parcial para obtenção do título de Mestre, ao Programa de Pós-Graduação em Modelagem Computacional do Instituto Politécnico, da Universidade do Estado do Rio de Janeiro.

Aprovado em 26 de fevereiro de 2021.

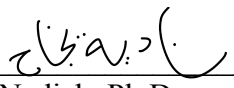
Banca examinadora:

Prof. Bernardo Sotto-Maior Peralva, D.Sc. – Orientador
Instituto Politécnico – UERJ

Prof. Eduardo Furtado de Simas Filho, D.Sc.
Universidade Federal da Bahia – UFBA



Prof. Roberto Pinheiro Domingos, D.Sc.
Instituto Politécnico – UERJ



Profa. Nadia Nedjah, Ph.D.
Universidade do Estadual do Rio de Janeiro – UERJ

Nova Friburgo

2021

DEDICATÓRIA

Aos meus pais, pelo apoio e incentivo constantes.

AGRADECIMENTOS

Primeiramente, agradeço a Deus, pela existência e pela oportunidade de mais uma vida. Sou grata por ter tido saúde e determinação para conseguir completar mais esta etapa.

Aos meus pais, que sempre me incentivaram e apoiaram, em todos os momentos, e sempre serviram como base para que eu pudesse chegar até aqui.

Ao meu orientador, Bernardo, por todo profissionalismo e humanidade. Agradeço pelos ensinamentos e por todas as vezes em que me ouviu pacientemente, tentando encontrar sempre a melhor solução para os desafios que se apresentavam. Esses anos, com certeza, se tornaram muito menos árduos graças ao seu apoio e a sua excelente orientação e didática.

Aos professores Luciano e Seixas, pelas conversas, dicas e orientações. Agradeço por terem acrescentado tanto a este trabalho, separando, diversas vezes, em meio a vidas tão corridas, algum tempo para ajudar e compartilhar conhecimento e experiências. Esta dissertação tem muito de vocês. Obrigada!

Ao grupo ATLAS Brasil, pela oportunidade de compartilhar e receber conhecimento.

À Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ), pelo apoio financeiro concedido durante todo esse período, importantíssimo para que este trabalho pudesse existir e ser feito com excelência.

A vitória aguarda aquele que tem tudo em ordem - *sorte* é como as pessoas chamam isso.

Roald Amundsen

RESUMO

RIMES, S. M. *Filtragem inversa não-linear para estimação de sinais em calorímetros operando a alta taxa de eventos.* 2021. 86 f. Dissertação (Mestrado em Modelagem Computacional) – Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, 2021.

Diversas áreas, atualmente, trabalham com problemas onde a estimação de parâmetros se coloca de forma crucial. Muitas vezes, a solução de tais problemas é dificultada pela presença de ruído nos dados observados. Para os casos onde esse ruído é Gaussiano, diversas técnicas lineares são bastante difundidas, porém, quando características Gaussianas são perdidas, a correta descrição do ruído se torna necessária. Em experimentos de física de altas energias, os sistemas de calorimetria são responsáveis por absorver e amostrar a energia de partículas provenientes das colisões. No LHC, com o crescente aumento da luminosidade, o fenômeno de empilhamento de sinais pode ser observado nos canais de leitura dos calorímetros do experimento ATLAS. Tal efeito acrescenta ao ruído, inicialmente apenas eletrônico e Gaussiano, uma componente não-linear que degrada a eficiência de métodos tipicamente empregados. Dessa forma, nesta dissertação, é apresentado um método não-linear baseado em um estimador de máxima verossimilhança utilizando uma distribuição Lognormal para modelar o ruído. Este método, chamado de MLE Lognormal, é comparado a outros três, lineares: OF2, atualmente utilizado no calorímetro de telhas (TileCal) do ATLAS, COF e MLE Gaussiano, o mesmo estimador, porém assumindo uma distribuição Gaussiana para o ruído. Além disso, uma análise sobre a dependência estatística das variáveis aleatórias do sinal é apresentada, utilizando-se a informação mútua presente nestas e a técnica da ICA para pré-processamento. Várias condições de ocupação e luminosidade foram consideradas nas análises e foi possível observar que a distribuição Lognormal apresenta um ajuste melhor aos dados de ruído quando comparada à distribuição Gaussiana. As análises de eficiência dos métodos condizem com tais observações. Foram observadas melhoras de 3,14%, 28,17% e 3,23% para o MLE Gaussiano, OF2 e COF, nos dados simulados, e 5,39%, 26,59% e 8,34%, nos dados reais, respectivamente. Também foi possível notar alta correlação entre as amostras do ruído, percebendo-se uma expressiva diminuição da dependência estatística entre as variáveis aleatórias após a aplicação da ICA, com redução, neste parâmetro, de 13,26% para 7,00% nos dados simulados e de 18,13% para 4,44% nos dados reais.

Palavras-chave: Estimação de parâmetros. Ruído lognormal. Estimadores de máxima verossimilhança. Calorimetria de altas energias.

ABSTRACT

RIMES, S. M. *Nonlinear inverse filtering for estimation of signals in calorimeters operating at high event rate.* 2021. 86 f. Dissertação (Mestrado em Modelagem Computacional) – Instituto Politécnico, Universidade do Estado do Rio de Janeiro, Nova Friburgo, 2021.

Currently, several areas deal with parameter estimation problems as a crucial step. Often, the solution of such problems is challenged by the presence of noise in the observed data. For cases where the noise is Gaussian, several linear techniques are extensively used, however, when Gaussianity properties are lost, a proper noise description is required. In high-energy experiments, the calorimeter systems are responsable for absorbing and sampling the energy from particled produced by collisions. In the LHC, with the increase of its luminosity level, the signal pile-up effect can be observed in the readout channels from the calorimeter systems in the ATLAS experiment. Such effect introduces a nonlinear component to the noise which was previously modeled by a Guassian distribution, degrading the efficieny of typical energy estimation methods. Therefore, in this work, a nonlinear method based on the Maximum Likelihooh Estimatior (MLE) that uses a Lognormal model for the noise is presented. The performance from the proposed method, called MLE Lognormal, is compared with thre other linear methods: OF2, which is currently used in the ATLAS main hadronic calorimeter (TileCal), COF and MLE Gaussian, which assumes a Gaussian noise for the MLE. Additionally, an analysis on the statistical dependence of the noise random variables is presented, using the ICA technique as a pre-processing in order to cope with the statistical dependency. Several channel occupancy conditions and luminosity were considered where it was possible to observe that the Lognormal distribution presents a better fitting when compared to the Gaussian approach. The performance achieved by the methods are consistent with such observations. Improvements of 3.14%, 28.17% and 3.23% were observed for the MLE Gaussian, OF2 and COF, in the simulated data, and 5.39%, 26.59% and 8.34%, in the real data, respectively. It was also possible to assess the statistical dependence between the noise time samples, where the ICA can be a promising alternative to decrease the mutual information between the noise samples. Reductions from 13.26% to 7.00% in the simulated data and from 18.13% to 4.44% in the real data were observed in this parameter.

Keywords: Parameter estimation. Lognormal noise. Maximum likelihood estimators. High-energy calorimetry.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

INTRODUÇÃO

Atualmente, o processo de estimação de parâmetros na área de análise e processamento de sinais tem se tornado cada vez mais utilizado, em diversos ramos das ciências exatas. Tal processo tem como intuito adquirir informações sobre fenômenos físicos lidos e amostrados por seus sistemas eletrônicos associados. Como exemplo, é possível citar sistemas de radares e sonares, assim como aqueles relacionados a reconhecimento de voz ou fala, além de análise de imagens. Um pouco mais distante, é possível citar, ainda, ciências como a biomedicina e a sismologia, que também se utilizam desses artifícios para estudos mais precisos.

Apesar de nos projetos de sistemas eletrônicos serem, no geral, utilizados, inicialmente, sinais analógicos, existe uma tendência de se trabalhar cada vez mais com sinais digitais. A partir daí, assim como é o caso dos exemplos citados no parágrafo anterior, a amostragem de problemas físicos reais é normalmente feita por um conversor analógico-digital, que permite que o sinal seja fornecido com base em sequências discretas no tempo.

Outra área que utiliza sinais eletrônicos com o fim de fazer a amostragem de seus estudos é a física de altas energias. Em grande parte, experimentos modernos aplicados nessa área fazem uso de sistemas de calorimetria como sua base. Aqui, o foco se encontra na absorção e amostragem da energia de partículas que são detectadas pelo calorímetro em questão, de forma que, a partir desses dados, a energia referente possa ser reconstruída e a partícula identificada.

Com o constante crescimento da capacidade de processamento e armazenamento de dados em processadores e memórias, aliado a experimentos cada vez mais complexos e que fornecem quantidades sempre maiores de informações, torna-se necessário o desenvolvimento de novas técnicas de análises de dados, de modo a estudar de forma fidedigna os dados coletados. Essas técnicas, no geral, exigem uma base matemática e uma aplicação computacional específicas para cada caso, que deve ser pensado e aprofundado cuidadosamente.

Em experimentos modernos na área de física de altas energias, ou física de partículas, a quantidade de dados vem aumentando em taxas significativas com o passar dos anos. Este é o caso do acelerador de partículas LHC (do inglês, *Large Hadron Collider*), ambiente de estudo desta dissertação, de onde provêm os dados utilizados para análise dos métodos aqui explorados. O LHC é o maior e mais energético acelerador de partículas do mundo e é composto por quatro experimentos principais, sendo o maior deles o ATLAS. Dentro dos anéis supercondutores do LHC, feixes de prótons são acelerados a velocidades próximas à da luz e colisões ocorrem em pontos estratégicos, produzindo subprodutos que são medidos pelos diversos sistemas altamente calibrados.

Os experimentos de física de altas energias, como o LHC, visam aumentar um

parâmetro chamado de luminosidade, que está diretamente relacionado à densidade dos feixes de partículas: quanto maior a densidade, maior a luminosidade presente no experimento. O aumento deste parâmetro, algo que está previsto para o LHC nos próximos anos, gera um crescimento expressivo na quantidade de dados produzidos pelas colisões. Como consequência disso, surgem problemas nos sistemas de aquisição de dados e, consequentemente, na estimação de parâmetros importantes, os quais são necessários para a correta identificação de partículas de interesse.

Motivação

O Calorímetro de Telhas, ou TileCal, é o maior calorímetro hadrônico do experimento ATLAS e responsável por fornecer os dados reais utilizados nesta dissertação, sendo os simulados gerados com base em seus parâmetros. No TileCal, as partículas provenientes das colisões são absorvidas por placas de aço e amostradas por telhas cintilantes, e sua eletrônica de leitura fornece, após todo um processo, um sinal digital de sete amostras que possui amplitude proporcional à energia da partícula correspondente.

O sinal digital surge corrompido por ruído, inicialmente, Gaussiano, proveniente da eletrônica do calorímetro. O aumento da luminosidade, no entanto, introduz ao sinal uma componente não-Gaussiana, que pode ser tratada como ruído adicional. Tal componente aparece devido a sinais de colisões adjacentes que se apresentam dentro de uma mesma janela de leitura, o que é chamado de *empilhamento de sinais*, e se torna responsável pela degradação de métodos lineares tipicamente utilizados, os quais apresentam dificuldades para trabalhar de forma eficiente com o ruído não-linear. Visto que o programa de atualização do LHC prevê um aumento contínuo da luminosidade, o fenômeno de empilhamento de sinais tende a aumentar nos canais de leitura do TileCal. Desta forma, novas abordagens estão sendo testadas a fim de mitigar este problema.

Objetivo

Com o intuito de tentar resolver o problema de estimação da energia em condições de empilhamento de sinais, que insere ao ruído características não-Gaussianas, esta dissertação apresenta um método baseado em um estimador de máxima verossimilhança (ou MLE, do inglês, *Maximum Likelihood Estimation*), utilizando uma distribuição Lognormal para modelar o ruído. Visto que o empilhamento de sinais corresponde à soma de sinais da colisão atual e de colisões adjacentes, e que a deposição de energia se assemelha à distribuição exponencial, a escolha da distribuição Lognormal se dá pela sua aproximação da distribuição Gama, a qual idealmente representa a soma de distribuições exponenciais,

porém não possui distribuição multivariada bem definida. Além disso, os parâmetros da distribuição Lognormal podem ser facilmente estimados através de um conjunto de dados. Na abordagem proposta, o parâmetro que se deseja estimar deve ser tal que a PDF (do inglês, *Probability Density Function*) escolhida para modelagem seja maximizada. Aqui, este parâmetro é a amplitude do sinal digitalizado adquirido.

Organização do Texto

O Capítulo 1 apresenta o ambiente de trabalho desta dissertação. Inicialmente, é feita uma exposição geral sobre o CERN, seguida de uma explicação sobre o LHC, tanto em termos físicos quanto em características de funcionamento. Posteriormente, o ATLAS é detalhado e é mostrado como funciona seu sistema de filtragem online. A seção seguinte é responsável pela parte de calorimetria de altas energias, expondo brevemente os calorímetros eletromagnético e hadrônico do ATLAS. Por último, o TileCal é apresentado de forma mais completa.

No Capítulo 2, são desenvolvidos métodos lineares que já são bem postos em problemas de estimação de parâmetros. O primeiro a ser apresentado é o Filtro Ótimo, que é atualmente utilizado no ATLAS, e, imediatamente após, o COF. Na seção seguinte, é mostrado como funcionam os estimadores baseados no MLE, e então o MLE Gaussiano é abordado.

O Capítulo 3 apresenta o método não-linear proposto para tentar resolver o problema de empilhamento de sinais, o MLE com uso da distribuição Lognormal. Além disso, um pré-processamento para este método também é abordado, tratando dos conceitos de dependência e independência estatística.

No Capítulo 4 são apresentados os resultados obtidos após a modelagem do ruído e a análise de eficiência dos métodos. Primeiramente, os parâmetros utilizados para tais avaliações são expostos. Em seguida, os resultados são mostrados para os dados simulados e reais, nesta ordem. Para ambos, são feitos, também, estudos da dependência estatística entre as amostras.

Por fim, em seguida, é apresentada uma discussão com as conclusões e trabalhos futuros que possam ser desenvolvidos a partir desta dissertação.

1 AMBIENTE DE FÍSICA EXPERIMENTAL DE ALTAS ENERGIAS

Neste capítulo, é apresentado o ambiente de trabalho utilizado como base para a análise de eficiência dos métodos estudados. Na Seção 1.1, é feita uma breve exposição do CERN, o maior laboratório de física de altas energias do mundo, enquanto a Seção 1.2 explica como funcionam os experimentos no LHC, apresentando de forma sucinta os principais detectores que o compõem. Já a Seção 1.3 apresenta o experimento ATLAS, seguida pela Seção 1.4, que detalha o sistema de calorimetria desse detector. O TileCal, maior calorímetro hadrônico do ATLAS e responsável por fornecer os dados usados para explorar e comparar os métodos propostos nesta dissertação, é desenvolvido na Seção 1.5.

1.1 O CERN

O CERN (CERN, 2020a) (do francês, *Organisation Européenne pour la Recherche Nucléaire*) é um centro de pesquisas na área de física nuclear e de partículas, que faz uso de equipamentos científicos da mais alta complexidade para estudar a fundo partículas fundamentais que compõem a estrutura da matéria, dando forma a tudo o que é conhecido atualmente. O início de sua construção se deu no ano de 1955, na cidade de Genebra, localizada na fronteira entre a França e a Suíça, após quatro anos da primeira resolução que estabelecia sobre a criação de um centro de pesquisas nucleares na Europa.

São 23 os estados membros que administram o CERN, ou seja, países que cooperam financeiramente com os programas de pesquisa e possuem, cada um, dois representantes no conselho do CERN, podendo votar e interferir em decisões importantes da Organização. O CERN recebe também financiamento de diversos outros países e instituições, membros ou não-membros. Além disso, alguns países e organizações estão na categoria de observadores, enquanto outros possuem acordos de cooperação ou têm contato científico ativo. A Figura 1 mostra uma imagem recente da vista aérea do CERN e sua estrutura de prédios, já que os experimentos são construídos abaixo da superfície terrestre.

Os experimentos do CERN têm como objetivo principal o estudo de partículas elementares da física, descritas como aquelas que não são formadas por quaisquer outras partículas. Pode-se destacar a descoberta dos Bósons W e Z (ARNISON et al., 1983; BANNER et al., 1983) e mais recentemente o bóson de Higgs (THE ATLAS COLLABORATION, 2016). No entanto, sua importância no desenvolvimento de novas tecnologias também é notável, como, por exemplo, o *World Wide Web* (CERN, 2020b), criado inicialmente para facilitar a comunicação e o compartilhamento de informações entre cientistas, mas que acabou se espalhando por todo o mundo.

As descobertas feitas por pesquisadores do CERN, juntamente com as tecnologias

Figura 1 - Vista aérea do CERN.

Fonte: CERN (2012).

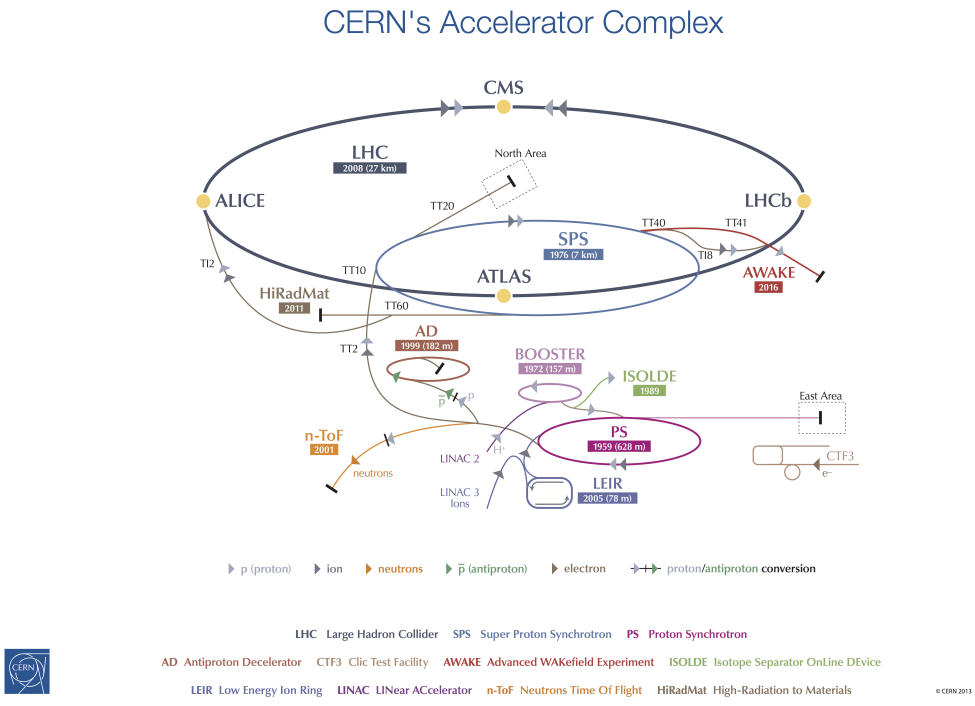
desenvolvidas e difundidas, vêm auxiliando há décadas na evolução de vários campos da ciência e da engenharia. Atualmente, o maior experimento em operação no CERN é o LHC (do inglês, *Large Hadron Collider*), que faz parte do ambiente de trabalho deste estudo e será detalhado na Seção 1.2. No entanto, o centro de pesquisa também inclui outros experimentos e campos de pesquisa, como pode ser observado na Figura 2.

1.2 **Large Hadron Collider (LHC)**

O LHC (EVANS; BRYANT, 2008), ou, traduzindo para o português, Grande Colisor de Hádrons, é uma máquina construída no CERN e conhecida por ser o maior e mais energético acelerador de partículas do mundo, sendo responsável por experimentos que levaram a importantes descobertas na área da física de altas energias (CERN, 2020a).

Composto por dois anéis supercondutores de cerca de 27 quilômetros de circunferência cada, o LHC é o principal experimento do CERN e fica localizado em um túnel construído a aproximadamente 100 metros abaixo do nível do solo. Dentro desses anéis, feixes de prótons são acelerados a velocidades próximas a da luz, em direções opostas, e colisões entre eles ocorrem a uma taxa máxima de 40 MHz, em intervalos de 25 nanossegundos, em locais estratégicos, onde detectores são instalados, de forma a coletar os dados de partículas provenientes de tais colisões. O LHC vem realizando experimentos com êxito desde 2010, colidindo grupos de até 10^{11} prótons, com energia de centro de massa de 14 TeV e luminosidades programadas de $10^{34} cm^{-2} s^{-1}$, além de íons pesados, com uma energia de 2,8 TeV por núcleo de chumbo e luminosidade de até $10^{27} cm^{-2} s^{-1}$. Há, ainda, um aumento da luminosidade previsto para os experimentos que serão realizados nos próximos anos (ROCCA; RIGGI, 2014; BRÜNING; ROSSI, 2019).

Figura 2 - Complexo do acelerador do CERN.



Fonte: CERN (2013a).

A luminosidade (HERR; MURATORI, 2003) tem relação direta com a razão entre o número de interações de prótons por segundo e a seção transversal do feixe, como pode ser visto da Equação (1), onde N é o número de prótons em cada feixe, t é o tempo entre as colisões e S é a seção transversal do feixe (RUGGIERO, 2004).

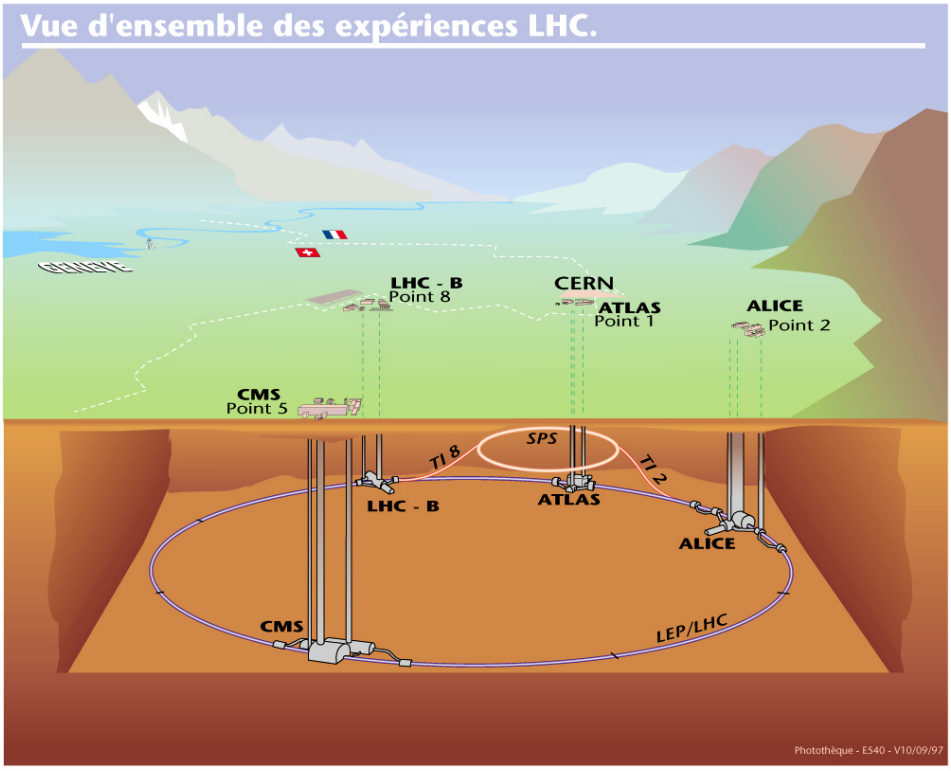
$$L \propto \frac{N^2}{t \cdot S}. \tag{1}$$

Aqui, considera-se que uma partícula de um feixe pode colidir com qualquer outra partícula de outro feixe que esteja sendo acelerado ao mesmo tempo na direção oposta. Sendo assim, contanto que os outros parâmetros que fazem parte da definição da luminosidade não sejam alterados, quanto maior a densidade do feixe, ou seja, quanto mais prótons estiverem contidos nele, mais interações ocorrerão em uma colisão. Atualmente, o LHC opera com um número médio de interações pp por colisão de aproximadamente 60, e a previsão é de que se atinja valores próximos a 200 nos próximos períodos de operação após as fases de atualizações dos experimentos serem completadas (HUFFMAN, 2014).

Quatro experimentos principais compõem o LHC: o ATLAS (do inglês, *A Toroidal LHC ApparatuS*) (THE ATLAS COLLABORATION, 2008), o CMS (do inglês, *Compact Muon Solenoid*) (THE CMS COLLABORATION, 2008), o ALICE (do inglês, *A Large*

Ion Collider Experiment) (THE ALICE COLLABORATION, 2008) e o LHCb (do inglês, *Large Hadron Collider-beauty*) (THE LHCb COLLABORATION, 2008). Os dois primeiros são detectores de propósito geral, ou seja, trabalham fazendo investigações diversas durante as operações, cobrindo um vasto programa de física. Já os dois subsequentes funcionam com objetivos específicos, sendo o ALICE projetado para a detecção de íons pesados, com foco em QCD (do inglês, *Quantum Chromodynamics*), que é o setor de interação forte do Modelo Padrão (HOLLIK, 2010), enquanto o LHCb tem como meta principal estudar a nova física em violações de CP (do inglês, *Charge+Parity*) (KHRIPLOVICH; LAMOREAUX, 1997), além de decaimentos raros de hádrons do tipo *beauty* e *charm*. O maior desses quatro experimentos é o ATLAS e será detalhado mais a frente como o ambiente de trabalho deste estudo. Na Figura 3 podem ser vistos o LHC e seus quatro experimentos principais, com seus respectivos posicionamentos.

Figura 3 - Anéis do LHC e seus experimentos principais.

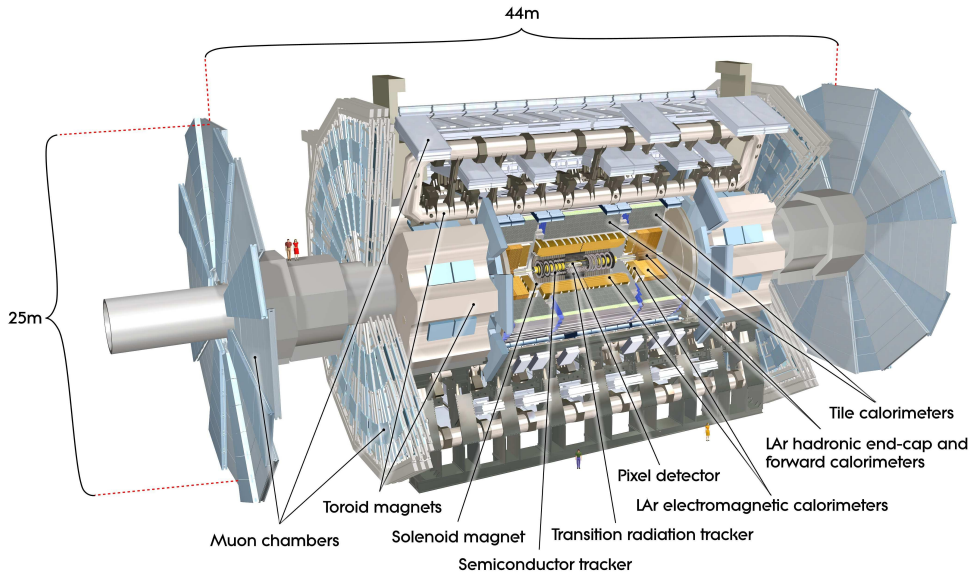


Fonte: CERN (1998).

1.3 A Toroidal LHC ApparatuS (ATLAS)

O ATLAS é o maior experimento do LHC e consiste em um grande detector de cerca de 25 metros de altura e 44 metros de comprimento, pesando aproximadamente 7.000 toneladas, como pode ser visto na Figura 4.

Figura 4 - Vista cortada lateral do detector ATLAS.



Fonte: THE ATLAS COLLABORATION (2008).

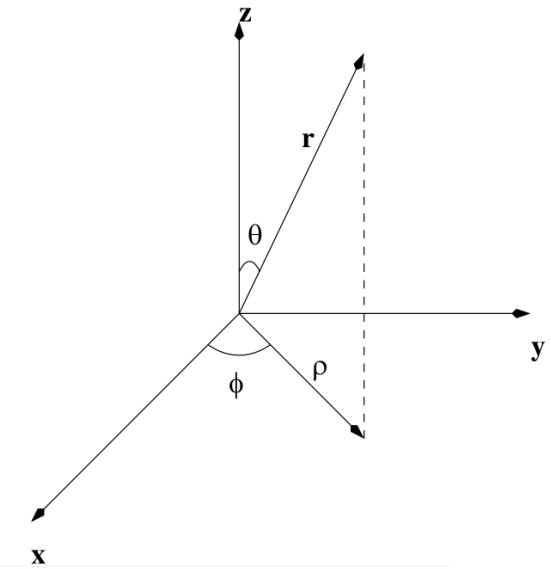
Em experimentos com feixes, no geral, o sistema de coordenadas utilizado não é polar, pois é mais adequado escolher um sistema que acompanhe a direção dos feixes das partículas que possam advir das colisões. Dessa forma, o sistema de coordenadas do ATLAS é apropriado ao formato cilíndrico de seus detectores, sendo definido de forma que o ponto de origem é fixado no ponto de interação. Ao longo da direção do feixe está a coordenada z e o plano $x - y$ é transversal ao feixe, sendo o eixo x positivo apontando da origem para o centro dos anéis do LHC e o eixo y positivo indo da origem para cima. O detector é dividido em dois lados, A e C, sendo o primeiro aquele que se encontra na parte positiva do eixo z e o segundo na parte negativa. Os ângulos ϕ e θ são medidos em volta do eixo do feixe, z, variando no plano $x - y$ (em azimute), e com relação ao eixo z, respectivamente. Existe ainda mais uma variável, definida como η e dada pela Equação (2), que é a chamada pseudo-rapidez. A Equação (3) apresenta uma relação de ϕ com x e y.

$$\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right] \tag{2}$$

$$\phi = \tan^{-1}\left(\frac{x}{y}\right) \tag{3}$$

A Figura 5 ilustra o esquema de coordenadas do ATLAS, onde **r** e $\boldsymbol{\rho}$ são vetores de exemplificação e os eixos se encontram rotacionados.

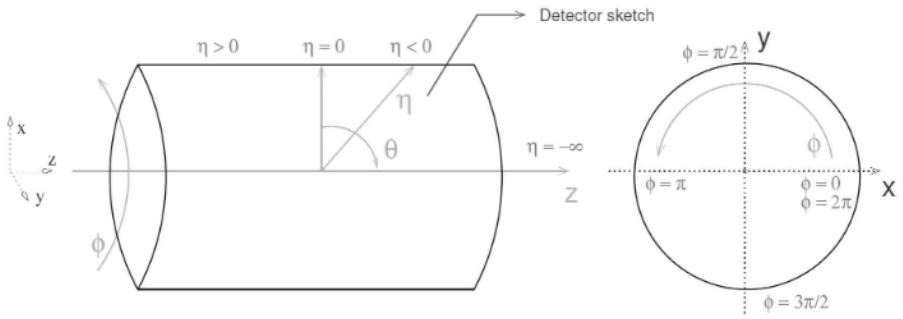
Figura 5 - Esquema do sistema de coordenadas do ATLAS.



Fonte: Rigolin e Rieznik (2005).

Na Figura 6, pode ser visto, através de um esboço do detector, como as coordenadas η e ϕ ficam localizadas ao longo do ATLAS.

Figura 6 - Coordenadas η e ϕ ao longo do detector ATLAS.



Fonte: Anjos (2006).

Com relação a sua estruturação, o ATLAS é composto por seis componentes (sistemas) principais. O Detector Central (*Inner Detector*) (ROS, 2003) é um cilindro de

7 metros de comprimento e 1,15 metros de raio que recebe o produto das colisões antes de qualquer outro elemento, e é envolto por um Solenoide (*Solenoidal Magnets*) que o fornece um campo magnético de 2 Teslas. Outras duas componentes do ATLAS são os seus calorímetros Eletromagnético e Hadrônico (*Electromagnetic* e *Hadronic Calorimeter*) (THE ATLAS COLLABORATION, 1997), dispositivos que desempenham um papel importante, pois sua resolução é melhorada com o aumento da energia e, consequentemente, a capacidade de detectar e amostrar os dados advindos das colisões também. Como o LHC trabalha com experimentos de altas energias, os calorímetros se fazem cruciais. Já o Espectrômetro de Múons (*Muon Spectrometer*) (PALESTINI, 2003) é o detector responsável por identificar o momento de múons, partículas que não são percebidas pelos detectores citados anteriormente, enquanto que o Toróide (*Toroid Magnets*) (ATLAS, 2020b) faz parte do sistema magnético do ATLAS, que ajuda a conter os rastros das partículas resultantes dos experimentos.

1.3.1 Sistema de filtragem de eventos *online* do ATLAS

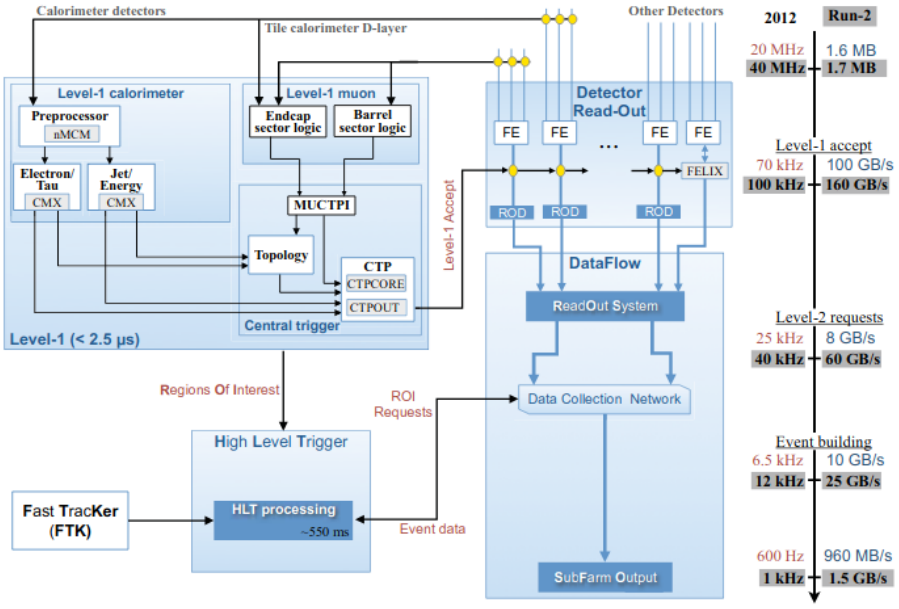
O termo *online* significa que os dados referentes às colisões são coletados e filtrados de forma concomitante aos experimentos que os geraram. Para a *Run*-2, o sistema de aquisição de dados *online* do ATLAS é dividido em duas etapas (NAKAHAMA, 2015): L1 (do inglês, *Level 1*) e HLT (do inglês, *High-Level Trigger*), que têm como objetivo o refinamento da imensa quantidade de dados que é recebida inicialmente. A Figura 7 mostra um diagrama de blocos que ilustra o sistema de filtragem e as divisões de níveis.

O L1 é um nível de *hardware* e é o primeiro do sistema de seleção de eventos do ATLAS. Este nível é composto pelo sistema de *trigger* dos calorímetros, chamado de L1Calo, do espectrômetro de múons, o L1Muon, pelos módulos de *trigger* topológicos, o L1Topo, e pelos Processadores de *Trigger* Central, o CTP (do inglês, *Central Trigger Processor*). Nesta etapa, é realizada uma triagem inicial dos dados, reduzindo a taxa das colisões, que antes era de 40 MHz, para 100 kHz.

No *Level* 1, a eletrônica utiliza dados dos calorímetros e do espectrômetro de múons para encontrar regiões de interesse (RoI, do inglês, *Region of Interest*) e é neste momento que é feita a maior parte da filtragem dos eventos no ATLAS.

O HLT é um nível de *software* onde algoritmos de alto desempenho acessam as informações contidas nos RoIs e fazem um refinamento mais detalhado dos dados obtidos. Nesta etapa, algoritmos *offline* também podem ser utilizados. Na *Run*-1, existiam o LVL2 (do inglês, *LeVeL 2*) e o EF (do inglês, *Event Filter*), que foram combinados, na *Run*-2, no HLT por questões de simplificação e melhor desempenho. Ao final deste processo, a taxa de eventos diminui de 100 kHz para 1 kHz, sendo esta a quantidade de eventos que é armazenada de forma permanente.

Figura 7 - Diagrama de blocos mostrando o sistema de aquisição de dados do ATLAS.



Fonte: Nakahama (2015).

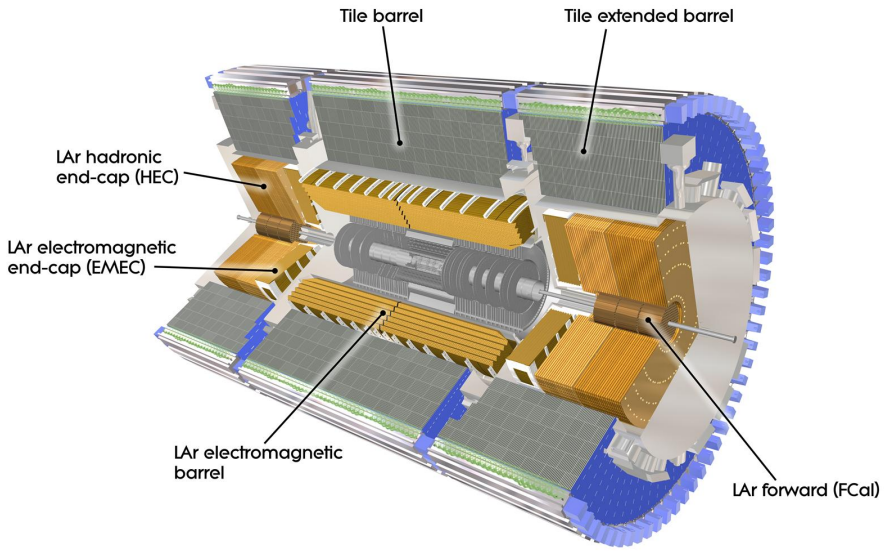
1.4 Calorimetria de Altas Energias

O estudo cada vez mais aprofundado da trajetória de partículas carregadas permite um contínuo aprendizado sobre suas naturezas e a forma como se comportam. Neste sentido, os calorímetros possuem um papel fundamental, já que são detectores capazes de complementar ou até mesmo substituir informações obtidas por outros detectores menos precisos (WIGMANS, 2000).

Em termos conceituais, o calorímetro é um instrumento que tem a capacidade de interceptar partículas primárias e que possui uma espessura que possibilita a interação de "chuvas" dessas partículas, com diminuição contínua de energia. O termo *shower*, em inglês, é utilizado neste contexto, pois, após uma colisão, as partículas se espalham de forma análoga a água saindo de um chuveiro, ou seja, tem-se uma cascata de partículas interagindo com o calorímetro neste momento. À medida que as partículas passam através das seções do calorímetro, sofrem perdas em sua energia, e essas perdas podem ser medidas pelo equipamento. Os calorímetros, no geral, são compostos por um material absorvedor e outro ativo, os quais possuem as funções de absorver a partícula e amostrar parcialmente sua energia, respectivamente (ATLAS, 2020a). A Figura 8 mostra o sistema de calorimetria do ATLAS e seus componentes.

Usualmente, os calorímetros são conhecidos por serem altamente segmentados. A obtenção das informações sobre a direção e a energia das partículas provindas das colisões

Figura 8 - Sistema de calorimetria do detector ATLAS.



Fonte: CERN (2008).

são dependentes de suas segmentações transversais, porém divisões longitudinais também podem ser projetadas, com o intuito de analisar a forma do chuveiro produzido pela partícula detectada.

É interessante pontuar algumas características dos calorímetros que fazem com que sejam largamente utilizados na área da física de partículas (FABJAN, 1985):

Sensibilidade de leitura: calorímetros são sensíveis tanto a partículas carregadas quanto a partículas neutras;

Eficiência de medição: o decaimento energético durante o decorrer da chuva de partículas é um processo estatístico, sendo o número médio de partículas secundárias proporcional a energia da partícula primária;

Tamanhos reduzidos: a profundidade dos calorímetros cresce em escala logarítmica com relação a deposição de energia das partículas, possibilitando a construção de detectores menores, se comparados a espectrômetros magnéticos, para uma dada resolução;

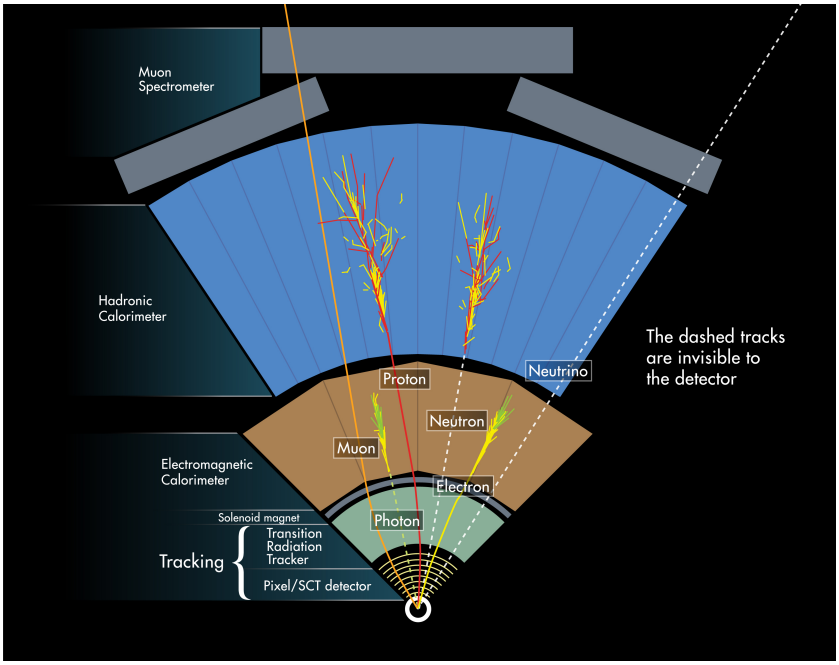
Medidas de posição e ângulo: devido ao fato de os calorímetros serem altamente segmentados, informações sobre a cascata de partículas podem ser extraídas com eficiência, permitindo medições precisas sobre posição e ângulo das partículas incidentes;

Identificação: calorímetros respondem de forma diferente a elétrons, múons e hádrons, o que pode ajudar na identificação das partículas;

Velocidade de resposta: a rápida taxa de resposta dos calorímetros permite que experimentos sejam feitos utilizando-se grandes quantidades de partículas, além da possibilidade de realização de seleção de eventos *online*.

O experimento ATLAS possui dois tipos de calorímetros: Eletromagnético e Hadrônico. Isso se deve ao fato de que diferentes tipos de partículas interagem de formas diferentes com os materiais e, por isso, sistemas específicos precisam ser implementados para estudá-las. A Figura 9 apresenta um esquema de como a energia se deposita ao longo das camadas dos calorímetros, através da ilustração de um corte transversal do detector ATLAS completo.

Figura 9 - Diagrama de deposição de energia no detector ATLAS.



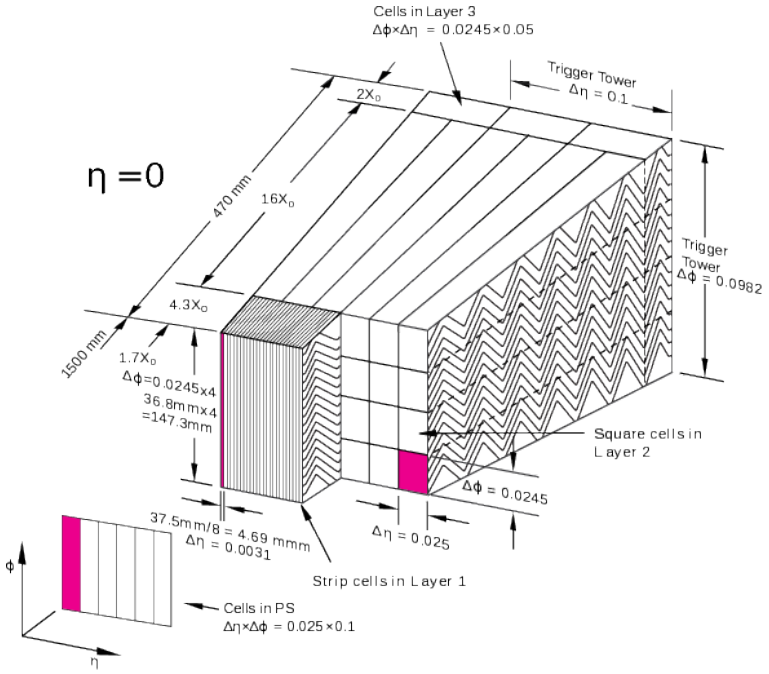
Fonte: CERN (2013b).

1.4.1 Calorímetro Eletromagnético

O Calorímetro Eletromagnético do ATLAS é também chamado de Calorímetro de Argônio Líquido (THE ATLAS COLLABORATION, 1996; ZHANG, 2011), fazendo alusão à técnica aplicada em seu funcionamento, que utiliza eletrodos de chumbo, em formato de acordeões, como material absorvedor, imersos em argônio líquido, que desempenha o papel de material ativo. Esse calorímetro tem ângulo azimutal, ϕ, completo e inclui uma faixa de pseudo-rapidez total de $|\eta| < 3,2$, sendo o barril eletromagnético

(EMB) (do inglês, *electromagnetic barrel*), com $0 < |\eta| < 1,475$, e duas tampas (EMEC) (do inglês, *end cap*), com a chamada exterior cobrindo uma faixa de $1,375 < |\eta| < 2,5$ e a interior alcançando $2,5 < |\eta| < 3,2$. Todas essas componentes, que juntas formam o Calorímetro Eletromagnético de Argônio Líquido, podem ser vistas na Figura 8, enquanto que a Figura 10 mostra sua divisão em camadas e sua granularidade, que é constante em ϕ e possui variação em η.

Figura 10 - Esboço das camadas e granularidade do Calorímetro Eletromagnético do ATLAS.



Fonte: Zhang (2011).

O Calorímetro Eletromagnético proporciona medições precisas sobre partículas eletromagnéticas, ou seja, fótons e elétrons (ver Figura 9). Isso se dá através de um pré-amostrador localizado em $|\eta| < 1,8$ e de uma camada de acordeões altamente regular até $|\eta| = 2,5$. As partículas provenientes das colisões interagem com o chumbo, que é o material absorvedor, e com o argônio líquido, produzindo partículas secundárias. Estas, por sua vez, geram cargas ionizadas, as quais criam um sinal elétrico nos eletrodos que estão imersos no argônio líquido. Esses sinais são então coletados pelas células, amplificados e amostrados pelo sistema do calorímetro. Ao todo, o calorímetro eletromagnético do ATLAS possui aproximadamente 200.000 canais de leitura.

1.4.2 Calorímetro Hadrônico

Os calorímetros hadrônicos do ATLAS (THE ATLAS COLLABORATION, 1997) possuem dois princípios de funcionamento diferentes, sendo projetados com base nos tipos de experimentos e nas partículas que se deseja detectar. Esses calorímetros têm como objetivo a detecção de hádrons, que são partículas compostas por quarks, nas quais pode-se incluir os prótons, nêutrons e píons (ver Figura 9).

Na faixa de $|\eta| < 1,6$ fica localizado o Calorímetro de Telhas, também chamado de TileCal, que utiliza placas de aço como material absorvedor e telhas cintilantes como material ativo. Esse calorímetro é composto por um barril central e um barril estendido, localizado nas extremidades do primeiro. Além disso, há também um Calorímetro de Telhas Intermediário (ITC, do inglês *Intermediate Tile Calorimeter*), que instrumenta parcialmente a divisão entre os barris e utiliza a mesma técnica de funcionamento.

Já o Calorímetro de Argônio Líquido tem como material absorvedor eletrodos de chumbo e como material ativo o argônio líquido, e é composto pelo calorímetro hadrônico de tampas (no inglês, *end-cap hadronic calorimeter*) e pelo calorímetro direto de alta densidade (no inglês, *high-density forward calorimeter*). Essas duas partes cobrem as faixas de $|\eta| < 3,2$ e $3,2 < |\eta| < 4,9$, respectivamente, sendo comportados no mesmo criostato que as tampas (*end-caps*) do calorímetro eletromagnético.

O ambiente de trabalho desta dissertação é dado especificamente pelo TileCal, que será, portanto, descrito na próxima seção.

1.5 **Calorímetro de Telhas (TileCal)**

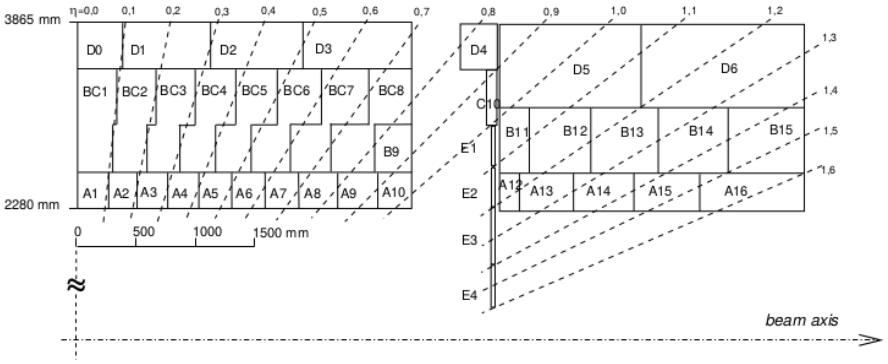
O maior calorímetro hadrônico do experimento ATLAS é o TileCal (do inglês, *Tile Calorimeter*) (AAD et al., 2010; FRANCAVILLA, 2012). Utilizando placas de aço como material absorvedor e telhas cintilantes como material ativo, o TileCal é um calorímetro de amostragem hadrônica que desempenha papel crucial na reconstrução de energia de partículas no LHC, tendo como foco principal a identificação de hádrons, jets e taus, além da medição de energia transversal que não pode ser detectada (no inglês, *missing transverse energy*), E_T^{miss}.

O TileCal engloba uma região onde a pseudo-rapidez é $|\eta| < 1,7$, dividindo-se em três partições: um barril central (LB, do inglês, *Long Barrel*) e dois barris externos (EB, do inglês, *Extended Barrels*). O barril central cobre a região central, onde $|\eta| < 1,0$, e é dividido, perpendicularmente à direção dos feixes, em duas partes independentes, que são chamadas LBA e LBC. Já os barris externos são chamados EBA e EBC e cobrem uma faixa onde $0,8 < |\eta| < 1,7$ (ver Figura 8).

Todas as quatro partes do TileCal são segmentadas em 64 módulos, com relação

ao ângulo azimutal, ϕ, produzindo uma granularidade de cerca de $\Delta\phi = 0,1$ radianos. Na direção radial, cada um dos módulos possui três camadas separadas, tendo as duas primeiras uma granularidade de $\Delta\eta = 0,1$ e a terceira, $\Delta\eta = 0,2$. A Figura 11 ilustra a segmentação de um módulo do barril central e um do barril estendido, onde a parte inferior da figura corresponde ao raio interno do cilindro e a variação é mostrada em η e também com relação a profundidade.

Figura 11 - Segmentação de um módulo do barril central (esquerda) e do barril estendido (direita), do TileCal.



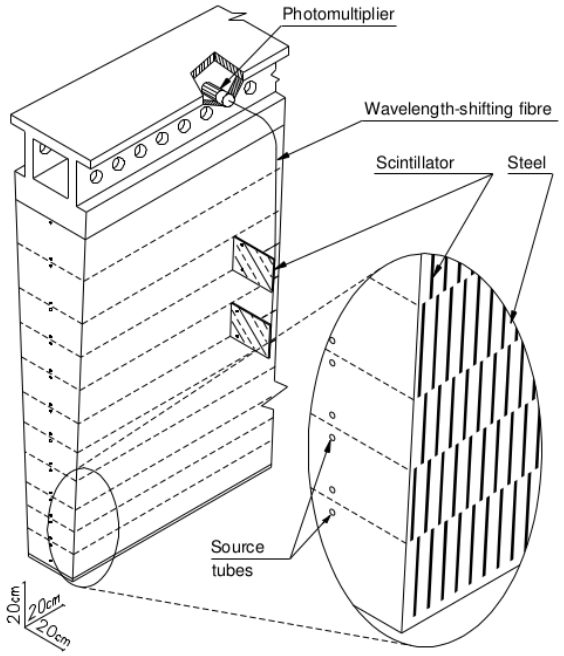
Fonte: Aad et al. (2010).

Para a coleta e arquivamento dos dados, duas fibras óticas são conectadas à cada célula do calorímetro, em lados diferentes com relação a ϕ, de forma a gerar uma redundância na leitura dos canais de saída e diminuir a perda de informação. As partículas provenientes das colisões são absorvidas pelas placas de aço e suas energias são parcialmente amostradas em forma de luz pelas telhas cintilantes. As fibras, então, coletam essa informação gerada e a carregam para diferentes tubos fotomultiplicadores (PMTs, do inglês, *photomultipliers*), os quais recebem sinais de múltiplas telhas agrupadas em células de tamanho variado. No total, o TileCal possui aproximadamente 10.000 canais, que produzem o sinal de resposta a cada colisão.

As células são alocadas em três camadas longitudinais, A, BC e D, e mais uma adicional, E, que é ligada aos módulos do barril estendido e possui apenas uma PMT conectada a cada célula. As dimensões das células das três primeiras camadas são definidas de forma a otimizar e se obter uma estrutura com granularidade $\Delta\eta \times \Delta\phi = 0,1 \times 0,1$ para as camadas A e BC, e $\Delta\eta \times \Delta\phi = 0,2 \times 0,1$ para a camada D. Ao todo, o TileCal possui 5.182 células e 9.852 canais de leitura. A Figura 12 mostra um esquema do sistema de coleta de dados do TileCal para um de seus módulos.

Ao sair das PMTs, já na eletrônica de leitura do calorímetro, os sinais são conformados. Em seguida, após serem gravados, os sinais são classificados como de alto ganho (HG, do inglês, *high gain*) ou baixo ganho (LG, do inglês, *low gain*). A partir daí, são

Figura 12 - Sistema de instrumentação para um módulo do TileCal.



Fonte: Aad et al. (2010).

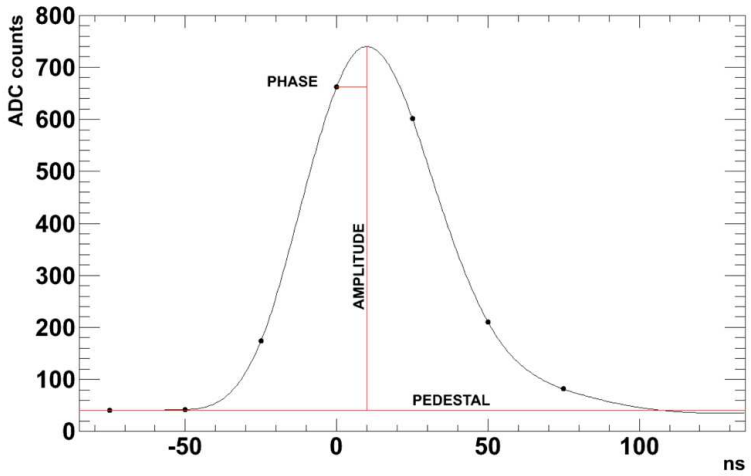
então amostrados a 40 MHz, que é a frequência utilizada no LHC, e o tempo de digitalização é ajustado de forma que a amostra central fique o mais próximo possível do pico do sinal e que todo o pulso seja coberto. Para cada colisão, a eletrônica de leitura produz um sinal de aproximadamente 150 ns e, portanto, sete amostras digitais são suficientes para representar o pulso analógico adquirido.

Os sinais de entrada de várias células dentro de uma granularidade de $\Delta\eta \times \Delta\phi = 0,1 \times 0,1$ são manipulados de forma a compor uma soma analógica. O resultado dessa soma é chamado Torre de Trigger e é enviado, através de cabos longos, até o *Level-1* (LVL1), que é o primeiro nível de seleção de eventos do ATLAS, passando em seguida pelos outros níveis de seleção.

O pulso digital contém sete amostras, número definido com base nos requisitos de largura de banda, as quais são enviadas através de fibras óticas para a eletrônica da parte interna do sistema (no inglês, *back-end electronics*), para os chamados RODs (do inglês, *Read Out Drivers*), que fica em um local fora da área do experimento, em uma central de controle. Nesta etapa, a energia da partícula detectada por uma determinada célula, de eventos que tenham sido aceitos pelo LVL1, pode ser estimada. Essa energia é proporcional à amplitude do sinal. A Figura 13 ilustra o pulso de referência do TileCal, onde o pedestal é dado como a linha de base do sinal, a fase é a distância entre a quarta amostra e o pico do sinal e a amplitude é a altura do sinal a partir do pedestal, ou seja,

da linha de base.

Figura 13 - Pulso do TileCal mostrando os parâmetros do pulso gerado pela eletrônica do TileCal: pedestal, fase e amplitude.



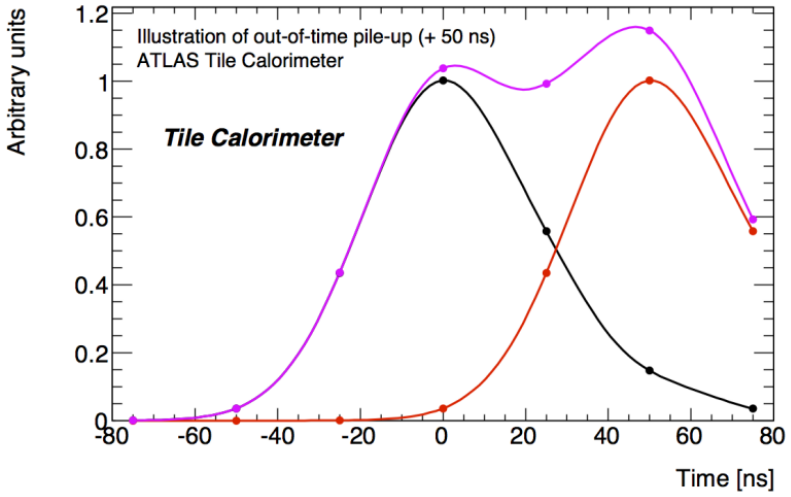
Fonte: Peralva (2013).

Para que a reconstrução do sinal original, obtido a partir de uma dada colisão e que descreve a energia da partícula referente, seja possível, é necessário que a estimação da amplitude, fase e pedestal seja feita de forma eficiente a partir das 7 amostras digitais recebidas por cada canal de leitura. Os algoritmos tipicamente utilizados para tal são apresentados no Capítulo 2.

Em condições de baixa luminosidade, e devido à alta segmentação do calorímetro, o efeito de empilhamento de sinais é pouco provável no TileCal, e a única fonte de ruído é proveniente da eletrônica. Esta componente de ruído é tipicamente caracterizada por uma distribuição de probabilidade Gaussiana, e métodos baseados na minimização da variância operam próximos a seus pontos ótimos. Entretanto, em condições de alta luminosidade, o fenômeno de empilhamento de sinais pode ser observado em alguns canais do sistema de calorimetria do ATLAS, ou seja, deformando o sinal recebido e degradando a eficiência da estimação da energia, conforme mostrado na Figura 14, que ilustra o problema no TileCal (BARBOSA et al., 2017).

Nota-se, da Figura 14, que o empilhamento ocorre quando um ou mais sinais de colisões adjacentes entram na janela de leitura do sinal de interesse que está sendo amostrado naquele momento. Isso se torna possível porque as colisões no LHC acontecem a cada 25 ns e a janela de leitura do TileCal possui um total de 150 ns. Este é o chamado *out-of-time pile-up*, ou, na tradução direta, empilhamento fora do tempo. Há, ainda, o *in-time pile-up*, ou, na tradução direta, empilhamento no tempo, que ocorre quando dois

Figura 14 - Efeito do empilhamento de sinais na eletrônica do TileCal. O pulso de interesse está centrado na janela de leitura, mas antes que fosse totalmente gerado, um segundo sinal é adquirido no instante +50 ns, resultando no sinal deformado.



Fonte: Seixas (2015).

sinais são gerados em uma mesma colisão e observados em um mesmo canal de leitura. O nível de empilhamento de um canal de leitura está associado à sua posição espacial no detector: quanto mais próximo do feixe, maior a probabilidade de se observar sinais empilhados.

Essa componente de empilhamento pode ser tratada como ruído adicional, o que modifica o ruído presente nos sinais, que era proveniente apenas da eletrônica de leitura do calorímetro e possuía características Gaussianas.

1.5.1 Sistema de Calibração do TileCal

O TileCal possui um sistema de calibração altamente eficiente, possibilitando maior estabilidade e confiança dos dados coletados (MARJANOVIĆ, 2019). O sinal gerado após as colisões precisa ser bem calibrado e minuciosamente monitorado e, para isso, são utilizados subsistemas, os quais compõem o sistema de calibração total, sendo os principais o Sistema de Cesium, de Laser, de Injeção de Carga e de Tempo.

O Sistema de Cesium (do inglês, *Cesium Calibration*) é responsável pela calibração dos componentes ópticos do TileCal (telhas cintilantes, fibras ópticas e eletrônicos) e das PMTs, verificando a qualidade de todo o sistema de leitura. Esse procedimento é realizado

através da utilização de uma fonte γ radioativa, que passa por todo o calorímetro emitindo raios γ de energia bem definida. Além disso, esse sistema também é encarregado de fazer a equalização da resposta de todos os canais de leitura, além de monitorar, no tempo, a escala eletromagnética da célula.

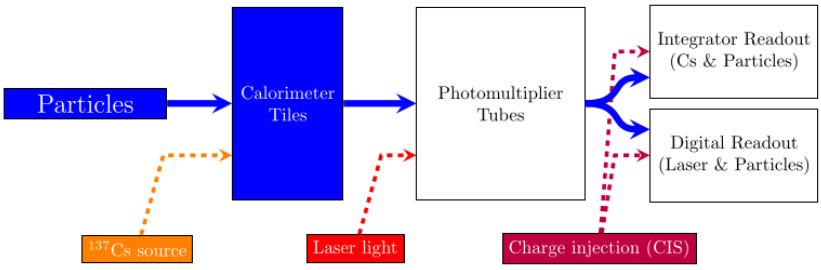
Já o Sistema de Laser (do inglês, *Laser Calibration*), por sua vez, tem a função de mensurar o desvio observado na resposta das PMTs, tendo como base a última medição feita pelo sistema de cesium. A variação de ganho média de todos os canais é feita célula por célula, a partir de uma quantidade controlada de luz que é enviada às PMTs através das fibras ópticas. As medidas feitas nessa etapa são responsáveis pela calibração de fase do sinal obtido.

O Sistema de Injeção de Carga (ou CIS, do inglês, *Charge Injection System*) realiza a medição do fator de conversão de ADC para pC, além de corrigir não-linearidades. Esse processo é feito introduzindo-se um sinal de carga conhecida e medindo a resposta eletrônica correspondente. Toda a faixa de valores em ADC está compreendida nesta etapa, ou seja, de 0 a 800 pC.

A precisão do tempo de medida também é crucial para obter-se descrições fidedignas dos eventos. Sendo assim, o Sistema de Tempo (do inglês, *Time calibration*), mais um subsistema que compõe o sistema de calibração do TileCal, define a fase do sinal de forma que a partícula que esteja chegando, a partir de uma colisão, produza um sinal com tempo igual a zero. Essa fase da calibração é monitorada durante a aquisição de dados usando o sistema de laser.

A Figura 15 apresenta uma ilustração de fluxo do sistema de calibração do TileCal. A junção desses diferentes subsistemas de calibração fornece, ao final, um pulso altamente preciso, tendo sua amplitude correspondente à energia da partícula que o gerou. A precisão do sinal recebido é essencial para a recuperação da energia e, consequentemente, para a identificação da natureza da partícula.

Figura 15 - Diagrama de fluxo do sistema de calibração do TileCal.



Fonte: Marjanović (2019).

2 ALGORITMOS PARA ESTIMAÇÃO DA ENERGIA

Neste capítulo, serão apresentados três dos métodos utilizados nas simulações computacionais desta dissertação. Nas Seções 2.1, 2.2 e 2.3.1 são detalhados métodos lineares já difundidos, sendo o primeiro atualmente utilizado no LHC. Já o Capítulo 3 aborda um método não-linear, tendo como base uma distribuição Lognormal, que é a alternativa proposta neste estudo para lidar com o problema de empilhamento de sinais no LHC. Este método, no entanto, pode ser aplicado em calorímetros no geral, sendo o TileCal utilizado apenas como base para as análises de eficiência.

2.1 Filtro Ótimo

Em um aspecto geral, filtros são utilizados como forma de eliminar impurezas, separar itens de interesse daqueles que não são necessários e acabam contaminando os desejados. No contexto de processamento de sinais em calorimetria, o raciocínio é o mesmo. Filtro Ótimo (OF, do inglês, *Optimal Filter*) (CLELAND; STERN, 1994) é um algoritmo baseado na minimização da variância que visa determinar os parâmetros de interesse através de uma combinação linear de amostras do sinal, diminuindo os efeitos de ruído. Para a aplicação deste método, é necessário o conhecimento da forma do pulso (CLEMENT; KLIMEK, 2011), informação esta que é bem estabelecida para o TileCal (ver Figura 13). Nesta abordagem, o sinal completo é tratado como sinal de interesse acrescido de um ruído Gaussiano, e, a partir deste conjunto, o sinal de interesse é filtrado e sua amplitude pode ser determinada.

Atualmente, no TileCal, é utilizada uma versão de Filtro Ótimo chamada OF2 (FULLANA et al., 2006). Esse algoritmo foi inicialmente desenvolvido para calorímetros de ionização líquida (CLELAND; STERN, 1994), sendo, em seguida, adaptado para aplicação no TileCal.

Com o OF2, na fase de reconstrução de energia nos RODs, as amostras do sinal recebido podem ser descritas pela Equação (4).

$$r_k = ped + As(t_k + \tau) + n_k, \tag{4}$$

onde r_k representa a amostra $k = 0, 1, ..., N-1$, sendo N o número total de amostras; o parâmetro *ped* é a variável correspondente ao pedestal do sinal, uma constante que é adicionada ao sinal analógico antes da digitalização; A é a amplitude verdadeira; s é o valor do pulso de referência do sinal, normalizado e sem ruído, tomado no tempo t_k; τ é definido como a fase do sinal e n_k é a componente de ruído presente.

Desenvolvendo o vetor do pulso de referência, $\mathbf{s}$, em uma série de Taylor de primeira ordem, obtém-se o formato final utilizado para descrever o sinal recebido, que é mostrado na Equação (5).

$$r_k = ped + As_k + A\tau\dot{s}_k + n_k, \tag{5}$$

onde $\dot{\mathbf{s}}$ é a derivada de $\mathbf{s}$, ou sua aproximação linear. Quanto maior a precisão na reconstrução da energia, mais próxima de zero estará a fase τ do sinal.

A amplitude pode, então, ser estimada através do somatório apresentado na Equação (6).

$$\hat{A}_{OF} = \sum_{k=0}^{N-1} w_k r_k, \tag{6}$$

na qual w_k é o coeficiente do filtro no instante k.

Assim, substituindo a Equação (5) na Equação (6), tem-se como resultado:

$$\hat{A}_{OF} = \sum_{k=0}^{N-1} \left(w_k ped + Aw_k s_k + Aw_k \tau \dot{s}_k + w_k n_k\right). \tag{7}$$

No entanto, a média do ruído será nula e, para que a amplitude calculada seja igual a amplitude verdadeira, ou seja, $\hat{A}_{OF} = A$, e o estimador seja independente da fase e do pedestal, as restrições apresentadas na Equação (8) são aplicadas.

$$\begin{aligned}
&\sum_{k=0}^{N-1} w_k s_k = 1;\\
&\sum_{k=0}^{N-1} w_k \dot{s}_k = 0;\\
&\sum_{k=0}^{N-1} w_k = 0.
\end{aligned} \tag{8}$$

A variância do estimador é definida como:

$$Var(A) = \sum_{k=0}^{N-1}\sum_{j=0}^{N-1} w_k w_j C_{kj}, \tag{9}$$

onde $\mathbf{C}$ é a matriz de covariância do ruído, dada pela Equação (10).

$$\mathbf{C} = \begin{bmatrix} E(\bar{n}_{11}) & E(\bar{n}_{12}) & \cdots & E(\bar{n}_{1N}) \\ E(\bar{n}_{21}) & E(\bar{n}_{22}) & \cdots & E(\bar{n}_{2N}) \\ \vdots & \vdots & \ddots & \vdots \\ E(\bar{n}_{N1}) & E(\bar{n}_{N2}) & \cdots & E(\bar{n}_{NN}) \end{bmatrix}, \tag{10}$$

sendo $\bar{n}_{kj} = [n_k - E(n_k)][n_j - E(n_j)]$, n_k o ruido no instante k e N o número total de amostras.

Os coeficientes do filtro são calculados através do uso de multiplicadores de Lagrange, minimizando a variância dada pela Equação (9) e satisfazendo as restrições da Equação (8). A Equação (11) mostra como os multiplicadores, representados por λ, κ e ϵ, se relacionam aos pesos que serão determinados.

$$I_A = \sum_{k=0}^{N-1}\sum_{j=0}^{N-1} w_k w_j C_{kj} - \lambda\left(\sum_{k=0}^{N-1} w_k s_k\right) - \kappa\left(\sum_{k=0}^{N-1} w_k \dot{s}_k\right) - \epsilon\left(\sum_{k=0}^{N-1} w_k\right). \tag{11}$$

Para o processo de minimização, a Equação (11) é derivada com relação aos coeficientes e igualada a zero. O resultado deste procedimento é apresentado na Equação (12).

$$\frac{\partial I_A}{\partial w_k} = 2\sum_{j=0}^{N-1} C_{kj} w_j - \lambda s_k - \kappa \dot{s} - \epsilon = 0. \tag{12}$$

O sistema de equações lineares composto pela Equação (12) e pelo conjunto de restrições da Equação (8) é representado em sua forma matricial pela Equação (13), onde o lado direito da igualdade relaciona-se com os valores impostos pelas restrições.

$$2\begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1N} & -s_1 & -\dot{s}_1 & -1 \\ C_{21} & C_{22} & \cdots & C_{2N} & -s_2 & -\dot{s}_2 & -1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ C_{N1} & C_{N2} & \cdots & C_{NN} & -s_N & -\dot{s}_N & -1 \\ s_1 & s_2 & \cdots & s_N & 0 & 0 & 0 \\ \dot{s}_1 & \dot{s}_2 & \cdots & \dot{s}_N & 0 & 0 & 0 \\ 1 & 1 & \cdots & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \\ \lambda \\ \kappa \\ \epsilon \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \tag{13}$$

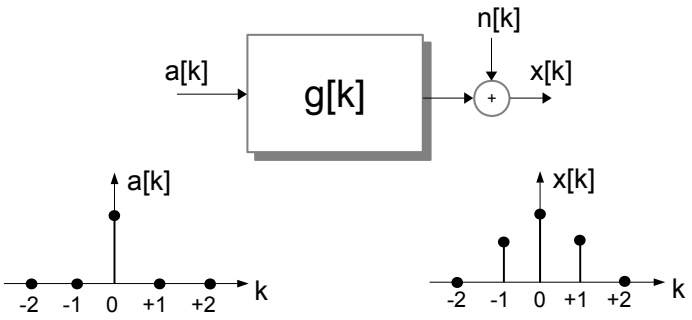
A partir da resolução deste sistema, os pesos do OF2 são calculados e a amplitude do sinal pode ser estimada. Como a correlação entre as amostras do sinal, encontrada na eletrônica do TileCal, é baixa e suas distribuições são aproximadas por Gaussianas, a matriz de covariância do ruído $\mathbf{C}$ pode ser considerada como uma matriz identidade. O OF2, no entanto, tende a sofrer uma diminuição da sua eficiência quando o LHC trabalha

com altas taxas de luminosidade e o problema de empilhamento de sinais surge, o que faz com que o sinal perca sua característica Gaussiana.

2.2 COF

O COF (do inglês, *Constrained Optimal Filter*) (ANDRADE FILHO et al., 2015) é um método linear que trabalha com uma técnica baseada na desconvolução aplicada aos sinais digitais recebidos. Nesta abordagem, o processo de produção do sinal no calorímetro é interpretado como um sistema linear e o intuito é recuperar, através do sinal imerso em ruído, o sinal original, a partir do qual a amplitude surge como parâmetro de interesse principal. A Figura 16 ilustra esse processo, onde x é o sinal recebido, que aqui será chamado de r.

Figura 16 - Sistema de geração do sinal de resposta de um canal do calorímetro.



Fonte: ANDRADE FILHO et al. (2015)

Para a aplicação do processo de desconvolução de forma completa, uma grande quantidade de dados, fornecidos de forma ininterrupta, é necessária. Para os estudos desta dissertação, no entanto, são utilizadas janelas de aquisição de sinais curtas. Assim, o COF apresentado aqui é um método alternativo, que pode ser usado no cenário de interesse descrito, onde os sinais não são fornecidos em grandes quantidades de maneira contínua.

Para o COF, o sinal digital de saída do calorímetro, r_k, pode ser modelado como um sinal conformado, cobrindo vários cruzamentos de feixes (colisões), somado ao ruído eletrônico do sistema. A Equação (14) demonstra o procedimento matematicamente, onde k é a amostra de tempo de um sistema de tempo discreto linear invariante no tempo, ou LTI (do inglês, *Linear time-invariant*), e s_k é a resposta de impulso desse sistema. A energia depositada na célula do calorímetro em uma dada colisão é modelada por um sinal impulso e representada pelas amostras a_k em seus respectivos tempos de

amostragem, enquanto que n_k refere-se ao ruído eletrônico do sistema, que é descrito por uma distribuição de probabilidade Gaussiana.

$$r_k = \sum_i (s_i a_{k-i}) + n_k. \tag{14}$$

Assim, a resposta do calorímetro pode ser dada por uma convolução entre os sinais de entrada e a resposta do sistema LTI, que corresponde ao pulso de referência do calorímetro.

Para problemas onde os dados são fornecidos de forma ininterrupta, o processo de desconvolução pode ser realizado com a utilização de uma versão causal de um filtro inverso com resposta de frequência $1/S(z)$, onde $S(z)$ é a transformada Z da resposta de impulso do sistema, s_k. Assim, se o pulso de referência se estende por D cruzamentos de feixes e são armazenadas N amostras do sinal de saída, a Equação (14) pode ser escrita na forma vetorial mostrada na Equação (15).

$$\mathbf{r}_{N\times 1} = \mathbf{S}_{N\times P}\mathbf{a}_{P\times 1} + \mathbf{n}_{N\times 1}, \tag{15}$$

onde $P = D + N - 1$, $\mathbf{r}$, $\mathbf{a}$ e $\mathbf{n}$ são as mesmas variáveis contidas na Equação (14) e $\mathbf{S}_{N\times P}$ é uma matriz com P versões modificadas (defasadas) do pulso de referência do calorímetro.

Como $P > N$, devido ao calorímetro usado como base para esta dissertação, no qual apenas uma janela contendo o sinal de interesse e alguns cruzamento adjacentes a esta podem ser armazenados, o sistema resultante da Equação (15) não possui solução única.

Visando implementar o COF no ambiente do presente estudo, considera-se, para a resolução do sistema, apenas os $p \leq N$ elementos principais no vetor do sinal de entrada, $\mathbf{a}_p$, fazendo com que as outras componentes, referentes ao ruído de empilhamento proveniente dos cruzamentos adjacentes, sejam absorvidas pelo vetor $\mathbf{n}_p$. Assim, $\mathbf{S}_{N\times P}$ se torna uma matriz $\mathbf{S}_{N\times p}$, contendo p versões modificadas dos sinais de referência do calorímetro, e a Equação (15) pode ser descrita na forma da Equação (16).

$$\mathbf{r} = \mathbf{S}_p\mathbf{a}_p + \mathbf{n}_p, \tag{16}$$

onde $\mathbf{a}_p$ é o vetor das amplitudes que devem ser estimadas pelo método. Assim, a amplitude estimada pode ser escrita como na Equação (17), onde $\mathbf{G}_p$ é uma matriz $N \times p$ que permite a estimação das componentes do vetor.

$$\hat{\mathbf{a}}_p = \mathbf{G}_p^T\mathbf{r}. \tag{17}$$

Para estimadores não-tendenciosos, é necessário que o valor esperado da variável

estimada seja igual ao valor da variável verdadeira, ou seja,

$$E\{\hat{\mathbf{a}}_p\} = \mathbf{G}_p^T E\{\mathbf{r}\} = \mathbf{a}_p. \tag{18}$$

Subtraindo-se o pedestal do ruído e fazendo com que o vetor resultante tenha média zero, a igualdade apresentada na Equação (19) se faz verdadeira.

$$E\{\mathbf{r}\} = E\{\mathbf{S}_p\mathbf{a}_p + \mathbf{n}_p\} = \mathbf{S}_p\mathbf{a}_p. \tag{19}$$

Combinando as Equações 18 e 19, chega-se a:

$$\begin{aligned}
&\mathbf{G}_p^T\mathbf{S}_p\mathbf{a}_p = \mathbf{a}_p \\
&\mathbf{G}_p^T\mathbf{S}_p\mathbf{a}_p\mathbf{a}_p^{-1} = \mathbf{a}_p\mathbf{a}_p^{-1} \\
&\mathbf{G}_p^T\mathbf{S}_p = \mathbf{I}_p.
\end{aligned} \tag{20}$$

Assim, é necessário minimizar a variância dos estimadores contidos em $\mathbf{G}_p$ sujeito às restrições impostas na Equação (20). Esse procedimento pode ser realizado utilizando-se multiplicadores de Lagrange (KAY, 1993) e o resultado é apresentado na Equação (21).

$$\mathbf{G}_p = \mathbf{C}_p^{-1}\mathbf{S}_p\left(\mathbf{S}_p^T\mathbf{C}_p^{-1}\mathbf{S}_p\right)^{-1}, \tag{21}$$

onde $\mathbf{C}_p$ é a matriz de covariância do ruído, de dimensão $N \times N$, responsável por absorver as estatísticas de segunda ordem dos sinais que não fizerem parte $\mathbf{S}_p$.

Para o Calorímetro de Telhas, utilizado como base para a análise dos métodos estudados nesta dissertação, o número de colisões é sempre igual ao número de componentes na janela de leitura, ou seja, $p = N$. Como a janela de leitura do sinal do TileCal é composta por 7 amostras digitais, $N = 7$ e, nesta situação, a matriz $\mathbf{S}$, de dimensão $N \times N$, é dada pela Equação (22) (GONÇALVES et al., 2020).

$$\mathbf{S} = \begin{bmatrix}
s_3 & s_4 & s_5 & s_6 & 0 & 0 & 0 \\
s_2 & s_3 & s_4 & s_5 & s_6 & 0 & 0 \\
s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & 0 \\
s_0 & s_1 & s_2 & s_3 & s_4 & s_5 & s_6 \\
0 & s_0 & s_1 & s_2 & s_3 & s_4 & s_5 \\
0 & 0 & s_0 & s_1 & s_2 & s_3 & s_4 \\
0 & 0 & 0 & s_0 & s_1 & s_2 & s_3
\end{bmatrix} \tag{22}$$

Substituindo a Equação (21) na Equação (17) e considerando as dimensões defini-

das acima, obtém-se:

$$\hat{\mathbf{a}} = \left(\mathbf{S}^T\mathbf{C}^{-1}\mathbf{S}\right)^{-1}\mathbf{S}^T\mathbf{C}^{-1}\mathbf{r}, \tag{23}$$

onde $\hat{\mathbf{a}}$ e $\mathbf{r}$ são vetores de dimensões $N \times 1$. Como, neste caso, o sinal se encontra completamente dentro da janela de aquisição, ou seja, a quantidade de cruzamentos de feixes é menor ou igual ao número de amostras discretas ($D \leq N$), o ruído de empilhamento contido na matriz de covariância pode ser desconsiderado. Isso faz com que apenas o ruído eletrônico, Gaussiano e descorrelacionado, permaneça, simplificando a Equação (23) para:

$$\hat{\mathbf{a}} = \mathbf{S}^{-1}\mathbf{r}. \tag{24}$$

O COF é um método desenvolvido de forma a lidar apenas com o empilhamento *out-of-time*. Por causa da sua estruturação, deve ser capaz de obter estimações precisas da amplitude, independentemente da luminosidade com a qual o experimento esteja trabalhando, fornecendo resultados confiáveis mesmo sob altas taxas de empilhamento de sinais.

2.3 Estimadores de Máxima Verossimilhança

O MLE (do inglês, *Maximum Likelihood Estimation*) (KAY, 1993) é um método baseado na maximização da verossimilhança amplamente utilizado na construção de estimadores práticos. É uma técnica que pode ser aplicada a problemas complexos e quando há grandes quantidades de dados.

Nesta abordagem, a função de verossimilhança (no inglês, *likelihood function*), definida na Equação (25), deve ser maximizada, utilizando-se algum método matemático para tal.

$$L(\mathbf{X}|\theta) = P(\theta|\mathbf{x} = \mathbf{X}), \tag{25}$$

onde L é a função de máxima verossimilhança e θ é o parâmetro que se deseja determinar quando $\mathbf{x}$ assume um determinado valor $\mathbf{X}$. Para a maximização, $\mathbf{X}$ deve representar os valores máximos do vetor de variáveis. Nesta dissertação, a Função Densidade de Probabilidade (PDF, do inglês, *Probability Density Function*) foi escolhida como função de verossimilhança.

O procedimento utilizado para a maximização é mostrado na Equação (26) e consiste na obtenção da derivada parcial da PDF, com relação ao parâmetro que se deseja

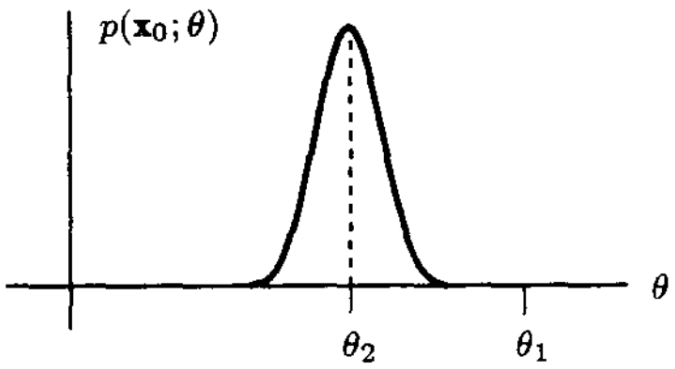
estimar, igualando-a a zero em seguida.

$$\frac{\partial p(\mathbf{x}|\theta)}{\partial \theta} = 0. \tag{26}$$

Aqui, p é a função de probabilidade, $\mathbf{x}$ é o vetor que corresponde ao processo que está sendo modelado e θ é o parâmetro a ser determinado. O valor de θ que fizer a Equação (26) chegar mais próximo de zero é escolhido.

A explicação para o MLE reside na percepção de que $p(\mathbf{x}|\theta)d\mathbf{x}$ fornece, para cada θ, a probabilidade de que $\mathbf{x}$ esteja próximo o suficiente de um determinado valor dada uma estimativa para θ. A Figura 17 apresenta uma visualização para este conceito, onde tem-se o gráfico de uma função probabilidade de $\mathbf{x} = \mathbf{x_0}$. Sabendo-se que $\mathbf{x}$ assume de fato os valores de $\mathbf{x_0}$, seria inadequado a escolha de $\theta = \theta_1$, já que para este valor de θ a probabilidade de $\mathbf{x} = \mathbf{x_0}$ é ínfima. Por outro lado, $\theta = \theta_2$ se torna uma boa estimativa, pois neste caso há uma grande chance de $\mathbf{x_0}$ ter sido o resultado observado para $\mathbf{x}$.

Figura 17 - Visualização para o conceito do método MLE.



Fonte: Kay (1993).

Nesta dissertação, o MLE é aplicado às amostras de ruído do sinal do TileCal, representadas por $\mathbf{n}$, o qual foi apresentado na Equação (5). Ao passar pela eletrônica de leitura do calorímetro, o sinal é validado por um complexo sistema de calibração, fazendo com que o pedestal seja subtraído no momento em que as amostras digitais são recebidas e com que a fase do sinal possa ser considerada nula ($\tau = 0$). Assim, o sinal recebido assume uma forma mais simples, que é mostrada na Equação (27) com o ruído em evidência.

$$n_k = r_k - As_k, \qquad k = 0, 1, ..., N-1. \tag{27}$$

Neste caso, $N = 7$ é o número total de amostras do sinal do TileCal e o parâmetro a ser determinado é a amplitude A do sinal de interesse

No geral, o MLE se apresenta como um método imparcial e é desenvolvido em conjunto com uma PDF Gaussiana. No entanto, a utilização de outras distribuições pode ser aplicada quando se mostrar mais adequado. Quando uma fórmula fechada para a estimação do parâmetro não for possível de ser encontrada, um processo iterativo deve ser aplicado, porém não existe garantia de convergência no caso de ferramentas que dependam dessa condição.

2.3.1 MLE Gaussiano

Nesta abordagem, será utilizada a PDF de uma distribuição Gaussiana multivariada, ou Normal multivariada, que é apresentada, em sua forma geral, na Equação (28) (ANDERSON, 2003).

$$p(\mathbf{x}) = \frac{1}{|\mathbf{C}|^{\frac{1}{2}}(2\pi)^{\frac{N}{2}}} \exp\left[-\frac{(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}\right], \tag{28}$$

onde $\mathbf{x}$ é o vetor de variáveis aleatórias do processo, N é o número de amostras e $\mathbf{C}$ e $\boldsymbol{\mu}$ são os parâmetros da distribuição, definidos, respectivamente, como a matriz de covariância, de dimensão $N \times N$, e o vetor de médias, de dimensão $N \times 1$.

Na presente dissertação, o processo a ser modelado é o ruído, dado como $\mathbf{n} = \mathbf{r} - A\mathbf{s}$, onde está contida a variável a ser determinada, A. Neste método, o ruído é tratado como Gaussiano e o vetor de médias é nulo. Assim, a PDF Gaussiana pode ser escrita especificamente para este caso como:

$$p(\mathbf{n}) = \frac{1}{|\mathbf{C}|^{\frac{1}{2}}(2\pi)^{\frac{N}{2}}} \exp\left[-\frac{(\mathbf{r}-A\mathbf{s})^T\mathbf{C}^{-1}(\mathbf{r}-A\mathbf{s})}{2}\right]. \tag{29}$$

Como o MLE consiste na maximização da função densidade de probabilidade e isso é feito através de sua derivada parcial, a Equação (29) será derivada com relação a A, que é o parâmetro para o qual deseja-se encontrar um valor que a maximize. Antes, no entanto, com o intuito de facilitar os cálculos, a função logaritmo natural pode ser aplicada. Este procedimento não altera o resultado final, já que a PDF Gaussiana aqui

utilizada se trata de uma função monolítica. A Equação (30) mostra este procedimento.

$$\begin{aligned}\ln\left[p(\mathbf{n})\right] &= \ln\left\{\frac{1}{|\mathbf{C}|^{\frac{1}{2}}(2\pi)^{\frac{N}{2}}}\exp\left[-\frac{(\mathbf{r}-A\mathbf{s})^T\mathbf{C}^{-1}(\mathbf{r}-A\mathbf{s})}{2}\right]\right\}\\ &= \ln(1) - \ln\left[|\mathbf{C}|^{\frac{1}{2}}(2\pi)^{\frac{N}{2}}\right] + \ln\left\{\exp\left[-\frac{(\mathbf{r}-A\mathbf{s})^T\mathbf{C}^{-1}(\mathbf{r}-A\mathbf{s})}{2}\right]\right\}\\ &= 0 - \left\{\ln\left(|\mathbf{C}|^{\frac{1}{2}}\right) + \ln\left[(2\pi)^{\frac{N}{2}}\right]\right\} - \frac{(\mathbf{r}-A\mathbf{s})^T\mathbf{C}^{-1}(\mathbf{r}-A\mathbf{s})}{2}\\ &= -\frac{1}{2}\left[\ln\left(|\mathbf{C}|\right) + N\ln(2\pi) + (\mathbf{r}-A\mathbf{s})^T\mathbf{C}^{-1}(\mathbf{r}-A\mathbf{s})\right]. \end{aligned} \tag{30}$$

Como o determinante da matriz $\mathbf{C}$, o número de amostras N e 2π são fatores constantes, não dependentes do parâmetro A, apenas o último termo desta expressão terá derivada parcial diferente de zero. A Equação (31), portanto, apresenta o processo de derivação aplicado.

$$\begin{aligned}\frac{\partial}{\partial A}\left\{\ln\left[p(\mathbf{n})\right]\right\} &= \frac{\partial}{\partial A}\left[-\frac{1}{2}(\mathbf{r}-A\mathbf{s})^T\mathbf{C}^{-1}(\mathbf{r}-A\mathbf{s})\right]\\ &= -\frac{1}{2}\frac{\partial}{\partial A}\left(\mathbf{r}^T\mathbf{C}^{-1}\mathbf{r} - 2A\mathbf{r}^T\mathbf{C}^{-1}\mathbf{s} + A^2\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}\right). \end{aligned} \tag{31}$$

Como $\mathbf{C}$ é uma matriz simétrica, a simplificação mostrada na Equação (32) pode ser aplicada.

$$\frac{\partial}{\partial A}\left\{\ln\left[p(\mathbf{n})\right]\right\} = \mathbf{r}^T\mathbf{C}^{-1}\mathbf{s} - A\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}. \tag{32}$$

Finalmente, igualando a derivada a zero, obtém-se uma expressão fechada para a amplitude A:

$$A = \frac{\mathbf{r}^T\mathbf{C}^{-1}\mathbf{s}}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}. \tag{33}$$

Este método também é considerado como um Filtro Ótimo (OF), assim como aquele apresentado na Seção 2.1. Contudo, neste caso, a quantidade de restrições aplicadas ao procedimento é menor, o que faz com que a estimação da amplitude seja mais eficiente. Para escrever a Equação (33) na forma de filtro, primeiramente define-se o coeficiente do filtro:

$$w_{OF1} = \frac{\mathbf{C}^{-1}\mathbf{s}}{\mathbf{s}^T\mathbf{C}^{-1}\mathbf{s}}, \tag{34}$$

e, em seguida, a amplitude pode ser expressada como na Equação (35).

$$A = \mathbf{r}^T w_{OF1}. \tag{35}$$

Apesar das vantagens que este método traz para a estimação da amplitude, nem sempre o ruído pode ser modelado adequadamente por uma distribuição Gaussiana, como no problema estudado nesta dissertação. Assim, uma alternativa não-linear, que descreva o ruído de forma mais apropriada, pode ser proposta.

3 MÉTODO PROPOSTO

Com o aumento da luminosidade no LHC, o ruído, no TileCal, perde as características Gaussianas. Isso se deve ao fato de que o fenômeno de empilhamento de sinais (do inglês, *pile-up*) acrescenta ao ruído eletrônico uma componente não-Gaussiana, dada por um somatório de exponenciais (KHANDAI et al., 2013), que pode ser tratada como ruído adicional.

Buscando uma alternativa que seja capaz de modelar o ruído de empilhamento de forma mais adequada, faz-se uma análise dos sinais amostrados pelo calorímetro que serviu como base para os estudos desta dissertação. No TileCal, como visto na Seção 1.5, o sinal é caracterizado por um pulso unipolar, o que tem como consequência o fato de que sinais sobrepostos somam-se entre si. Desta forma, o ruído de empilhamento resulta em um somatório de várias exponenciais, o que leva a um tipo específico de distribuição Gama, a distribuição Erlang (FORBES et al., 2011). Apesar do ruído de empilhamento e o eletrônico estarem combinados juntamente, a energia do primeiro é geralmente muito superior a do segundo, e, portanto, este pode ser desconsiderado para fins de simplificação e testes (BARBOSA et al., 2017).

No entanto, cada sinal gerado pela eletrônica de leitura do TileCal possui um total de 7 amostras, o que leva as simulações ao campo da estatística multivariável. Como a distribuição Gama multivariada ainda não é bem definida matematicamente, uma alternativa frequentemente utilizada é sua aproximação por uma distribuição Lognormal. Essas duas funções possuem características semelhantes (ALZAID; SULTAN, 2009) e, para as aplicações desta dissertação, a função Lognormal multivariada se apresenta suficientemente adequada, como será mostrado mais à frente. A Figura 18 mostra exemplos de distribuições Gama e Lognormal.

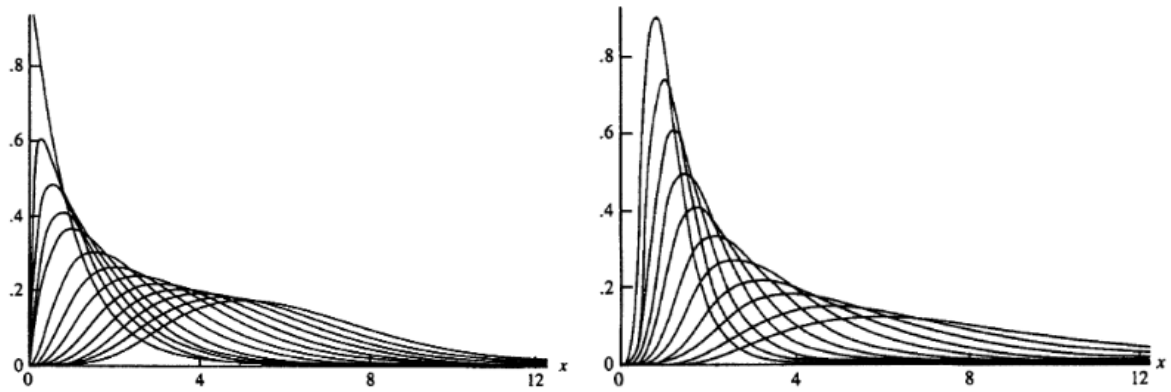
A distribuição Lognormal multivariada pode ser deduzida a partir da mudança de variáveis em uma distribuição Gaussiana multivariada, aplicando-se a função exponencial (TARMAST, 2001):

$$Y_i = \exp\left(X_i\right), \tag{36}$$

onde X_i é o *i-ésimo* elemento do vetor de variáveis aleatórias descritas pela distribuição Gaussiana e Y_i é o *i-ésimo* elemento do vetor das novas variáveis aleatórias, que serão modeladas pela distribuição Lognormal.

O procedimento de transformação das múltiplas variáveis (PEEBLES, 1987), que será descrito a seguir, pode ser aplicado a este caso devido ao fato de a função exponencial ser contínua e de valor único, além de possuir derivada parcial em toda parte de seu

Figura 18 - Exemplos de distribuições (a) Gama e (b) Lognormal



(a) (b)

Fonte: Cho, Bowman e North (2004).

domínio e também uma função inversa contínua, a qual é definida pela Equação (37).

$$X_i = \ln(Y_i)\,. \tag{37}$$

Essas condições fazem com que para cada valor no espaço amostral conjunto de $\mathbf{X}$ exista apenas um valor correspondente no espaço do novo vetor de variáveis $\mathbf{Y}$.

Assim, definindo uma região limitada de pontos no espaço de $\mathbf{X}$ como S_x, e S_y como a região correspondente de pontos no espaço de $\mathbf{Y}$, gerado a partir de $\mathbf{X}$, tem-se:

$$\int_{S_x} \cdots \int p_{x_1,...,x_N}(x_1,...,x_N)dx_1...dx_N = \int_{S_y} \cdots \int p_{y_1,...,y_N}(y_1,...,y_N)dy_1...dy_N, \tag{38}$$

ou seja, a probabilidade de um ponto em S_x é a mesma de seu ponto correspondente em S_y. A Equação (38) mostra isso em termos da probabilidade conjunta.

Aplicando a mudança de variáveis definida na Equação (37) ao lado esquerdo da Equação (38), tem-se:

$$\begin{aligned}&\int_{S_x} \cdots \int p_{x_1,...,x_N}(x_1,...,x_N)dx_1...dx_N \\ &= \int_{S_y} \cdots \int p_{x_1,...,x_N}(x_1 = \ln Y_1, ...x_N = \ln Y_N)|J|dy_1...dy_N,\end{aligned} \tag{39}$$

onde $|J|$ é a magnitude do determinante da matriz de transformação, definida como:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \ln Y_1}{\partial Y_1} & \frac{\partial \ln Y_1}{\partial Y_2} & \cdots & \frac{\partial \ln Y_1}{\partial Y_N} \\ \frac{\partial \ln Y_2}{\partial Y_1} & \frac{\partial \ln Y_2}{\partial Y_2} & \cdots & \frac{\partial \ln Y_2}{\partial Y_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \ln Y_N}{\partial Y_1} & \frac{\partial \ln Y_N}{\partial Y_2} & \cdots & \frac{\partial \ln Y_N}{\partial Y_N} \end{bmatrix} = \begin{bmatrix} \frac{1}{Y_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{Y_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{Y_N} \end{bmatrix} \tag{40}$$

Igualando o lado direito das Equações (38) e (39), chega-se à relação apresentada na Equação (41).

$$p_{y_1,...,y_N}(y_1,...,y_N) = p_{x_1,...,x_N}(x_1 = \ln Y_1,...,x_N = \ln Y_N)|J|, \tag{41}$$

na qual o lado direito corresponde a PDF da distribuição Gaussiana, avaliada na variável descrita na Equação (37), multiplicada pela magnitude do jacobiano.

Finalmente, a PDF da distribuição Lognormal multivariada, em sua forma geral, é apresentada na Equação (42).

$$p(\mathbf{y}) = \frac{1}{\left(\prod_{k=0}^{N-1} y_k\right) |\mathbf{C}|^{\frac{1}{2}} (2\pi)^{\frac{N}{2}}} \exp\left[-\frac{(\ln \mathbf{y} - \boldsymbol{\mu})^T \mathbf{C}^{-1} (\ln \mathbf{y} - \boldsymbol{\mu})}{2}\right], \tag{42}$$

em que $\mathbf{y}$ é o vetor de variáveis aleatórias do processo que está sendo modelado, N é o número total de variáveis, $\mathbf{C}$ é a matriz de covariância e $\boldsymbol{\mu}$ é o vetor de médias.

Como na presente dissertação o processo a ser modelado é o ruído do sinal recebido, o qual foi definido na Equação (27) utilizando-se uma simplificação que fornece $\mathbf{n} = \mathbf{r} - A\mathbf{s}$, neste caso, a PDF Lognormal pode ser escrita especificamente como:

$$p(\mathbf{n}) = \frac{1}{\left[\prod_{k=0}^{N-1} (r_k - As_k)\right] |\mathbf{C}|^{\frac{1}{2}} (2\pi)^{\frac{N}{2}}} \exp\left\{-\frac{[\ln(\mathbf{r} - A\mathbf{s}) - \boldsymbol{\mu}]^T \mathbf{C}^{-1} [\ln(\mathbf{r} - A\mathbf{s}) - \boldsymbol{\mu}]}{2}\right\}. \tag{43}$$

A técnica de maximização através da determinação da raiz da derivada não é muito útil quando se trata da Equação (43), pois não é possível encontrar uma expressão fechada para a amplitude A a partir de sua derivada. Assim, nesta dissertação, optou-se pela utilização de um método conhecido como *busca exaustiva*. Nesse procedimento, os valores do parâmetro de interesse são variados, dentro de uma faixa de valores factível, e aquele que der como resultado o maior valor da PDF é escolhido. Neste trabalho, essa faixa foi escolhida tomando-se como base o valor da amplitude estimada pelo método MLE Gaussiano, e fazendo uma variação entre $\pm f$, em que f é um valor que define os limites superior e inferior para teste do parâmetro de interesse e foi selecionado com base no conhecimento do comportamento das amplitudes dos sinais e em uma análise empírica dos métodos.

3.1 **Pré-processamento para o MLE**

O empilhamento de sinais introduz ao ruído uma componente com características não-Gaussianas. Assim, a matriz de covariância, presente nas PDFs Gaussiana e Log-normal, se torna ineficaz no processo de descorrelação das variáveis aleatórias no sentido mais amplo (RIMES et al., 2020). Visando estudar a possível influência da dependência estatística, presente nos resultados, um pré-processamento foi aplicado aos dados.

3.1.1 Informação Mútua

Inicialmente, consideram-se duas variáveis aleatórias X e Y. Essas variáveis são *estatisticamente independentes* (HYVÄRINEN; OJA, 2000) se os valores assumidos por X não informam nada sobre os valores que Y poderá assumir, e vice-versa. Matematicamente, isso significa que a probabilidade de ocorrência simultânea das duas é igual à multiplicação das probabilidades de ocorrência individuais. Para variáveis aleatórias discretas, isso pode ser escrito como:

$$P(X = x, Y = y) = P(X = x)P(Y = y) \tag{44}$$

ou

$$p(x, y) = p(x)p(y). \tag{45}$$

Neste caso, $p(x, y)$ é chamada de *probabilidade conjunta*, enquanto $p(x)$ e $p(y)$ são as *probabilidades marginais*, que podem ser definidas da seguinte forma:

$$p(x) = P(X = x) = \sum_{y} p(x, y) \tag{46}$$

$$p(y) = P(Y = y) = \sum_{x} p(x, y) \tag{47}$$

Nesta dissertação, como será mostrado posteriormente, as amostras de sinais de ruído são compostas por variáveis aleatórias *estatisticamente dependentes*, ou seja, que possuem probabilidades conjunta e marginais que não satisfazem a Equação (45). Para o desenvolvimento do trabalho, torna-se útil obter-se uma maneira de quantificar a dependência entre as variáveis, e isso pode ser feito utilizando-se o conceito de Informação Mútua.

A Informação Mútua (ou MI, do inglês, *Mutual Information*) (COVER; THOMAS, 2006) contida entre duas variáveis aleatórias fornece uma métrica para determinar quão dependentes essas variáveis são entre si. Se, e somente se, as variáveis forem indepen-

dentes, a informação mútua será igual a zero. A definição matemática desta informação, para duas variáveis aleatórias discretas X e Y, é dada pela Equação (48).

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}, \tag{48}$$

onde $p(x,y)$ é a função de probabilidade conjunta das variáveis e $p(x)$ e $p(y)$ são as probabilidades marginais respectivas. O conceito de informação mútua está diretamente ligado ao de entropia, que estima a quantidade de informação necessária para descrever uma dada variável.

Nesta dissertação, a Equação (48) foi utilizada e a informação mútua foi calculada dois-a-dois, ou seja, o sinal é composto por 7 variáveis aleatórias e foi feita uma análise de como cada variável depende de cada uma das outras, separadamente.

3.1.2 Análise de Componentes Independentes (ICA)

O método conhecido como ICA (do inglês, *Independent Component Analysis*) (COMON, 1994) baseia-se na procura de uma transformação linear que seja capaz de minimizar a dependência entre componentes aleatórias estatisticamente dependentes de um vetor. A ICA é semelhante a PCA (do inglês, *Principal Component Analysis*), no entanto, esta só é capaz de fazer a descorrelação até estatísticas de segunda ordem, enquanto que a ICA funciona para estatísticas de ordem superior.

A definição precisa é dada, primeiramente, considerando-se o modelo estatístico linear dado pela Equação (49), onde $\mathbf{y}$ é o vetor de observações, $\mathbf{n}$ é o vetor do ruído presente nos dados, $\mathbf{x}$ é um vetor de componentes estatisticamente independentes e $\mathbf{M}$ é uma matriz retangular, tendo um número de colunas menor ou igual ao número de linhas.

$$\mathbf{y} = \mathbf{M}\mathbf{x} + \mathbf{n}. \tag{49}$$

Os parâmetros $\mathbf{y}$, $\mathbf{x}$ e $\mathbf{n}$ são vetores aleatórios com média zero e covariância finita, que podem ser definidos no conjunto dos reais ou dos complexos. A ICA tem como objetivo estimar a matriz $\mathbf{M}$ e o vetor $\mathbf{x}$, dada uma certa quantidade de observações $\mathbf{y}$. No entanto, a presença do ruído geralmente torna impossível que $\mathbf{x}$ seja completamente recuperado. Assim, de forma alternativa, modela-se o problema pela Equação (50),

$$\mathbf{y} = \mathbf{F}\mathbf{z}, \tag{50}$$

e sua resolução se torna dependente da maximização de uma *função contraste*, que, como

mostrado por Comon (1994), ocorre quando as componentes de $\mathbf{z}$ são estatisticamente independentes.

Na literatura, existem alguns algoritmos para resolver o problema da ICA. Um dos mais eficientes é o FastICA (HYVÄRINEN; OJA, 2000), que trabalha de forma a encontrar n vetores unitários de peso, $\mathbf{w}$, em uma rede com n neurônios capazes de atualizar esses vetores utilizando uma regra de aprendizado, de modo que a multiplicação entre a forma transposta de cada vetor e as ocorrências, ou seja, $\mathbf{w}_k^T\mathbf{y}$, $k = 1, 2, \ldots, n$, maximize a não-Gaussianidade. Nesta dissertação, o FastICA foi aplicado através do uso de um algoritmo já desenvolvido e difundido, que é brevemente detalhado nesta seção.

Como pré-processamento para a ICA, para que a estimação seja mais simples e melhor condicionada, é necessário que os dados tenham média zero e, em seguida, passem por um processo de branqueamento (do inglês, *whitening*). Na primeira parte, basta que a média dos dados seja calculada e subtraída de cada ocorrência. Já para o branqueamento, o procedimento é um pouco mais longo, sendo necessário encontrar, a partir do vetor de observações $\mathbf{y}$, utilizando uma transformação linear, um novo vetor $\widetilde{\mathbf{y}}$, que tenha componentes descorrelacionadas e variâncias iguais a 1. Isto significa que a matriz de covariância de $\widetilde{\mathbf{y}}$ será igual à matriz identidade, ou seja,

$$E\left\{\widetilde{\mathbf{y}}\widetilde{\mathbf{y}}^T\right\} = \mathbf{I}. \tag{51}$$

Uma opção bastante utilizada para este procedimento, e que foi empregue no algoritmo aqui descrito, é a decomposição de autovalores (ou EVD, do inglês, *eigenvalue decomposition*) da matriz de covariância na forma

$$E\left\{\mathbf{y}\mathbf{y}^\mathbf{T}\right\} = \mathbf{O}\mathbf{D}\mathbf{O}^T, \tag{52}$$

onde $\mathbf{O}$ e $\mathbf{D}$ são, respectivamente, a matriz ortogonal de autovetores e a matriz diagonal de autovalores de $E\left\{\mathbf{y}\mathbf{y}^\mathbf{T}\right\}$. Assim, o processo de branqueamento pode ser feito utilizando-se a Equação (53).

$$\widetilde{\mathbf{y}} = \mathbf{E}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^T\mathbf{x}, \tag{53}$$

onde a matriz $\mathbf{D}^{\frac{1}{2}}$ é calculada simplesmente pegando-se o inverso da raiz de cada elemento de sua diagonal principal, ou seja, $\mathbf{D}^{-\frac{1}{2}} = diag(d_1^{-\frac{1}{2}}, d_2^{-\frac{1}{2}}, \ldots, d_n^{-\frac{1}{2}})$. Combinando as Equações (50) e (53), pode-se escrever

$$\widetilde{\mathbf{y}} = \mathbf{E}\mathbf{D}^{-\frac{1}{2}}\mathbf{E}^T\mathbf{F}\mathbf{z} = \widetilde{\mathbf{F}}\mathbf{z}. \tag{54}$$

Assim,

$$E\left\{\widetilde{\mathbf{y}}\widetilde{\mathbf{y}}^T\right\} = \widetilde{\mathbf{F}}E\left\{\mathbf{z}\mathbf{z}^T\right\}\widetilde{\mathbf{F}}^T$$
$$= \widetilde{\mathbf{F}}\widetilde{\mathbf{F}}^T$$
$$= \mathbf{I}, \tag{55}$$

e, portanto, a nova matriz $\widetilde{\mathbf{F}}$ é ortogonal. Desta forma, o branqueamento dos dados reduz pela metade o número de parâmetros a ser determinado.

Após o tratamento dos dados, o algoritmo FastICA pode ser aplicado. Primeiramente, o método será desenvolvido para a versão de apenas uma unidade computacional, ou seja, um neurônio artificial. Para maximizar a não-Gaussianidade através da projeção $\mathbf{w}^T\mathbf{y}$, sua medida é feita utilizando-se a aproximação de negentropia, dada pela Equação (56).

$$J\left(\mathbf{w}^T\mathbf{y}\right) \propto \left[E\left\{G\left(\mathbf{w}^T\mathbf{y}\right)\right\} - E\left\{G(\upsilon)\right\}\right]^2, \tag{56}$$

onde G é praticamente qualquer função não-quadrática e υ é uma variável Gaussiana de média zero e variância unitária. A escolha certa de G é importante. É possível obter-se estimadores mais robustos optando-se por funções que não cresçam muito rapidamente. As Equações (57) e (58) apresentam dois exemplos de escolhas de G que vêm se mostrando bastante úteis.

$$G_1(u) = \frac{1}{a_1}\log(\cosh(a_1 u)) \tag{57}$$

$$G_2(u) = -\exp\left(-\frac{u^2}{2}\right), \tag{58}$$

onde a_1 é uma constante tal que $1 \leq a_1 \leq 2$.

Para encontrar o máximo de não-Gaussianidade, o FastICA utiliza um método de iteração de ponto fixo, podendo também ser utilizada uma derivação de Newton aproximada. As derivadas das funções dadas nas Equações (57) e (58) são:

$$\dot{G}_1 = \tanh(a_1 u) \tag{59}$$

$$\dot{G}_2 = u\exp\left(-\frac{u^2}{2}\right), \tag{60}$$

onde a_1 segue a mesma condição que para a função primitiva referente, sendo, aqui, geralmente escolhida como $a_1 = 1$.

Assim, o algoritmo do FastICA para a iteração de ponto fixo consiste em

1. Escolha um vetor aleatório $\mathbf{w}$ inicial;

2. Faça $\mathbf{w}^+ = E\left\{\mathbf{y}\dot{G}\left(\mathbf{w}^T\mathbf{y}\right)\right\} - E\left\{\ddot{G}\left(\mathbf{w}^T\mathbf{y}\right)\right\}\mathbf{w}$;
3. Faça $\mathbf{w}^+ = \mathbf{w}^+/\|\mathbf{w}^+\|$;
4. Se o critério de convergência
   $\mathbf{w} \cdot \mathbf{w}^+ \approx 1$
   for satisfeito, o algoritmo pode ser encerrado. Senão, faça $\mathbf{w} = \mathbf{w}^+$ e volte para o passo 2.

Nesta dissertação, o desenvolvimento do FastICA utilizando o método de Newton será omitido, já que o objetivo principal é mostrar como o algoritmo funciona e sua linha de raciocínio. No entanto, toda sua construção é desenvolvida de forma detalhada por HyvÄrinen e Oja (2000).

Para estimar um número n de componentes independentes, é necessário adaptar este algoritmo, de forma a utilizar n neurônios e n vetores de peso. Para aplicar o algoritmo do ponto fixo neste caso, é necessário, a cada iteração, descorrelacionar as saídas $\mathbf{w}_1^T\mathbf{y}, \mathbf{w}_2^T\mathbf{y}, \ldots, \mathbf{w}_n^T\mathbf{y}$. Uma das formas para isso é estimando as componentes uma por uma, um esquema que tem como base o processo de Gram–Schmidt.

A explicação deste método é dada pensando na estimativa da p-*ésima* componente, sendo $p = 1, 2, \ldots, n-1$. Com $\mathbf{w}_p$ definido, $\mathbf{w}_{p+1}$ é estimado utilizando-se o algoritmo do ponto fixo para uma unidade computacional e, em seguida, subtrai-se deste o somatório das projeções $\mathbf{w}_{p+1}^T\mathbf{w}_j\mathbf{w}_j$, $j = 1, 2, \ldots, p$. Logo após, $\mathbf{w}_{p+1}$ é renormalizado. Assim, o algoritmo se torna:

1. Escolha um vetor aleatório $\mathbf{w}_{p+1}$ inicial;
2. Faça $\mathbf{w}_{p+1}^+ = E\left\{\mathbf{y}\dot{G}\left(\mathbf{w}_{p+1}^T\mathbf{y}\right)\right\} - E\left\{\ddot{G}\left(\mathbf{w}_{p+1}^T\mathbf{y}\right)\right\}\mathbf{w}_{p+1}$;
3. Faça $\mathbf{w}_{p+1}^+ = \mathbf{w}_{p+1}^+/\|\mathbf{w}_{p+1}^+\|$;
4. Faça $\mathbf{w}_{p+1}^+ = \mathbf{w}_{p+1}^+ - \sum_{j=1}^{p} \mathbf{w}_{p+1}^{+T}\mathbf{w}_j\mathbf{w}_j$;
5. Faça $\mathbf{w}_{p+1}^+ = \mathbf{w}_{p+1}^+/\sqrt{\mathbf{w}_{p+1}^{+T}\mathbf{w}_{p+1}}$
6. Se o critério de convergência
   $\mathbf{w}_{p+1} \cdot \mathbf{w}_{p+1}^+ \approx 1$
   for satisfeito, o algoritmo pode ser encerrado. Senão, faça $\mathbf{w}_{p+1} = \mathbf{w}_{p+1}^+$ e volte para o passo 2.

Esta é uma das formas de estimar várias componentes independentes com o FastICA. Outras duas maneiras alternativas são apresentadas por HyvÄrinen e Oja (2000).

4 RESULTADOS

Neste capítulo, são apresentados os resultados obtidos das análises para dados simulados e experimentais. Primeiramente, a Seção 4.1 apresenta os parâmetros que foram utilizados para avaliar o desempenho dos métodos estudados. Em seguida, na Seção 4.2, são mostrados a etapa de geração dos dados simulados e toda a análise correspondente. Já a Seção 4.3 traz os dados reais provenientes do TileCal e suas respectivas modelagem e análise de eficiência. Por último, na Seção 4.4, encontra-se um estudo sobre a dependência estatística das variáveis de ruido, simulados e reais, a fim de observar como esta característica poderia estar afetando os resultados obtidos.

4.1 Parâmetros de avaliação

Antes de estudar o desempenho dos métodos quando aplicados aos dados, é importante apresentar os parâmetros que serão usados em suas comparações. As medidas definidas aqui são utilizadas tanto para os dados simulados quanto para os reais.

- Erro:

Para cada simulação, uma amplitude verdadeira foi gerada e cada um dos métodos estimou sua amplitude correspondente. Para o cálculo do erro, a amplitude verdadeira foi subtraída das amplitudes estimadas, dando como resultado um valor em contagens de ADC:

$$e = A_{est} - A_{verd}. \tag{61}$$

Este valor é calculado para cada sinal individualmente e, ao final, uma distribuição de erro pode ser apresentada. Destes histogramas, foram computados média e desvio padrão, juntamente com suas barras de erro correspondentes, quando são considerados os dados simulados, para os quais foram gerados 50 conjuntos de 100.000 sinais cada, para cada ocupação.

- Média:

Uma média aritmética simples foi utilizada para calcular a média do erro para cada um dos métodos, ou seja, a todos os valores foram dados pesos iguais, e então o resultado final pode ser gerado fazendo-se um somatório dos termos e dividindo-se pela quantidade destes:

$$m = \frac{1}{n}\sum_{i=1}^{n} e_i, \tag{62}$$

onde n é o número total de erros calculados, que, aqui, é igual à quantidade de sinais no conjunto de teste, ou seja, 50.000 para os dados simulados e 321.977, 280.012 e 73.715 para os dados reais nas luminosidades relativas a $\mu = 30$, 50 e 90, respectivamente.

A média é um parâmetro importante quando se trata de distribuições de erro, pois fornece uma informação do quão próximos de zero os erros calculados estão. Quanto menor o valor do erro, mais eficiente é o método estudado.

- Desvio padrão:

O desvio padrão é uma medida que fornece uma informação sobre o quão dispersos os dados de um conjunto estão em torno da média. Quando associado ao erro, quanto menor o valor deste parâmetro, menor o espalhamento dos dados e, usualmente, mais adequados estes se mostram. A Equação (63) mostra como o desvio padrão pode ser calculado.

$$\sigma = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(e_i - m)^2}, \tag{63}$$

onde m é a média, definida na Equação (62), e n, novamente, é o número total de ocorrências.

Assim, o cálculo da medida da dispersão de um conjunto de amostras é dado considerando-se o valor de cada amostra subtraído da média aritmética de todo o conjunto. O desvio padrão é um conceito útil na análise de distribuições de erro, principalmente quando este se mostra simetricamente distribuído e com características Gaussianas.

- Matriz de covariância:

A matriz de covariância condensa em si os valores da covariância entre variáveis aleatórias. Sua diagonal é composta pela variância de cada uma das variáveis, enquanto que os elementos acima ou abaixo descrevem o quão dependentes estas são umas das outras. A covariância será zero se, e somente se, as variáveis forem linearmente descorrelacionadas.

Uma matriz de covariância geral é definida na Equação (64).

$$C = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1N} \\ c_{21} & c_{22} & \cdots & c_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{N1} & c_{N2} & \cdots & c_{NN} \end{bmatrix}, \tag{64}$$

onde

$$c_{ij} = E\left\{\left(X_i - E\left\{X_i\right\}\right)\left(X_j - E\left\{X_j\right\}\right)\right\}.$$

Assim, um dado $c_{i,j}$ é capaz de quantificar a correlação (relação linear) entre as variáveis aleatórias i e j correspondentes.

4.2 **Dados Simulados**

Esta seção apresenta todos os aspectos dos dados simulados, os quais são baseados nos parâmetros do calorímetro de telhas do ATLAS, ou TileCal. Na Seção 4.2.1 é explicado como os sinais simulados foram gerados, mostrando também a forma do pulso e como o problema do empilhamento de sinais pode afetar na estimação da amplitude. A Seção 4.2.2 mostra as distribuições de ruído para ocupações selecionadas e motiva, através de parâmetros estatísticos, o desenvolvimento do método proposto nesta dissertação. Finalmente, a Seção 4.2.3 apresenta os resultados obtidos quando os métodos são aplicados aos dados simulados, abordando aspectos positivos e negativos de cada um.

4.2.1 Banco de Dados

Para as simulações e análise dos métodos descritos, foram gerados sinais simulados, devido ao fato de que, sobre estes, é possível ter-se um maior controle, o que torna mais fácil a averiguação das vantagens e desvantagens de cada abordagem. Para definir os parâmetros destes sinais, optou-se por utilizar o calorímetro TileCal como base, já que os dados reais, que serão analisados posteriormente nesta dissertação, são provenientes deste detector.

Os sinais aqui estudados são formados pelas componentes do sinal de interesse, do ruído eletrônico e do ruído empilhamento. Este último ocorre devido ao fenômeno de empilhamento de sinais, que foi descrito na Seção 1.5. Assim, os sinais de ruído simulados são compostos de duas partes: uma Gaussiana, que descreve o ruído eletrônico, e uma não-Gaussiana, que representa o ruído de empilhamento.

Para simular o ruído eletrônico, uma distribuição Gaussiana de média zero e desvio padrão de 1,5 contagens de ADC foi utilizada. A sigla ADC refere-se a um tipo de medida digital que é fornecida a partir de dados originalmente analógicos. Essa unidade foi usada pois o sistema de calibração do TileCal mede e converte de forma eficiente informações que estejam em contagens de ADC para a unidade de energia, GeV (do inglês, *Giga electron-Volt*) (MARJANOVIĆ, 2019). O sinal recebido está, inicialmente, na forma analógica, porém um conversor analógico-digital (do inglês, *Analog to Digital Converter*) (MARSHALL, 2014) faz a conversão.

O ruído de empilhamento, por outro lado, depende da posição espacial do canal de leitura e das condições de operação do LHC, principalmente da luminosidade utilizada

nas colisões. O conceito de luminosidade foi definido anteriormente e é importante, pois afeta diretamente a eficiência dos métodos aplicados na estimação da amplitude do sinal de interesse.

A distribuição para o ruído de empilhamento está relacionada com a distribuição de energia no calorímetro. O espectro de hádrons em regime de elevado momento transverso (p_T) depende das partículas utilizadas no experimento e das características do acelerador. No entanto, as regiões de baixo momento transverso, onde interações de espalhamentos múltiplos predominam, são responsáveis pela maior parte dos espectros e possuem características exponenciais (KHANDAI et al., 2013).

Devido a esses aspectos do TileCal, o ruído de empilhamento utilizado nas simulações foi gerado segundo uma distribuição exponencial, de média igual a 100 contagens de ADC e acrescida de um pedestal de 50 contagens de ADC. Com o intuito de analisar várias condições de luminosidade, um parâmetro chamado de *ocupação* foi variado dentro de uma faixa de 0% a 100%. Este parâmetro diz respeito à probabilidade de uma célula do calorímetro receber um sinal quando ocorre uma colisão, ou seja, quanto maior a ocupação, maior a chance de um sinal ser lido. Em ocupações muito altas, a possibilidade de haver empilhamento de sinais também aumenta, pois se torna mais provável que um segundo sinal chegue à janela de leitura antes que o primeiro, aquele de interesse, seja completamente amostrado.

Para cada condição de ocupação, foram gerados 50 conjuntos distintos de sinais de ruído (eletrônico + empilhamento), com 100.000 sinais cada. Cada um desses conjuntos foi dividido em dois outros, formando dois conjuntos de 50.000 sinais. O primeiro, chamado de *conjunto de desenvolvimento*, foi utilizado para calcular os parâmetros das distribuições, ou seja, as médias e as matrizes de covariância. Para a distribuição lognormal, esses fatores são computados utilizando-se o logaritmo dos dados, o que levou à necessidade do descarte dos valores negativos. Já para a Gaussiana, os cálculos podem ser feitos usando-se diretamente os dados. O segundo conjunto de dados de ruído foi chamado de *conjunto de teste* e utilizado para a análise de eficiência dos métodos.

O sinal de interesse, por sua vez, foi simulado utilizando-se uma distribuição exponencial com média igual àquela calculada do conjunto de teste, que foi determinada após os sinais terem sido gerados e fazendo-se a subtração do pedestal nos mesmos. Sob estas condições, foi gerado, então, um conjunto de 50.000 sinais, chamado *conjunto de interesse*.

Para a análise de eficiência dos métodos, o conjunto de interesse foi multiplicado por um vetor contendo as amostras do pulso de referência (amplitude igual a 1) e essa multiplicação foi somada ao conjunto de teste. O resultado dessas operações é o sinal completo, que contem o sinal de interesse imerso em ruído, do qual a amplitude deverá ser estimada pelos métodos.

4.2.2 Modelagem do ruído

Com o intuito de mitigar o problema da não-Gaussianidade do ruído, aspecto que afeta a eficiência dos métodos de estimação já difundidos e utilizados, optou-se por buscar uma distribuição que o descrevesse de forma mais adequada. Como o pulso do TileCal é unipolar (ver Seção 1.5), sempre positivo, e seus sinais possuem assinatura exponencial, o ruído de empilhamento pode ser visto como uma soma de exponenciais. A distribuição resultante é a Erlang, um tipo específico de distribuição Gama (ver Capítulo 3).

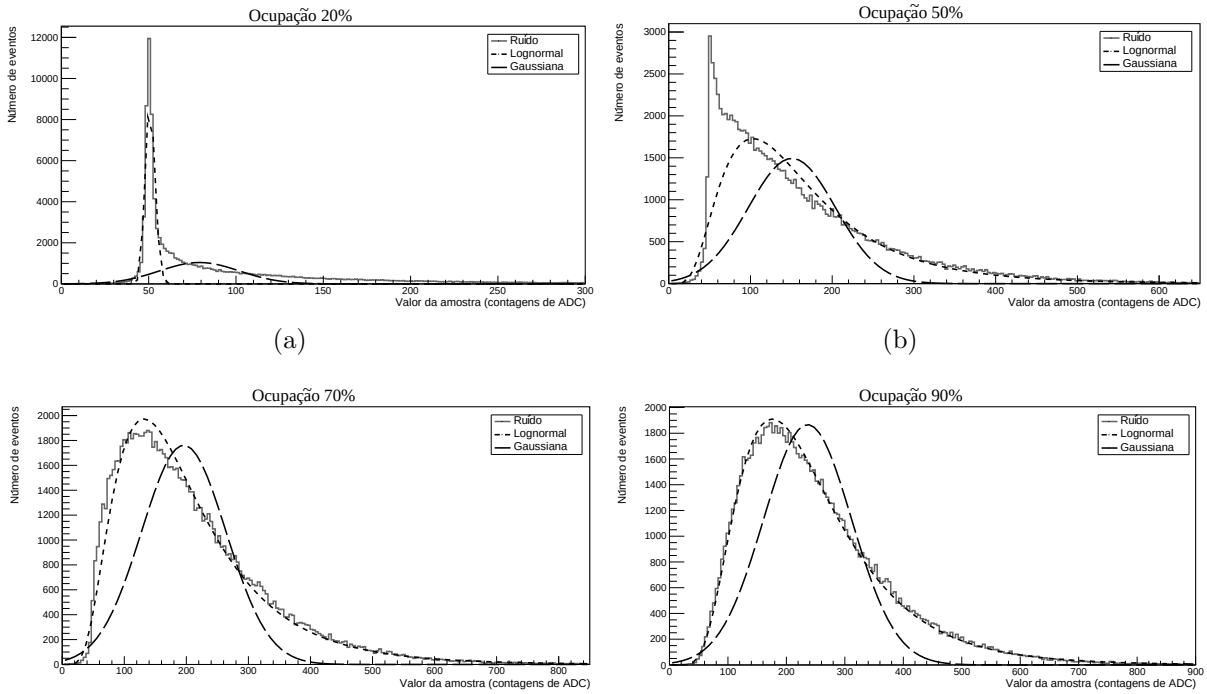
Os sinais provenientes do TileCal são compostos por 7 amostras, ou seja, 7 variáveis aleatórias. A distribuição Gama, no entanto, não possui uma PDF multivariada bem estabelecida. Assim, decidiu-se pela utilização de uma distribuição Lognormal, que possui uma PDF multivariada com parâmetros relativamente simples de serem calculados (ver Capítulo 3).

A Figura 19 mostra as distribuições de ruído da terceira amostra para quatro ocupações diferentes, juntamente com os ajustes para as curvas das distribuições lognormal e Gaussiana. Como todas as 7 amostras possuem distribuições bastante semelhantes, esse ajuste, aqui, será mostrado apenas para a amostra 3. A binagem de cada histograma foi escolhida de acordo com a razão entre o χ^2 (PAPOULIS; PILLAI, 2002) e o número de graus de liberdade, ndf: quanto mais próximo de 1 estiver o valor da razão, mais adequado é o ajuste da função nos dados. O número de bins foi variado em uma faixa de 40 a 200, em cada ocupação, e o χ^2/ndf foi calculado em todos os casos, de forma a encontrar o menor valor para ambas as distribuições. O χ^2/ndf é um parâmetro que diz o quão próximo da função utilizada estão os dados que se deseja modelar, fornecendo uma medida da qualidade do ajuste da distribuição.

A Figura 20 mostra como o valor do χ^2/ndf varia para a faixa de ocupação de 10% a 100%, tendo em mente que na ocupação de 0% não há empilhamento de sinais e, portanto, o único ruído presente é o eletrônico, que é perfeitamente Gaussiano, não se aplicando o ajuste da função lognormal. As barras de erro foram aumentadas em um fator de 5, para que pudessem ser observadas melhor. Seus cálculos se deram através do desvio padrão do parâmetro χ^2/ndf, considerando os 50 arquivos de ruído gerados para cada ocupação.

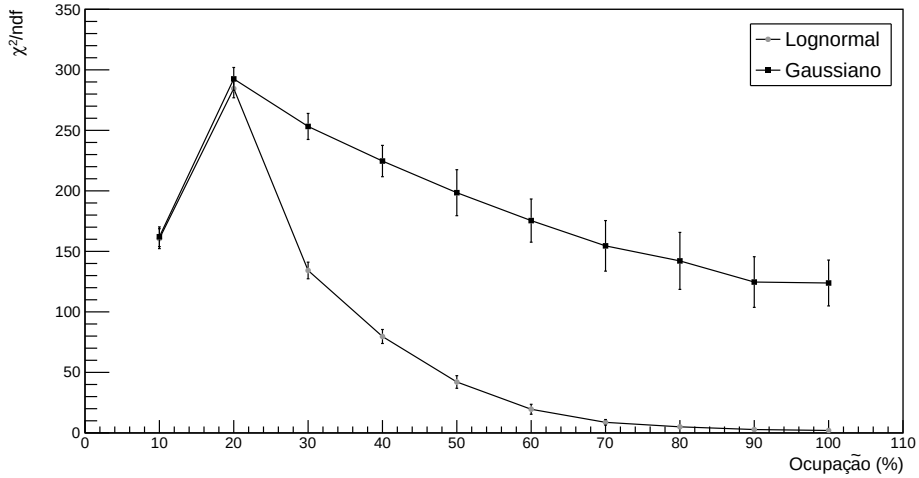
A Tabela 1 apresenta os valores exatos dos pontos que a Figura 20 mostra, juntamente com seus erros nos valores originais, para todas as ocupações. É possível notar que, a partir da ocupação de 20%, o valor de χ^2/ndf diminui cada vez mais, para ambas as distribuições. Na ocupação 10% este valor é menor porque quase não há empilhamento, o que faz com que o ruído seja bastante Gaussiano. Nesse caso, ambas as funções conseguem descrevê-lo com praticamente a mesma eficiência. Na ocupação de 20%, a componente não-linear começa a afetar de forma mais contundente, criando uma cauda positiva maior no histograma do ruído, como pode ser visto na Figura 19(a).

Figura 19 - Distribuições de ruído e ajuste lognormal e Gaussiano para dados simulados nas ocupações (a) 20%, (b) 50%, (c) 70% e (d) 90%.



Fonte: A autora (2021).

Figura 20 - Parâmetro χ^2/ndf para dados simulados em toda a faixa de ocupações, comparando os ajustes com as distribuições lognormal e Gaussiana.



Fonte: A autora (2021).

Tabela 1 - Valores de χ^2/ndf e erros, para todas as ocupações.

	χ^2/ndf	
Ocupação	Gaussiano	Lognormal
10%	162,085 ± 1,627	160,500 ± 1,644
20%	292,518 ± 1,888	284,564 ± 1,526
30%	253,234 ± 2,151	134,226 ± 1,371
40%	224,644 ± 2,587	79,663 ± 1,155
50%	198,596 ± 3,800	42,076 ± 1,043
60%	175,434 ± 3,567	19,524 ± 0,821
70%	154,559 ± 4,178	8,714 ± 0,465
80%	142,124 ± 4,709	5,009 ± 0,369
90%	124,654 ± 4,187	2,773 ± 0,290
100%	123,869 ± 3,789	2,012 ± 0,238

Fonte: A autora (2021).

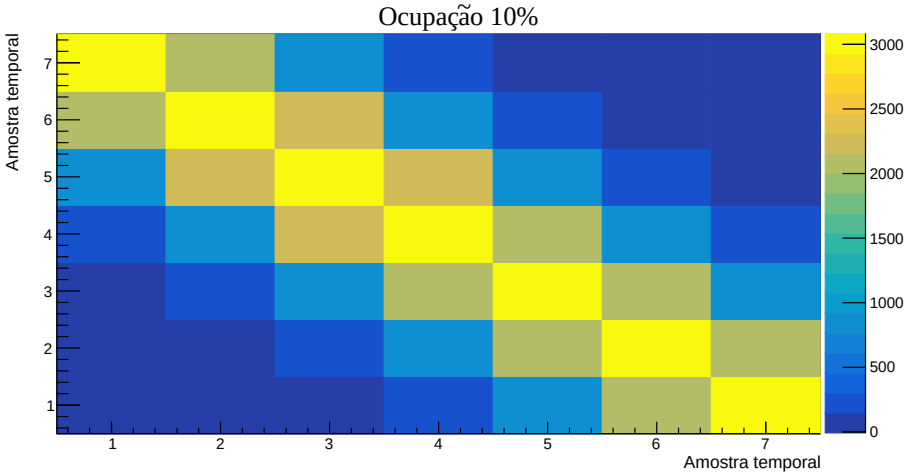
Conforme a ocupação aumenta, a componente do ruído de empilhamento se torna mais presente, fazendo com que a cauda positiva aumente, enquanto a distribuição adquire um formato cada vez mais próximo da Lognormal. Esse processo, que pode ser observado na Figura 19, faz com que os valores do parâmetro χ^2/ndf fiquem cada vez mais próximos da unidade quando se tratando do ajuste Lognormal. Apesar de também ocorrer diminuição para a distribuição Gaussiana, a Lognormal parece conferir uma representação mais eficiente aos dados.

As Figuras 21 e 22 mostram os gráficos relativos às matrizes de covariância do ruído para as ocupações de 10% e 50%, respectivamente. Percebe-se que os valores fora da diagonal principal, os quais dão uma noção de quantificação para a dependência entre as variáveis aleatórias, são significativamente maiores para a ocupação de 50%, onde o empilhamento deve ser maior do que na condição de 10%. Isto se deve ao fato de o empilhamento de sinais aumentar a correlação entre as amostras, além de estas perderem suas características Gaussianas. A consequência disso é que a inversa da matriz de covariância, que é utilizada nos métodos baseados no MLE e tem como uma de suas funções a descorrelação das variáveis, tem sua eficiência diminuída, deteriorando o desempenho de tais métodos. As matrizes que deram origem às Figuras 21 e 22 podem ser encontradas no Apêndice B.

4.2.3 Análise de Eficiência

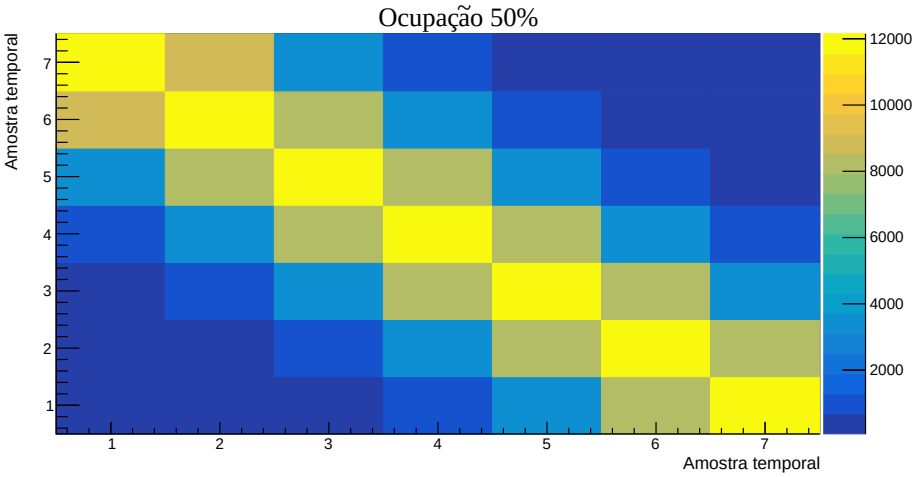
Após a modelagem do ruído, foi feita a análise de eficiência do método não-linear proposto, juntamente com os lineares, de forma a definir se alguma melhora foi apresen-

Figura 21 - Gráfico relativo à matriz de covariância do ruído simulado para a ocupação de 10%.



Fonte: A autora (2021).

Figura 22 - Gráfico relativo à matriz de covariância do ruído simulado para a ocupação de 50%.



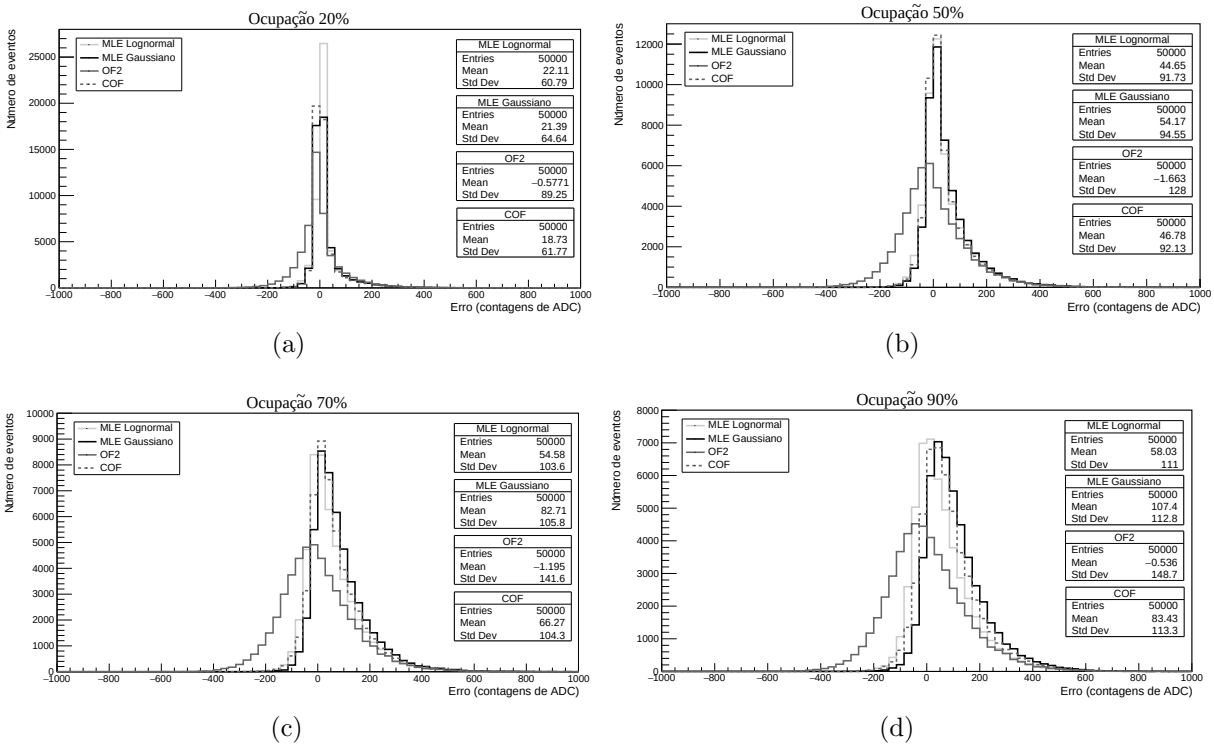
Fonte: A autora (2021).

tada e também observar possíveis dificuldades.

Inicialmente, todos os métodos foram aplicados aos dados e as amplitudes segundo cada um foram calculadas. Uma amplitude simulada foi gerada aleatoriamente para cada sinal, segundo uma distribuição exponencial (ver Seção 4.2.1), e os métodos deveriam ser capazes de estimar valores o mais próximos possíveis destes originais. Assim, os erros dos métodos foram calculados de acordo com a definição dada na Seção 4.1. A Figura 23 mostra os histogramas das distribuições dos erros para quatro ocupações selecionadas: 20%, 50%, 70% e 90%. Com o intuito de facilitar a análise dos resultados, as Tabelas 2 e

3 apresentam, respectivamente, os valores exatos de média e desvio padrão dos erros para todas as ocupações.

Figura 23 - Histogramas das distribuições dos erros para todos os métodos nas ocupações (a) 20%, (b) 50%, (c) 70% e (d) 90%.



Fonte: A autora (2021).

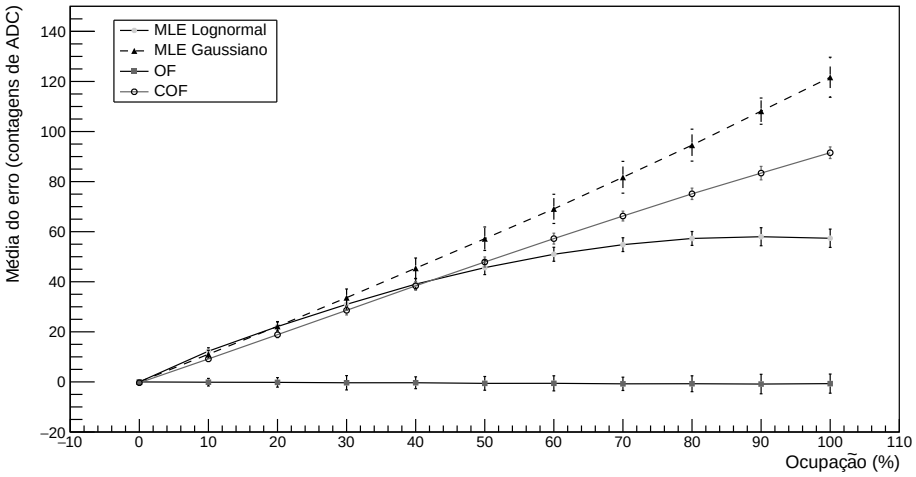
Uma distribuição de erro que demonstre boa eficiência de um método deve ter média centrada o mais próximo possível de zero e baixo valor de desvio padrão. Olhando para a Figura 23, nota-se que o método OF2, principalmente, sofre uma grande perda de eficiência conforme o aumento da ocupação ocorre, demonstrada pela dispersão de sua distribuição de erro. Sua média, por outro lado, se mantém próxima de zero, o que pode ser confirmado na Tabela 2 e é esperado, já que este método não apresenta tendência, devido às restrições que são impostas ao procedimento que calcula os coeficientes do filtro.

Os outros três métodos, no entanto, não apresentam uma perda de eficiência tão grande quanto o OF2, e evoluem de forma semelhante conforme a ocupação aumenta. O MLE Lognormal, quando comparado ao MLE Gaussiano e ao COF, ambos lineares, parece apresentar uma distribuição de erro com menor dispersão e com a média mais próxima de zero, principalmente em condições de ocupação mais altas, onde o empilhamento de sinais é maior.

Para uma análise mais completa, a Figura 24 apresenta os valores da média do erro para todas as ocupações, em cada um dos métodos. Esses valores foram calculados utilizando-se todos os 50 conjuntos gerados para cada ocupação, da seguinte forma: para

cada um desses conjuntos, a análise de eficiência foi feita. Ao final, a média das médias dos erros foi calculada, assim como o desvio padrão entre as 50 medidas, que foi usado para que fosse possível ter-se uma noção do erro estatístico contido nos métodos. Esse erro estatístico é representado por barras de erro, as quais podem ser vistas no gráfico referido. Com o intuito de facilitar a visualização, nesta figura, são mostradas barras de erro com valores multiplicados por um fator de 5, já que os números originais eram pequenos e a análise acabaria sendo prejudicada.

Figura 24 - Média dos erros para todas as ocupações, comparando cada um dos métodos.



Fonte: A autora (2021).

Na Figura 24 observa-se que, para baixas ocupações, onde o empilhamento de sinais é menor, os métodos MLE Gaussiano, COF e MLE Lognormal apresentam valores de média bastante próximos, com o COF se sobressaindo aos outros. A partir da ocupação de 30%, o MLE Gaussiano começa a sofrer um aumento significativo, o que acontece para o COF na condição de 50%. Enquanto isso, o MLE Lognormal parece estabilizar os valores da média de seus erros quando o empilhamento ultrapassa certo patamar. Essa análise pode ser confirmada e melhor explorada quando a Tabela 2 é incluída no estudo.

É possível notar também, ainda na Figura 24, que, para o método MLE Gaussiano, o desvio padrão entre as medidas, representado pelas barras de erro do gráfico, cresce com o aumento da ocupação, enquanto que para os outros três métodos essa dispersão é mantida mais ou menos constante.

Uma fato importante, que deve ser citado, é que o valor da média dos erros pode ser parametrizado e subtraído quando se tratando dos métodos baseados no MLE.

A Figura 25 mostra as curvas de desvio padrão do erro para todos os métodos, em toda a faixa de ocupação. Assim como para a média, todos os 50 conjuntos foram utilizados, sendo feita a análise de eficiência para cada um deles, onde os desvios padrões

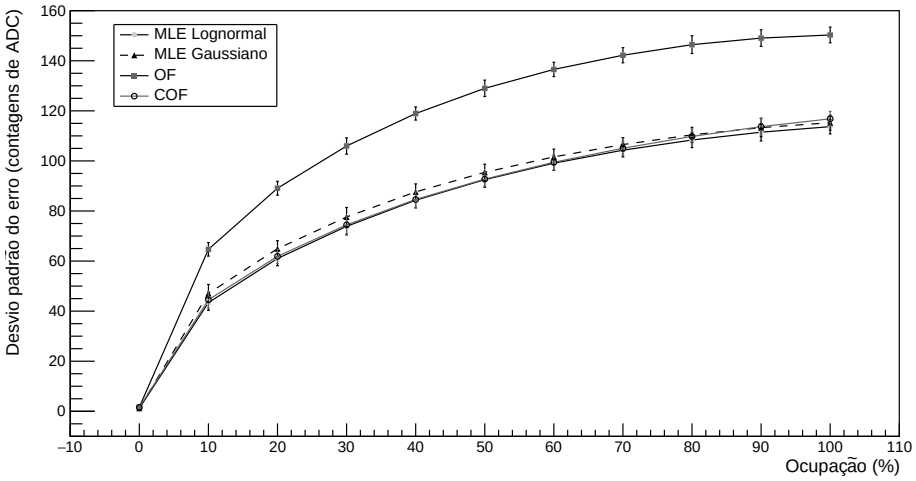
Tabela 2 - Média dos erros, em contagens de ADC, para todas as ocupações.

Parâmetro	Ocupação	MLE Logn	MLE Gauss	OF2	COF
Média	0%	0,004	0,004	0,002	-0,334
	10%	12,600	11,160	-0,253	9,096
	20%	22,110	21,390	-0,577	18,730
	30%	30,740	33,220	-1,016	28,530
	40%	38,340	45,260	-0,329	38,000
	50%	44,650	54,170	-1,663	46,780
	60%	50,500	69,110	-0,820	57,020
	70%	54,580	82,710	-1,195	66,270
	80%	57,720	95,580	-1,060	74,570
	90%	58,030	107,400	-0,536	83,430
	100%	57,330	121,100	-1,060	90,820

Fonte: A autora (2021).

foram calculados e, ao final, a média entre os 50 valores resultantes foi determinada. As barras de erro também foram feitas utilizando-se o desvio padrão das medidas dos conjuntos e um fator de 5 foi utilizado para aumento e facilitação da visualização.

Figura 25 - Desvio padrão dos erros para todas as ocupações, comparando cada um dos métodos.



Fonte: A autora (2021).

Aqui, pode-se perceber que, enquanto o OF apresenta uma grande discrepância, os outros três métodos aparentam valores de desvio padrão bem próximos, principalmente o MLE Lognormal e o COF. A Tabela 3 confirma essas observações.

Para quantificar a eficiência do método MLE Lognormal quando comparado aos outros, foi feito, primeiramente, o cálculo da divisão do valor do desvio padrão do erro do MLE Lognormal pelo desvio padrão do erro do método em questão. Esse valor foi

Tabela 3 - Desvio padrão dos erros, em contagens de ADC, para todas as ocupações.

Parâmetro	Ocupação	MLE Logn	MLE Gauss	OF2	COF
Desvio padrão	0%	1,209	1,209	1,637	1,567
	10%	42,990	46,790	64,940	44,250
	20%	60,790	64,640	89,250	61,770
	30%	73,690	77,350	105,600	74,050
	40%	83,710	86,900	118,700	84,040
	50%	91,730	94,550	128,000	92,130
	60%	98,940	101,300	136,000	99,270
	70%	103,600	105,800	141,600	104,300
	80%	109,000	110,900	147,700	110,300
	90%	111,000	112,800	148,700	113,300
	100%	114,100	115,700	150,900	117,300

Fonte: A autora (2021).

subtraído da unidade e então multiplicado por 100, para obter-se uma medida de porcentagem. A Equação (65) apresenta matematicamente este raciocínio.

$$E_{Ln} = \left(1 - \frac{Std_{erroLn}}{Std_{erroX}}\right) \times 100\%, \tag{65}$$

onde E_{Ln} é a eficiência do método MLE Lognormal, Std_{erroLn} é o desvio padrão do erro calculado para este método e Std_{erroX} é o desvio padrão do método sobre o qual a comparação está sendo feita. Os resultados para estes cálculos são apresentados na Tabela 4, para cada uma das ocupações.

Primeiramente, analisando o MLE Gaussiano, é possível notar que, em ocupações mais baixas, o MLE Lognormal apresenta uma eficiência maior na comparação, a não ser em 0%, onde nenhum empilhamento é acrescentado. Isso pode ser explicado pelo fato de que, apesar do pouco empilhamento, como tem-se uma distribuição de ruído menos esparsa, a componente não-linear que aparece acaba incluindo uma cauda positiva que afeta significativamente o ajuste à curva de uma distribuição Gaussiana (ver Figura 19, na Seção 4.2.2), acarretando em uma modelagem melhor com o uso da distribuição Lognormal. Conforme a ocupação aumenta, a distribuição do ruído se torna mais larga e o ajuste com ambas as distribuições é facilitado. Essas alterações também fazem com que a eficiência do MLE Lognormal comece a diminuir, já que a distribuição de ruído tende para uma Gaussiana, devido ao teorema do limite central.

No caso do OF2, mesmo na ocupação de 0% o MLE Lognormal se mostra mais interessante. A eficiência comparativa aumenta na condição de 10% e então começa a cair, também devido ao teorema do limite central. Ainda que essa diminuição ocorra, o MLE Lognormal segue superando o OF2 em cerca de 20%. Apesar de o OF2 ser um método

Tabela 4 - Eficiência do método MLE Lognormal para dados simulados, em porcentagem, quando comparado aos outros métodos, usando a medida de desvio padrão como base.

Parâmetro (%)	Ocupação	MLE Gauss	OF2	COF
E_{Ln}	0%	0,00	26,14	22,85
	10%	8,12	33,80	2,85
	20%	5,96	31,89	1,59
	30%	4,73	30,21	0,49
	40%	3,67	29,48	0,39
	50%	2,98	28,34	0,43
	60%	2,33	27,25	0,33
	70%	2,08	26,84	0,67
	80%	1,71	26,20	1,18
	90%	1,60	25,35	2,03
	100%	1,38	24,39	2,73

Fonte: A autora (2021).

linear, assim como o MLE Gaussiano, as restrições impostas ao cálculo dos coeficientes do filtro aumentam a variância do estimador, o que acaba afetando negativamente a eficiência do método. Isso explica seu desempenho inferior.

Por último, para o COF, apesar de se mostrar superior em toda a faixa de ocupação, o MLE Lognormal parece não levar tanta vantagem em ocupações intermediárias. Inicialmente sua eficiência é considerável, diminuindo até a condição de 60%, e então voltando a aumentar. A vantagem do COF sobre o MLE Gaussiano, mesmo que sejam ambos métodos lineares, é que o primeiro não é tão dependente do ajuste da distribuição, ao ruído, quanto o segundo. Assim, em ocupações mais baixas, onde o ajuste Gaussiano está mais prejudicado, o COF se apresenta como uma opção mais interessante.

O desempenho inferior do método MLE Lognormal, mesmo com uma modelagem do ruído aparentemente tão mais adequada, pode ser proveniente de problemas numéricos ou da forte correlação entre as variáveis do sinal. A primeira opção se deve à dificuldade de se trabalhar com PDFs multivariadas de 7 dimensões, o que fornece probabilidades muito pequenas e pode acabar afetando os resultados finais. Já a questão da correlação vem do fato de que as amostras de ruído não são independentes e a matriz de covariância não é capaz de lidar com estatísticas de ordem superior, aspectos que aparecem quando a característica linear é perdida.

4.3 **Dados Reais**

Nesta seção, de forma similar ao que foi feito na Seção 4.2, serão apresentados os resultados de modelagem do ruído e análise de eficiência dos métodos quando aplicados aos dados reais. Primeiramente, a Seção 4.3.1 mostra como o ruído real pode ser modelado, comparando as distribuições Gaussiana e Lognormal. Em seguida, a Seção 4.3.2 faz comparações entre os resultados obtidos de cada método, analisando suas eficiências e ponderando sobre o que se observa. Os dados desta seção são provenientes do TileCal e divididos em três condições: $\mu = 30$, $\mu = 50$ e $\mu = 90$, que se referem ao número médio interações por colisão. O pedestal foi retirado em uma fase inicial de tratamento dos dados e, para a análise de eficiência, foi adicionado um valor de 50 a este parâmetro.

Os dados experimentais de ruído são coletados diretamente do TileCal, que amostra os sinais após receber as informações sobre as colisões que foram detectadas pelo sistema do calorímetro. Dessa forma, apesar dos dados serem apenas de ruído, é importante estabelecer a possibilidade de que hajam dados de sinais de interesse imersos neles.

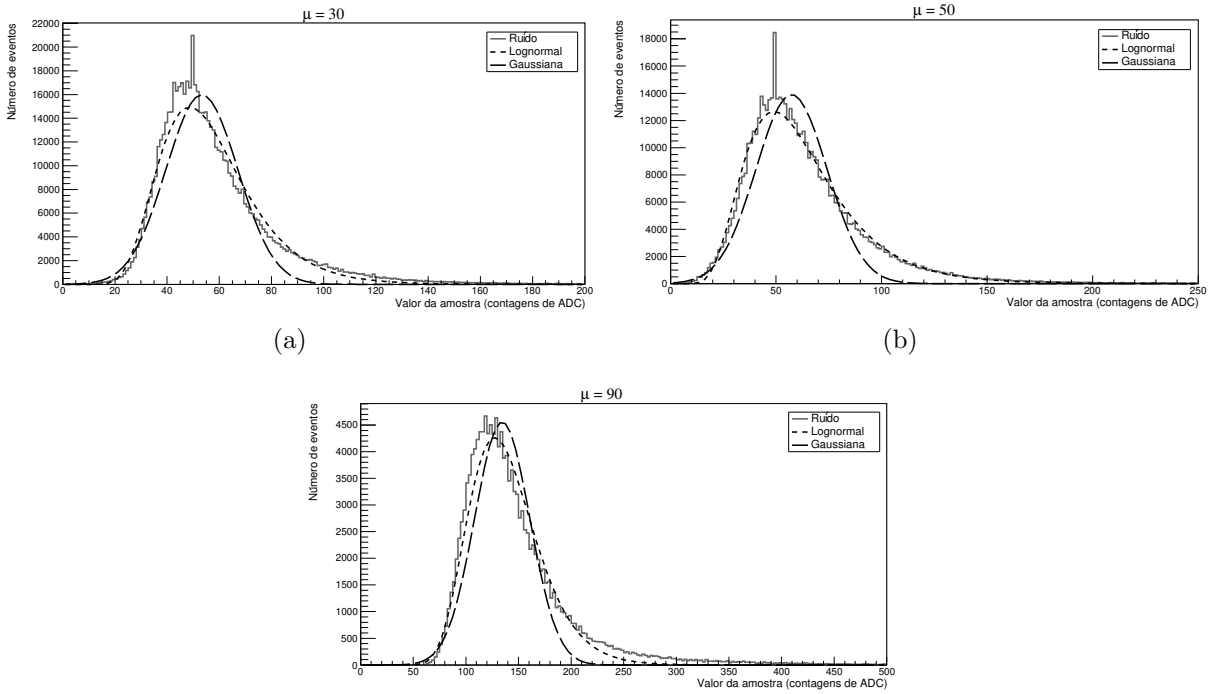
4.3.1 Modelagem do ruído

Assim como na análise dos dados simulados, no caso dos dados reais, a distribuição Lognormal foi utilizada para tentar modelar o ruído total de forma mais adequada. É importante citar que as três condições de luminosidade utilizadas nesta dissertação foram escolhidas por conterem uma quantidade maior de dados do que outras que também poderiam ter sido selecionadas. A Figura 26 mostra o ajuste das curvas Lognormal e Gaussiana aos dados reais de ruído para todas as três luminosidades, considerando a amostra 3. Mais uma vez, a binagem foi escolhida de forma que fosse utilizada aquela que retornasse um melhor ajuste, variando em uma faixa de 40 a 200 bins.

Analisando os histogramas, é possível notar que a cauda positiva, presente na distribuição e causada devido ao empilhamento, faz com que a distribuição Gaussiana encontre dificuldades para se ajustar aos dados. A Lognormal, por outro lado, consegue chegar mais próximo de uma modelagem eficaz, mesmo para $\mu = 90$, onde a Gaussiana parece se comportar melhor do que nas condições de empilhamento menos severo.

Para quantificar a qualidade do ajuste das curvas aos dados, a Figura 27 apresenta o gráfico dos valores de χ^2/ndf para as três luminosidades. Lembrando que, quanto mais próximo de 1 estiver este parâmetro, melhor a modelagem testada. Na Tabela 5 encontram-se os valores exatos para os pontos plotados na Figura 27. Nota-se que, para todas as luminosidades, a distribuição Lognormal fornece valores significativamente menores, chegando bem mais próximo da unidade. Este fato indica que o ajuste dos dados de ruído com esta função parece ser mais apropriado do que com a função Gaussiana. É

Figura 26 - Distribuições de ruído e ajuste Lognormal e Gaussiano para dados reais nas condições de empilhamento (luminosidade) (a) $\mu = 30$, (b) $\mu = 50$ e (c) $\mu = 90$.



Fonte: A autora (2021).

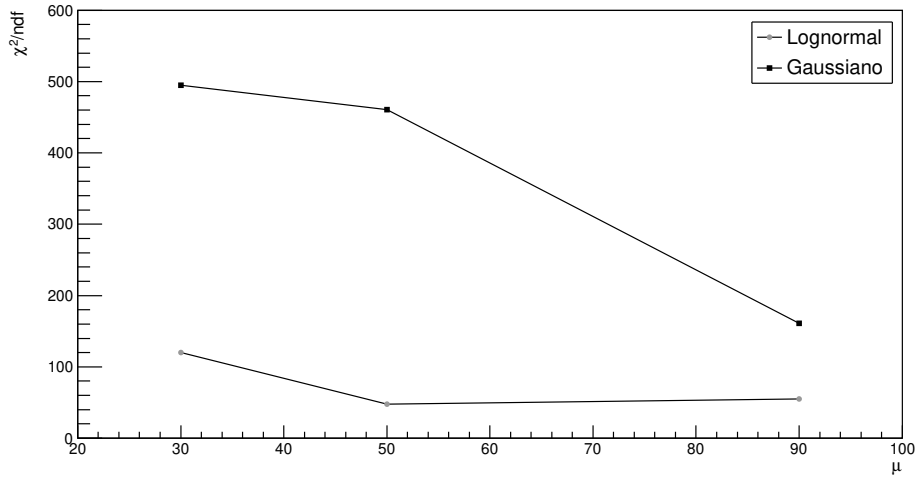
interessante observar que, quando a luminosidade chega a $\mu = 90$, a Gaussiana parece conseguir um desempenho superior àquele observado em valores mais baixos, apesar de ainda não alcançar a performance da Lognormal.

Na Figura 28 é mostrado o gráfico relativo à matriz de covariância do ruído para a condição de luminosidade quando $\mu = 50$. Também para os dados reais, é possível notar a grande correlação existente entre as amostras, o que ocorre principalmente devido à componente não-linear acrescentada pelo ruído de empilhamento. Quanto maior a presença de ruído não-Gaussiano, mais influência de estatísticas de ordem superior os dados têm, e a matriz de covariância acaba se tornando insuficiente para fazer a descorrelação. A equação da matriz que deu origem à Figura 28 pode ser encontrada no Apêndice B.

4.3.2 Análise de Eficiência

A análise de eficiência para os dados reais foi feita de forma similar àquela aplicada aos dados simulados, diferindo apenas por causa da não possibilidade de uma análise estatística com barra de erro, já que se trata de dados provenientes de experimentos verdadeiros. Após a divisão dos dados em conjunto de desenvolvimento e conjunto de

Figura 27 - Parâmetro χ^2/ndf para dados reais nas luminosidades $\mu = 30$, $\mu = 50$ e $\mu = 90$, comparando os ajustes com as distribuições Lognormal e Gaussiana.



Fonte: A autora (2021).

Tabela 5 - Valores de χ^2/ndf para todas as luminosidades.

	χ^2/ndf	
Luminosidade	Gaussiano	Lognormal
$\mu = 30$	494,886	120,317
$\mu = 50$	460,742	47,663
$\mu = 90$	161,068	54,832

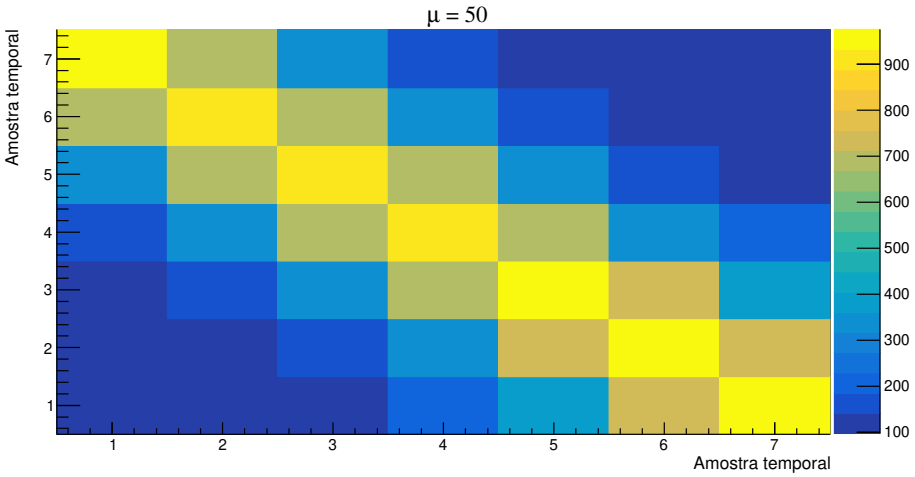
Fonte: A autora (2021).

teste e o cálculo dos parâmetros utilizados nos métodos, uma amplitude simulada foi gerada segundo uma distribuição exponencial de média igual à média calculada dos sinais de ruído. Os sinais com tais amplitudes foram embebidos nos dados reais e, então, os métodos foram empregues, com o objetivo de recuperar tais valores. Os erros, as médias e os valores de desvio padrão foram calculados de acordo com as definições dadas na Seção 4.1.

O erro é dado pela diferença entre a amplitude real e a estimada. A Figura 29 apresenta os histogramas das distribuições de erro de todos os métodos para as três luminosidades: $\mu = 30$, $\mu = 50$ e $\mu = 90$. Observando seus formatos, é possível notar que o MLE Lognormal apresenta um desvio padrão menor do que os outros três métodos, apesar de a média não seguir o mesmo padrão.

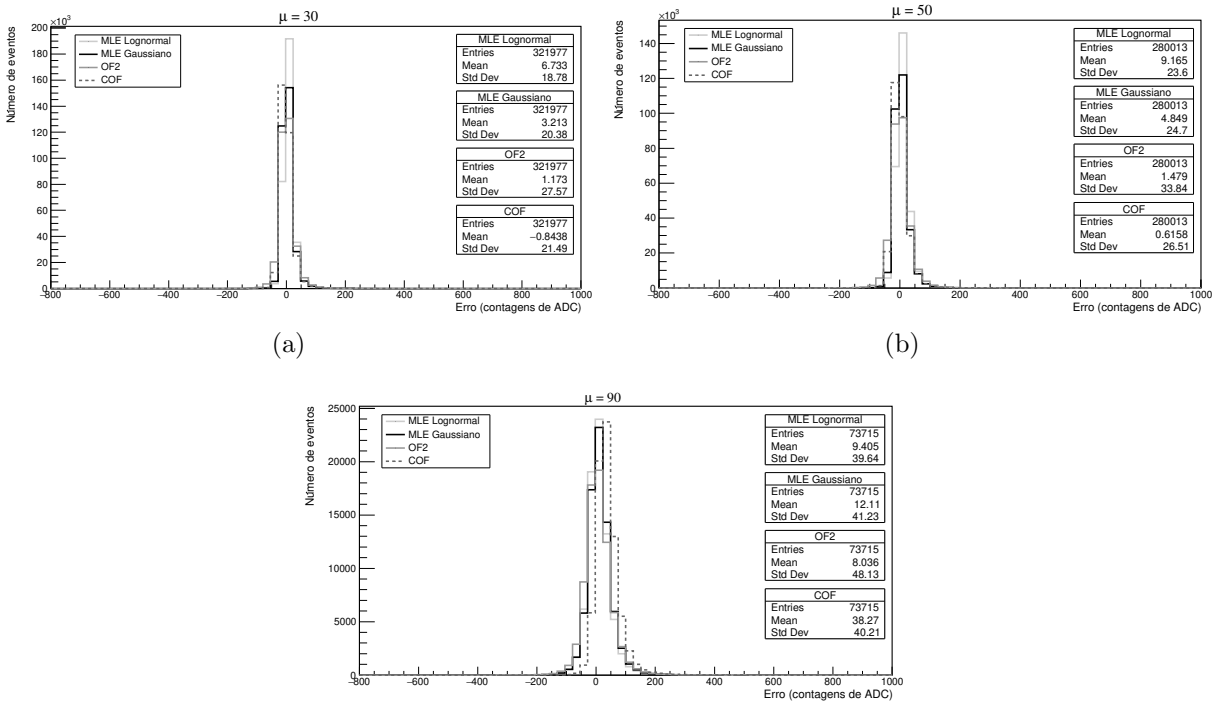
A Figura 30 mostra como o valor da média do erro de cada método se comporta com o aumento da luminosidade. Inicialmente, para as luminosidades onde $\mu = 30$ e $\mu = 50$,

Figura 28 - Gráfico relativo à matriz de covariância do ruído real para a luminosidade de $\mu = 50$.



Fonte: A autora (2021).

Figura 29 - Histogramas das distribuições dos erros para todos os métodos para dados reais nas luminosidades (a) $\mu = 30$, (b) $\mu = 50$ e (c) $\mu = 90$.



(a) (b)

(c)

Fonte: A autora (2021).

o OF2 apresenta números mais altos, seguido pelo MLE Gaussiano, MLE Lognormal e, finalmente, o COF. No entanto, quando chega-se a $\mu = 90$, há uma aparente mudança no

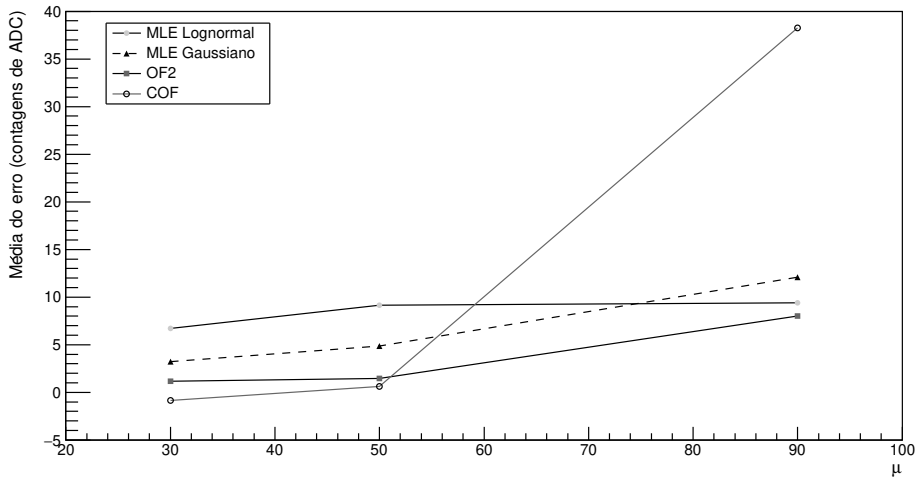
Tabela 6 - Média dos erros, em contagens de ADC, para todas as luminosidades.

Parâmetro	Luminosidade	MLE Logn	MLE Gauss	OF2	COF
Média	$\mu = 30$	6,733	3,213	1,173	-0,844
	$\mu = 50$	9,165	4,849	1,479	0,616
	$\mu = 90$	9,405	12,110	8,036	38,270

Fonte: A autora (2021).

desempenho dos métodos, quando se olha para este parâmetro. Neste caso, o menor valor é alcançado pelo OF2, aparecendo em segundo lugar o MLE Lognormal e em terceiro, o MLE Gaussiano, com o COF tendo um aumento significativo e se sobrepondo aos demais. A Tabela 6 mostra os valores das médias para todos os métodos e confirma essa análise. É importante lembrar que, para a utilização dos métodos baseados no MLE, a média pode ser parametrizada e subtraída, se necessário.

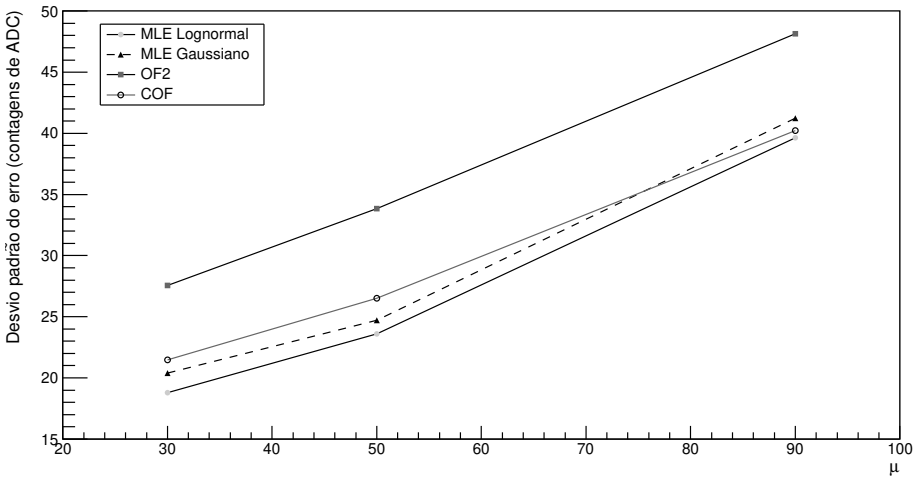
Figura 30 - Média dos erros para todas as luminosidades, comparando cada um dos métodos.



Fonte: A autora (2021).

Para o estudo do desvio padrão dos erros, o gráfico apresentado na Figura 31 foi plotado. Aqui, os métodos parecem se comportar de forma distinta àquela observada para as médias. Nas luminosidades menores ($\mu = 30$ e $\mu = 50$), o MLE Lognormal demonstra certa vantagem, com o MLE Gaussiano vindo em seguida e então o COF e o OF2 apresentando eficiências menores. Já na luminosidade de $\mu = 90$, o COF parece se comportar melhor, chegando a ultrapassar o MLE Gaussiano e quase alcançar o MLE Lognormal, que continua com o melhor desempenho. A Tabela 7 expõe os valores exatos para estas medidas e permite um melhor estudo das comparações.

Figura 31 - Desvio padrão dos erros para todas as luminosidades, comparando cada um dos métodos.



Fonte: A autora (2021).

Tabela 7 - Desvio padrão dos erros, em contagens de ADC, para todas as luminosidades.

Parâmetro	Luminosidade	MLE Logn	MLE Gauss	OF2	COF
Desvio padrão	$\mu = 30$	18,780	20,380	27,570	21,490
	$\mu = 50$	23,600	24,700	33,840	26,510
	$\mu = 90$	39,640	41,230	48,130	40,210

Fonte: A autora (2021).

Tendo como base a Equação (65), que utiliza a medida de desvio padrão, a eficiência do método MLE Lognormal, quando comparado aos outros, foi calculada. Os resultados são mostrados na Tabela 8. Nota-se que, quando a luminosidade aumenta, apesar de o MLE Lognormal se manter à frente dos outros, essa vantagem parece diminuir em todos os casos. Isto não surpreende, já que, para os dados simulados, já foram obtidos resultados semelhantes a este (ver Seção 4.2).

Assim como para os dados simulados, no entanto, os resultados, de certa forma, surpreendem, já que a distribuição Lognormal aparenta descrever o ruído com superioridade significativa, quando comparada à Gaussiana. Esta verificação não é refletida na análise de eficiência feita posteriormente. Mais uma vez, as explicações possíveis são a questão dos problemas numéricos apresentados por uma PDF multivariada de sete dimensões e também a dependência entre as variáveis aleatórias, com a qual a matriz de covariância já não é capaz de lidar de forma ótima em condições de não-Gaussianidade.

Tabela 8 - Eficiência do método MLE Lognormal para dados reais, em porcentagem, quando comparado aos outros métodos, usando a medida de desvio padrão como base.

Parâmetro (%)	Luminosidade	MLE Gauss	OF2	COF
E_{Ln}	$\mu = 30$	7,85	31,88	12,61
	$\mu = 50$	4,45	30,26	10,98
	$\mu = 90$	3,86	17,64	1,42

Fonte: A autora (2021).

4.4 **Análise de pré-processamento**

Com o intuito de investigar o quanto a dependência entre as variáveis aleatórias dos sinais de ruído podem estar influenciando nos resultados obtidos com o método MLE Lognormal, decidiu-se por calcular a Informação Mútua existente entre as amostras. Este cálculo foi feito através da discretização das variáveis (LIU et al., 2002), que eram, inicialmente, contínuas.

Primeiramente, os dados de cada amostra foram normalizados em valores contidos no intervalo fechado $[-1, 1]$, utilizando-se a Equação (66).

$$v_n = \frac{2(v - \min v)}{\max v - \min v} - 1, \tag{66}$$

onde v_n e v são, respectivamente, a variável normalizada e a original. Em seguida, o intervalo $[-1, 1]$ foi dividido em 100 subintervalos, ou seja, com um passo de $h = 0,02$. Os valores da variável normalizada foram então alocados nesses subintervalos e contabilizou-se a quantidade de ocorrências em cada um. Assim, foi possível obter-se os dados discretizados e a informação mútua pode ser calculada de acordo com a teoria para variáveis discretas.

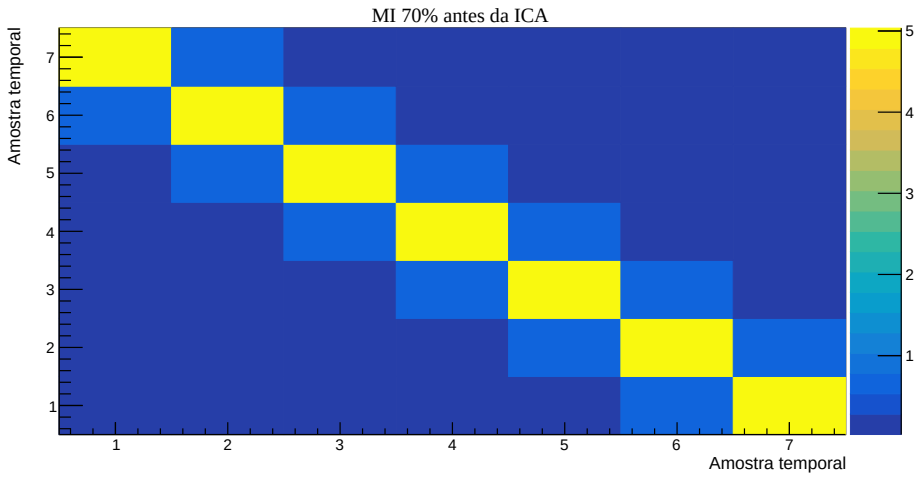
4.4.1 Dados simulados

Os resultados dos cálculos da informação mútua podem ser apresentados na forma matricial e, para os dados simulados, são ilustrados tendo como exemplo as matrizes representadas pelas Figuras 32 e 33. As linhas e colunas dessas matrizes representam as amostras do sinal e os elementos são a informação mútua correspondente. Por exemplo, a componente localizada na linha 2 e coluna 3 representa a quantidade de informação mútua entre as amostras 2 e 3. Por isso, nota-se que as matrizes são simétricas, já que a informação entre 2 e 3 é a mesma que entre 3 e 2, e que os maiores valores encontram-se

em suas diagonais principais, pois a correlação entre uma variável e ela mesma é total.

Para ilustrar como a dependência estatística está presente nas variáveis, a Figura 32 mostra o gráfico relativo à matriz de informação mútua do ruído original, sem nenhum pré-processamento, para a ocupação de 70%. É possível notar que as amostras mais próximas apresentam uma informação mútua maior, enquanto que com o distanciamento das mesmas esse fator vai diminuindo. A matriz que deu origem à Figura 32 se encontra no Apêndice B.

Figura 32 - Gráfico relativo à matriz de informação mútua do ruído simulado para a ocupação de 70%, antes do uso da ICA.



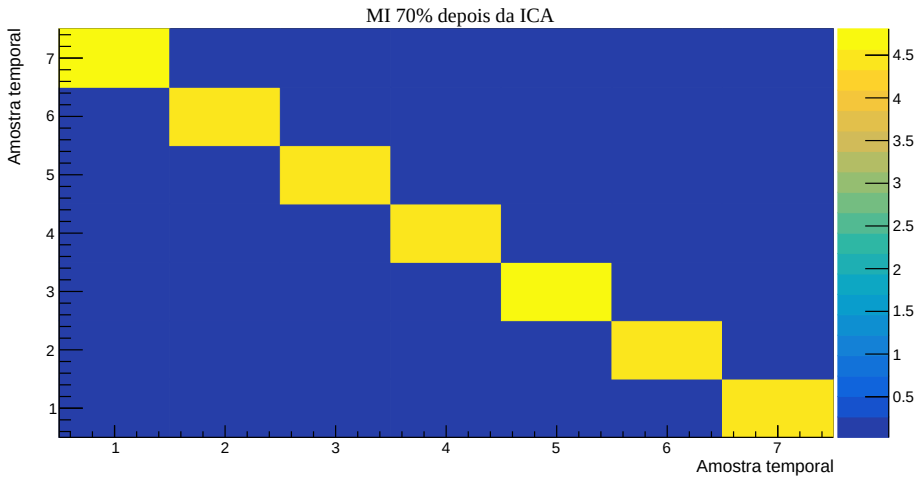
Fonte: A autora (2021).

Apenas estes dados, no entanto, não permitem uma quantificação mais precisa da informação mútua presente. Assim, visando tornar as variáveis estatisticamente independentes, um método foi utilizado, objetivando-se uma comparação. O recurso escolhido foi um algoritmo conhecido como FastICA (HYVÄRINEN; OJA, 2000). Este procedimento é baseado na técnica ICA, que é utilizada para buscar uma transformação tornando as componentes (variáveis) deste novo espaço o mais estatisticamente independentes possível. No entanto, embora a independência total nem sempre possa ser alcançada, a ICA pode ser testada a fim de diminuir a dependência existente. Vale ressaltar que o número de componentes independentes esperado é o tamanho da dimensão dos dados, ou seja, sete. Além disso, para a configuração do algoritmo FastICA, a abordagem baseada na deflação foi utilizada, onde os componentes são estimados um a um, de forma sequencial (HYVÄRINEN, 1999). A não-linearidade escolhida foi a função cúbica.

A Figura 33 mostra o gráfico referente à matriz relativa aos mesmos sinais que deram origem à Figura 32, mas agora após a aplicação da ICA. Ao comparar as duas matrizes, percebe-se uma diminuição significativa nos valores das componentes de informação mútua para amostras adjacentes, e uma redução menos acentuada para amostras que este-

jam a duas linhas (ou colunas) de distância. Para as amostras mais afastadas, no entanto, esse padrão não é observado, ocorrendo até um aumento em alguns elementos. Como a correlação é menor nestes casos, o algoritmo utilizado pode estar gerando apenas flutuações estatísticas, sem ter efetividade de fato, já que as variáveis já são descorrelacionadas de início.

Figura 33 - Gráfico relativo à matriz de informação mútua do ruído simulado para a ocupação de 70%, após o uso da ICA.



Fonte: A autora (2021).

Uma quantidade final para descrever a informação mútua em cada ocupação é útil para a análise. Este valor pode ser calculado através da matriz correspondente, somando-se os elementos da matriz triangular superior, sem a diagonal, e dividindo pela soma dos elementos da matriz triangular inferior, incluindo a diagonal:

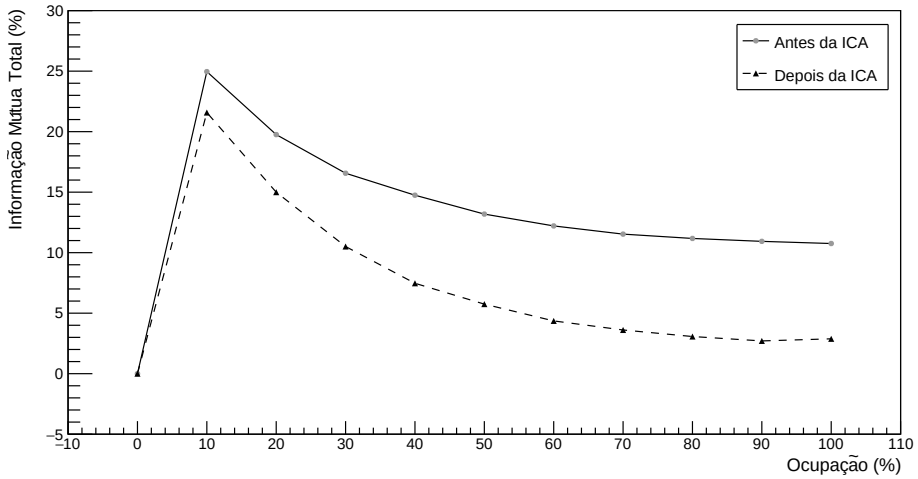
$$MI_{total} = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} MI(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{i} M(i,j)}, \tag{67}$$

onde N é o número de amostras, ou seja, $N = 7$.

Esse cálculo foi feito para cada uma das ocupações, antes e depois da aplicação da ICA, e o resultado obtido pode ser visto na Figura 34. Para a ocupação de 0% o algoritmo não tem êxito, pelo fato de os dados serem totalmente Gaussianos. Assim, a informação mútua é zero antes e depois da ICA. Para as demais condições, nota-se uma expressiva diferença nos valores de informação mútua total. O gráfico sugere que, quanto mais sinais empilhados, maior a descorrelação obtida e mais distantes estão os valores comparando-se o antes e depois. A diminuição da dependência conforme a ocupação aumenta pode ser

explicada pelo teorema do limite central, que faz os dados tenderem à linearidade.

Figura 34 - Informação mútua total para todas as ocupações, comparando dados de ruído antes e depois da aplicação da ICA.



Fonte: A autora (2021).

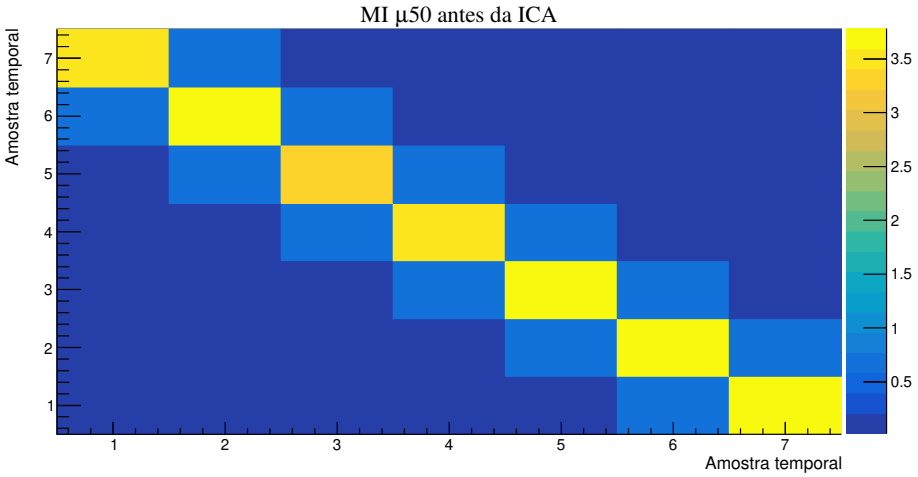
4.4.2 Dados reais

Para os dados reais, as Figuras 35 e 36 apresentam os gráficos relativos às matrizes de informação mútua do ruído na luminosidade de $\mu = 50$ antes e depois da aplicação da ICA, respectivamente. As matrizes referentes, em seu formato numérico, podem ser encontradas no Apêndice B.

Assim como para os dados simulados, a correlação parece ser alta em amostras adjacentes e um pouco menor quando há um salto de uma variável entre as amostras. Nestes casos, a descorrelação após a aplicação da ICA é evidente. Já em distâncias maiores, a dependência não se mostra tão significativa, e a ICA se torna menos efetiva. Em alguns elementos, os valores podem ser até maiores, devido provavelmente à flutuações estatísticas causadas pelo algoritmo.

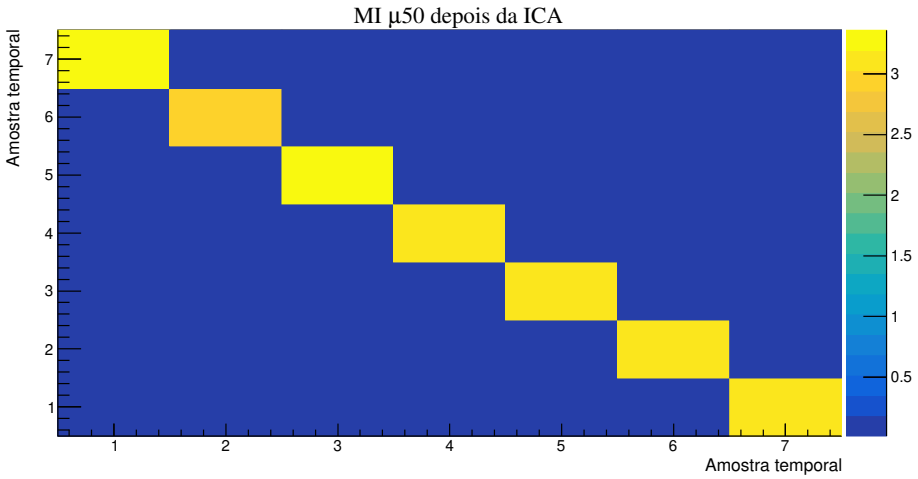
Para quantificar a correlação total entre as amostras, em cada luminosidade, a Equação (67) foi utilizada. A Figura 37 apresenta os resultados, mostrando o quão dependentes são as variáveis do sinal e como a ICA consegue trabalhar para torná-las tão independentes quanto possível. Analisando o gráfico, é possível observar que os números de correlação iniciais não apresentam grande variação quando a luminosidade é modificada. Quando a ICA é aplicada, eles caem consideravelmente e continuam não diferindo muito entre si. Vale destacar que, uma vez que as variáveis se tornam mais independen-

Figura 35 - Gráfico relativo à matriz de informação mútua do ruído real para a luminosidade $\mu = 50$, antes do uso da ICA.



Fonte: A autora (2021).

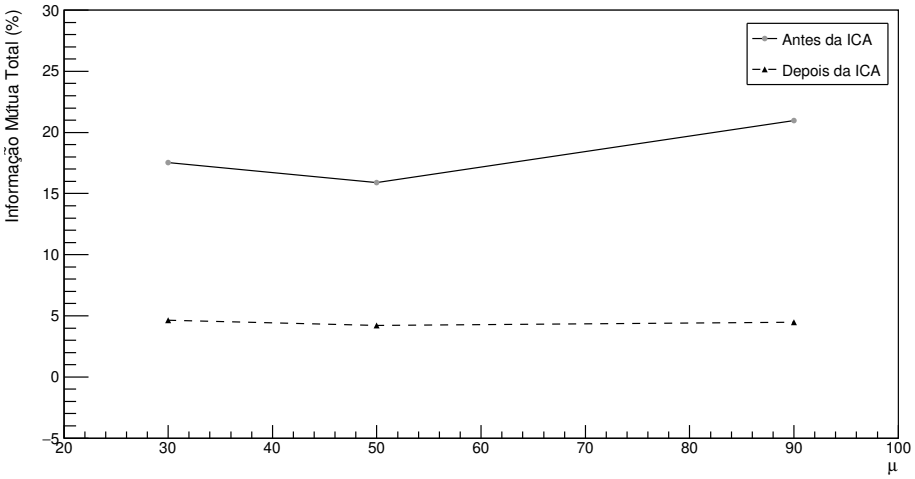
Figura 36 - Gráfico relativo à matriz de informação mútua do ruído real para a luminosidade $\mu = 50$, após o uso da ICA.



Fonte: A autora (2021).

tes, o cálculo da função densidade de probabilidade multivariada utilizada pelo método MLE pode ser aproximado pelo produtório das probabilidades das variáveis aleatórias individuais. Desta forma, o problema numérico para calcular o valor da PDF multivariada Lognormal é minimizado, além de otimizar a estimação da máxima verossimilhança para o parâmetro de interesse.

Figura 37 - Informação mútua total para todas as luminosidades, comparando dados reais de ruído antes e depois da aplicação da ICA.



Fonte: A autora (2021).

CONCLUSÃO

A estimação de parâmetros surge em diversos tipos de problemas, nas mais variadas áreas de estudo e pesquisa. No LHC, o maior e mais energético acelerador de partículas do mundo, o ATLAS é o maior experimento, tendo o TileCal como seu principal calorímetro hadrônico. No TileCal, a estimação da energia é feita através do processamento de sinais que são gerados a partir das colisões das quais partículas são observadas. A amplitude do sinal produzido pela amostragem da partícula é proporcional a energia da mesma. Assim, por meio da estimação da amplitude do pulso é possível dizer qual a natureza da partícula que o gerou.

O sinal fornecido pela eletrônica de leitura do TileCal é formado por sete amostras digitais, que são lidas em intervalos de 25 ns, formando uma janela de leitura total de 150 ns. As colisões no LHC ocorrem a cada 25 ns, criando, assim, a possibilidade de que sinais de colisões temporalmente adjacentes sejam observados numa mesma janela de leitura. Nem toda colisão produz uma partícula detectável por um mesmo sensor do calorímetro, mas com o aumento da luminosidade no LHC, essa probabilidade se torna cada vez maior. Esse empilhamento de sinais introduz ao ruído, inicialmente Gaussiano, uma componente não-linear, que diminui a eficiência de métodos lineares tipicamente utilizados, dificultando a estimação da amplitude do sinal.

Nesta dissertação, foi apresentada uma alternativa, um método baseado em uma abordagem não-linear, que tenta trabalhar melhor com o ruído adicional não-Gaussiano. Pode ser observado que, devido à assinatura exponencial que possuem os sinais do TileCal, o ruído de empilhamento apresenta características de uma distribuição Gama específica, a Erlang, a qual é formada a partir do somatório de exponenciais. Em razão da dificuldade de se trabalhar com a distribuição Gama multivariada, o que seria necessário, a princípio, já que o sinal é composto por sete variáveis aleatórias, optou-se pelo uso da Lognormal.

A partir dos estudos realizados, notou-se uma superioridade na descrição do ruído fornecida pela distribuição Lognormal, quando comparada à Gaussiana, tanto para os dados simulados quanto para os reais. Foi possível notar que a Lognormal se apresenta como uma boa alternativa para a distribuição Gama, nos casos aqui apresentados.

Quatro métodos foram comparados no presente estudo, sendo três lineares: OF2, COF e MLE Gaussiano. O primeiro é um método baseado em um filtro ótimo, extremamente eficiente para ruídos Gaussianos e atualmente utilizado no TileCal. O COF tem como base uma técnica de desconvolução, que é aplicada aos sinais recebidos. Já o MLE Gaussiano, utiliza um estimador de máxima verossimilhança com uma distribuição Gaussiana para estimar o parâmetro de interesse. O quarto método estudado foi o MLE Lognormal, que é não-linear e faz uso de uma distribuição Lognormal no estimador MLE.

Os resultados da análise de eficiência para os dados simulados, comparando todos

os métodos, mostraram que o MLE Lognormal apresenta resultados mais precisos do que os métodos lineares. Para o MLE Gaussiano e o OF2, essa melhora foi observada principalmente em ocupações intermediárias, um comportamento esperado, devido ao teorema do limite central, que prevê que o aumento da quantidade de distribuições exponenciais somando-se entre si fará esta soma tender para uma distribuição Gaussiana. A eficiência do MLE Lognormal com relação a esses dois métodos ficou em cerca de 3,14% e 28,17%, respectivamente, em média. Para o COF, a melhora foi observada nas ocupações iniciais e mais uma vez em ocupações muito altas, com cerca de 3,23%, em média, de eficiência. Mesmo em ocupações que apresentaram uma vantagem comparativa mais baixa do método não-linear, este continuou se mostrando superior a todos os lineares.

Já para os dados reais, a eficiência do método MLE Lognormal seguiu o mesmo padrão que para os dados simulados, quando comparado ao MLE Gaussiano e ao OF2, com 5,39% e 26,59% de melhora, em média, respectivamente. Para o COF, a maior luminosidade apresentou a menor eficiência comparativa do método não-linear. A média de melhoria, neste caso, foi de cerca de 8,34%. Para todos os casos, o MLE Lognormal se mostrou superior.

Com o intuito de verificar possíveis dificuldades que poderiam estar influenciando na eficiência do MLE Lognormal, optou-se pelo estudo da dependência estatística presente entre as variáveis aleatórias de ruído. Através do cálculo da informação mútua, mostrou-se que existe dependência entre as amostras, a qual varia de acordo com a densidade dos feixes de partículas. Apesar de a luminosidade parecer aumentar a dependência, não se pode dizer que quanto maior a luminosidade mais dependentes as variáveis serão. O que pode-se observar é que a informação mútua parece diminuir quando a distribuição está mais próxima de uma Gaussiana, ou seja, aumenta diretamente com a não-linearidade. A técnica da ICA foi aplicada e notou-se uma redução significativa na informação mútua entre as amostras, seguindo o mesmo padrão da informação mútua pré-ICA.

Para os dados simulados, a média da informação mútua total antes do uso da ICA, considerando todas as ocupações, ficou em cerca de 13,26%. Após o uso da ICA, esse valor foi para 7,00%. Já para os dados reais, considerando as três luminosidades estudadas, observou-se uma diminuição na informação mútua total de 18,13% para 4,44%, na média, antes e depois do uso da ICA, respectivamente.

A continuação deste trabalho prevê a análise de eficiência dos dados após a aplicação da ICA para tornar as variáveis aleatórias o mais independentes possível. A redução dimensional é cogitada, tendo em vista possíveis melhorias numéricas no cálculo das probabilidades. Além disso, métodos numéricos também devem ser estudados, de forma que a estimação da amplitude que maximiza a PDF Lognormal não dependa exclusivamente da busca exaustiva.

REFERÊNCIAS

AAD, G. et al. Readiness of the ATLAS Tile Calorimeter for LHC collisions. *European Physical Journal C*, v. 70, p. 1193 – 1236, 2010.

ALZAID, A.; SULTAN, K. Discriminating between gamma and lognormal distributions with applications. *Journal of King Saud University - Science*, v. 21, p. 99 – 108, 2009.

ANDERSON, T. W. *An Introduction to Multivariate Statistical Analysis*. 3. ed. Stanford, CA: Wiley-Interscience, 2003.

ANDRADE FILHO, L. M. de et al. Calorimeter Response Deconvolution for Energy Estimation in High-Luminosity Conditions. *IEEE Transactions on Nuclear Science*, v. 62, n. 6, p. 3265–3273, 2015.

ANJOS, A. *Online Filtering System Operating on a High Event Rate Environment*. Tese (Tese de Ph.D.) — Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro - Brasil, 2006.

ARNISON, G. et al. Experimental observation of isolated large transverse energy electrons with associated missing energy at s=540 GeV. *Physics Letters B*, v. 122, n. 1, p. 103 – 116, 1983. ISSN 0370-2693. Disponível em: ⟨http://www.sciencedirect.com/science/article/pii/0370269383911772⟩.

ATLAS. *Calorimeters*. 2020. Disponível em: ⟨https://atlas.cern/discover/detector/calorimeter⟩. Acesso em: 30 jul. 2020.

ATLAS. *Magnet System*. 2020. Disponível em: ⟨https://atlas.cern/discover/detector/magnet-system⟩. Acesso em: 28 jul. 2020.

BANNER, M. et al. Observation of single isolated electrons of high transverse momentum in events with missing transverse energy at the CERN pp collider. *Physics Letters B*, v. 122, n. 5, p. 476 – 485, 1983. ISSN 0370-2693.

BARBOSA, D. P. et al. Sparse Representation for Signal Reconstruction in Calorimeters Operating in High Luminosity. *IEEE Transactions on Nuclear Science*, v. 64, n. 7, p. 1942–1949, 2017.

BRÜNING, O.; ROSSI, L. The High-Luminosity Large Hadron Collider. *Nat Rev Phys 1*, p. 241–243, 2019.

CERN. *Overall view of LHC experiments.* 1998. Disponível em: ⟨https://cds.cern.ch/record/841555⟩. Acesso em: 21 jul. 2020.

CERN. *Computer Generated image of the ATLAS calorimeter*. 2008. Disponível em: ⟨https://cds.cern.ch/record/1095927/⟩. Acesso em: 31 jul. 2020.

CERN. *Aerial view of CERN site of Meyrin and Globe of Innovation.* 2012. Disponível em: ⟨https://cds.cern.ch/record/1476896/⟩. Acesso em: 20 jul. 2020.

CERN. *The CERN accelerator complex.* 2013. Disponível em: ⟨https://cds.cern.ch/record/1621894/⟩. Acesso em: 20 jul. 2020.

CERN. *How ATLAS detects particles: diagram of particle paths in the detector*. 2013. Disponível em: ⟨https://cds.cern.ch/record/1505342⟩. Acesso em: 31 jul. 2020.

CERN. 2020. Disponível em: ⟨https://home.cern/⟩. Acesso em: 20 jul. 2020.

CERN. 2020. Disponível em: ⟨https://home.cern/science/computing/birth-web⟩. Acesso em: 20 jul. 2020.

CHO, H.-K.; BOWMAN, K. P.; NORTH, G. R. A Comparison of Gamma and Lognormal Distributions for Characterizing Satellite Rain Rates from the Tropical Rainfall Measuring Mission. *Journal of Applied Meteorology*, v. 43, p. 1586 – 1597, 2004.

CLELAND, W.; STERN, E. Signal processing considerations for liquid ionization calorimeters in a high rate environment. *Nuclear Instruments and Methods in Physics Research A*, v. 338, p. 467 – 497, 1994.

CLEMENT, C.; KLIMEK, P. Identification of pile-up using the quality factor of pulse shapes in the ATLAS Tile Calorimeter. 2011.

COMON, P. Independent component analysis, A new concept? *Signal Processing*, v. 36, p. 287 – 314, 1994.

COVER, T. M.; THOMAS, J. A. *Elements of Information Theory*. 2. ed. Hoboken, NJ: Wiley, 2006.

EVANS, L.; BRYANT, P. LHC Machine. *Journal of Instrumentation*, v. 3, p. S08001, 2008.

FABJAN, C. Calorimetry in high-energy physics. *Experimental Techniques in High-Energy Nuclear and Particle Physics*, 1985. T. Ferbel, ed. (Plenum Pub. Corp., 1985).

FORBES, C. et al. *Statistical Distributions*. 4. ed. Hoboken, NJ: Wiley, 2011.

FRANCAVILLA, P. The ATLAS Tile Hadronic Calorimeter performance at the LHC. *Journal of Physics: Conference Series*, v. 404, p. 012007, 2012.

FULLANA, E. et al. Digital Signal Reconstruction in the ATLAS Hadronic Tile Calorimeter. *IEEE Transactions on Nuclear Science*, v. 53, n. 4, 2006.

GONÇALVES, G. I. et al. Desempenho de Algoritmos de Estimação de Energia para o Calorímetro de Telhas do Experimento ATLAS. In: *Congresso Brasileiro de Automática – Porto Alegre: CBA, 2020*. Porto Alegre, RS: [s.n.], 2020.

HERR, W.; MURATORI, B. Concept of luminosity. In: *CERN Accelerator School and DESY Zeuthen: Accelerator Physics*. [S.l.: s.n.], 2003. p. 361–377.

HOLLIK, W. Quantum field theory and the Standard Model. *Scientific Information Service Cern*, 2010. Disponível em: ⟨https://cds.cern.ch/record/1281946/files/p1.pdf⟩. Acesso em: 20 jul. 2020.

HUFFMAN, B. T. Plans for the phase II upgrade to the ATLAS detector. *Journal of Instrumentation*, IOP Publishing, v. 9, n. 02, p. C02033–C02033, 2014.

HYVÄRINEN, A. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, v. 10, n. 3, p. 626–634, 1999.

HYVÄRINEN, A.; OJA, E. Independent component analysis: algorithms and applications. *Neural Networks*, v. 13, p. 411 – 430, 2000.

KAY, S. M. *Fundamentals of Statistical Processing, Volume I: Estimation Theory*. New Jersey: Prentice-Hall Inc, 1993.

KHANDAI, P. K. et al. Hadron Spectra in p+p Collisions at RHIC and LHC Energies. *International Journal of Modern Physics A*, v. 28, n. 16, 2013.

KHRIPLOVICH, I. B.; LAMOREAUX, S. *CP Violation Without Strangeness - Electric Dipole Moments of Particles, Atoms, and Molecules*. New York, NY: Springer, 1997.

LIU, H. et al. Discretization: An enabling technique. *Data Mining and Knowledge Discovery*, v. 6, p. 393–423, 2002.

MARJANOVIĆ, M. ATLAS Tile Calorimeter Calibration and Monitoring Systems. *IEEE Transactions on Nuclear Science*, v. 66, n. 7, 2019.

MARSHALL, Z. Simulation of Pile-up in the ATLAS Experiment. *Journal of Physics: Conference Series*, v. 513, p. 022024, 2014.

NAKAHAMA, Y. The ATLAS Trigger System: Ready for Run-2. *Journal of Physics: Conference Series*, v. 664, p. 082037, 2015.

PALESTINI, S. The Muon Spectrometer of the ATLAS Experiment. *Nuclear Physics B - Proceedings Supplements*, v. 125, p. 337 – 345, 2003.

PAPOULIS, A.; PILLAI, S. U. *Probability, Random Variables, and Stochastic Processes*. 4. ed. New York, NY: McGraw-Hill, 2002.

PEEBLES, J. P. Z. *Probability, Random Variables, and Random Signal Principles*. 2. ed. US: McGraw-Hill, 1987.

PERALVA, B. S. The Tilecal Energy Reconstruction for Collision Data Using the Matched Filter. *IEEE Nuclear Science Symposium and Medical Imaging Conference*, p. 1 – 6, 2013.

RIGOLIN, G.; RIEZNIK, A. A. Introdução à criptografia quântica. *Revista Brasileira de Ensino de Física*, v. 120, n. 4, p. 517 – 526, 2005.

RIMES, S. de M. et al. Estimação da amplitude de sinais em calorimetria de altas energias em condições de alta ocupação de eventos no detector. In: *XXXVIII Simpósio Brasileiro de Telecomunicações e Processamento de Sinais – Florianópolis: SBrT, 2020*. Florianópolis, SC: [s.n.], 2020.

ROCCA, P. L.; RIGGI, F. The upgrade programme of the major experiments at the Large Hadron Collider. *Journal of Physics: Conference Series*, IOP Publishing, v. 515, p. 012012, 2014.

ROS, E. ATLAS Inner detector. *Nuclear Physics B - Proceedings Supplements*, v. 27, p. 235 – 238, 2003.

RUGGIERO, F. LHC accelerator R&D and upgrade scenarios. *The European Physical Journal C - Particles and Fields*, v. 34, p. 433 – 442, 2004.

SEIXAS, J. Quality Factor for the Hadronic Calorimeter in High Luminosity Conditions. *Journal of Physics: Conference Series*, v. 608, p. 012044, 2015.

TARMAST, G. Multivariate Log-normal Distribution. *ISI Proceedings: 53º Session Seoul*, 2001.

THE ALICE COLLABORATION. The ALICE Experiment at the CERN LHC. *Journal of Instrumentation*, v. 3, p. S08002, 2008.

THE ATLAS COLLABORATION. Liquid Argon Calorimeter. *Technical Design Report*, 1996. CERN/LHCC 96-41.

THE ATLAS COLLABORATION. Calorimeter Performance. *Technical Design Report*, 1997. CERN/LHCC/96-40.

THE ATLAS COLLABORATION. The ATLAS Experiment at the CERN Large Hadron Collider. *Journal of Instrumentation*, v. 3, p. S08003, 2008.

THE ATLAS COLLABORATION. Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV. *J. High Energ. Phys.*, 2016.

THE CMS COLLABORATION. The CMS Experiment at the CERN LHC. *Journal of Instrumentation*, v. 3, p. S08004, 2008.

THE LHCb COLLABORATION. The LHCb Detector at the LHC. *Journal of Instrumentation*, v. 3, p. S08005, 2008.

WIGMANS, R. *Calorimetry - Energy Measurement in Particle Physics*. [S.l.]: Oxford University Press, 2000.

ZHANG, H. The ATLAS Liquid Argon Calorimeter: Overview and Performance. *Journal of Physics: Conference Series*, v. 293, p. 012044, 2011.

APÊNDICE A – Trabalhos Submetidos e Apresentados em Eventos

Estimação da Amplitude Utilizando o Estimador de Máxima Verossimilhança em Condições de Empilhamento de Sinais

Autores: Sarita de Miranda Rimes, Lucas de Souza Gomes Nolla, Gabriel Cezar De Biase, Bernardo Sotto-Maior Peralva, Luciano Manhães de Andrade Filho, Augusto Santiago Cerqueira e José Manoel de Seixas.

Evento: XXII ENMC – Encontro Nacional de Modelagem Computacional e X ECTM – Encontro de Ciências e Tecnologia de Materiais, 2019.

Resumo: Sistemas de reconstrução de sinais se apoiam na estimação de parâmetros a partir de amostras temporais recebidas. Em calorimetria de altas energias, o sinal é conformado de modo que seu formato seja fixo, e a amplitude corresponda ao parâmetro a ser estimado. Tipicamente, métodos lineares são empregados, visto que o ruído eletrônico presente no sinal recebido pode ser modelado por uma função Gaussiana. Entretanto, na operação em alta taxa de eventos, como as presentes no LHC, no CERN, o sinal recebido é deformado pelo efeito de empilhamento de sinais, comprometendo a eficiência de métodos lineares. Desta forma, este trabalho avalia o uso de estimadores baseados na teoria do estimador de máxima verossimilhança (MLE) em que o ruído é descrito por funções multivariadas Gaussiana e Lognormal. Para comparar a eficiência, um conjunto de dados contendo diferentes condições de empilhamento de sinais foi gerado considerando o calorímetro de telhas (TileCal) do ATLAS no LHC. Os resultados mostram que os métodos baseados no MLE apresentam melhores eficiências quando comparados com o método atualmente utilizado no TileCal.

Estimação da Amplitude de Sinais em Calorimetria de Altas Energias em Condições de Alta Ocupação de Eventos no Detector

Autores: Sarita de Miranda Rimes, Bernardo Sotto-Maior Peralva, Luciano Manhães de Andrade Filho, Augusto Santiago Cerqueira e José Manoel de Seixas.

Evento: XXXVIII Simpósio Brasileiro de Telecomunicações e Processamento de Sinais - SBrT, 2020.

Resumo: Aplicações que têm como foco a estimação de parâmetros podem ser encontradas em diversas áreas do conhecimento. Na calorimetria de altas energias, por exemplo, a amplitude de um sinal conformado corresponde a energia da partícula absorvida. No entanto, em condições de alta luminosidade, o sinal de interesse surge deformado por um efeito de empilhamento de sinais, que possui características não-Gaussianas, degradando a eficiência de métodos tipicamente utilizados. Neste trabalho, é proposto um estimador projetado para um ruído que pode ser descrito por um modelo lognormal. Este é comparado com dois métodos lineares comumente utilizados. Sob diferentes condições de operação, os resultados mostram que o estimador projetado para o ruído lognormal apresenta uma melhor eficiência na estimação da amplitude do sinal adquirido pelo calorímetro considerado.

APÊNDICE B – Equações Auxiliares

Matriz de covariância para dados simulados na ocupação de 10%:

$$C_{10} = \begin{bmatrix} 3.042,9 & 2.140,5 & 848,2 & 239,1 & 52,6 & 1,4 & -7,7 \\ 2.140,5 & 2.976,0 & 2.111,8 & 850,2 & 232,1 & 38,4 & -8,4 \\ 848,2 & 2.111,8 & 2.980,2 & 2.122,5 & 842,3 & 227,4 & 32,2 \\ 239,1 & 850,2 & 2.122,5 & 3.014,3 & 2.170,4 & 877,0 & 233,5 \\ 52,6 & 232,1 & 842,3 & 2.170,4 & 3.084,0 & 2.164,3 & 859,0 \\ 1,4 & 38,4 & 227,4 & 877,0 & 2.164,3 & 3.000,5 & 2.151,6 \\ -7,7 & -8,4 & 32,2 & 233,5 & 859,0 & 2.151,6 & 3.055,6 \end{bmatrix} \tag{68}$$

Matriz de covariância para dados simulados na ocupação de 50%:

$$C_{50} = \begin{bmatrix} 11.823,6 & 8.324,5 & 3.361,4 & 993,2 & 273,7 & 113,0 & 76,5 \\ 8.324,5 & 11.787,2 & 8.401,6 & 3.399,1 & 943,9 & 221,6 & 98,5 \\ 3.361,4 & 8.401,6 & 11.805,5 & 8.313,9 & 3.298,4 & 914,6 & 246,1 \\ 993,2 & 3.399,1 & 8.313,9 & 11.684,9 & 8.343,0 & 3.349,1 & 954,0 \\ 273,7 & 943,9 & 3.298,4 & 8.343,0 & 11.828,2 & 8.381,2 & 3.382,7 \\ 113,0 & 221,6 & 914,6 & 3.349,1 & 8.381,2 & 11.893,4 & 8.548,4 \\ 76,5 & 98,5 & 246,1 & 954,0 & 3.382,7 & 8.548,4 & 12.152,1 \end{bmatrix} \tag{69}$$

Matriz de covariância para dados reais na luminosidade de $\mu = 50$:

$$C_{\mu 50} = \begin{bmatrix} 975,673 & 731,570 & 370,522 & 193,778 & 133,763 & 112,375 & 96,954 \\ 731,570 & 959,453 & 720,625 & 354,723 & 183,414 & 131,717 & 110,690 \\ 370,522 & 720,625 & 949,057 & 700,444 & 339,237 & 177,214 & 127,645 \\ 193,778 & 354,723 & 700,444 & 923,548 & 682,483 & 332,335 & 176,330 \\ 133,763 & 183,414 & 339,237 & 682,483 & 909,879 & 679,044 & 339,320 \\ 112,375 & 131,717 & 177,214 & 332,335 & 679,044 & 915,552 & 695,291 \\ 96,954 & 110,690 & 127,645 & 176,330 & 339,320 & 695,291 & 934,327 \end{bmatrix} \tag{70}$$

Matriz de informação mútua para dados simulados na ocupação de 70%, antes do uso da

ICA:

$$MI_{a70} = \begin{bmatrix} 4,95462 & 0,58924 & 0,12220 & 0,04230 & 0,03071 & 0,02943 & 0,02812 \\ 0,58924 & 4,91510 & 0,59491 & 0,12255 & 0,04037 & 0,03004 & 0,02776 \\ 0,12220 & 0,59491 & 5,03516 & 0,59293 & 0,12473 & 0,04297 & 0,03212 \\ 0,04230 & 0,12255 & 0,59293 & 4,95309 & 0,59676 & 0,12223 & 0,04114 \\ 0,03071 & 0,04037 & 0,12473 & 0,59676 & 4,89309 & 0,59073 & 0,11902 \\ 0,02943 & 0,03004 & 0,04297 & 0,12223 & 0,59073 & 4,92771 & 0,58883 \\ 0,02812 & 0,02776 & 0,03212 & 0,04114 & 0,11902 & 0,58883 & 4,91150 \end{bmatrix} \quad (71)$$

Matriz de informação mútua para dados simulados na ocupação de 70%, após o uso da ICA:

$$MI_{d70} = \begin{bmatrix} 4,54649 & 0,09962 & 0,04036 & 0,10812 & 0,05099 & 0,05439 & 0,03764 \\ 0,09962 & 4,49256 & 0,04946 & 0,05645 & 0,08841 & 0,04021 & 0,02747 \\ 0,04036 & 0,04946 & 4,80792 & 0,03277 & 0,09738 & 0,02786 & 0,02583 \\ 0,10812 & 0,05645 & 0,03277 & 4,49015 & 0,03868 & 0,10350 & 0,05057 \\ 0,05099 & 0,08841 & 0,09738 & 0,03868 & 4,49224 & 0,02872 & 0,02452 \\ 0,05439 & 0,04021 & 0,02786 & 0,10350 & 0,02872 & 4,49902 & 0,11225 \\ 0,03764 & 0,02747 & 0,02583 & 0,05057 & 0,02452 & 0,11225 & 4,57429 \end{bmatrix} \quad (72)$$

Matriz de informação mútua para dados reais na luminosidade $\mu = 50$, antes do uso da ICA:

$$MI_{a50} = \begin{bmatrix} 3,65859 & 0,69400 & 0,17097 & 0,04935 & 0,02483 & 0,02078 & 0,01887 \\ 0,69400 & 3,68515 & 0,65417 & 0,13550 & 0,03830 & 0,02402 & 0,02014 \\ 0,17097 & 0,65417 & 3,61783 & 0,60481 & 0,11834 & 0,03672 & 0,02279 \\ 0,04935 & 0,13550 & 0,60481 & 3,43629 & 0,58503 & 0,11786 & 0,03650 \\ 0,02483 & 0,03830 & 0,11834 & 0,58503 & 3,40140 & 0,60070 & 0,12858 \\ 0,02078 & 0,02402 & 0,03672 & 0,11786 & 0,60070 & 3,78134 & 0,64045 \\ 0,01887 & 0,02014 & 0,02279 & 0,03650 & 0,12858 & 0,64045 & 3,52156 \end{bmatrix} \quad (73)$$

Matriz de informação mútua para dados reais na luminosidade $\mu = 50$, após o uso da ICA:

$$MI_{d50} = \begin{bmatrix} 3,09350 & 0,02801 & 0,06809 & 0,07537 & 0,02725 & 0,02652 & 0,05088 \\ 0,02801 & 3,03069 & 0,03020 & 0,04606 & 0,01711 & 0,05768 & 0,02614 \\ 0,06809 & 0,03020 & 3,08956 & 0,04858 & 0,05020 & 0,02047 & 0,06878 \\ 0,07537 & 0,04606 & 0,04858 & 3,07504 & 0,03055 & 0,04489 & 0,10401 \\ 0,02725 & 0,01711 & 0,05020 & 0,03055 & 3,35998 & 0,01693 & 0,04944 \\ 0,02652 & 0,05768 & 0,02047 & 0,04489 & 0,01693 & 2,91812 & 0,06768 \\ 0,05088 & 0,02614 & 0,06878 & 0,10401 & 0,04944 & 0,06768 & 3,22081 \end{bmatrix} \quad (74)$$