



## QUALIDADE DE DADOS BASEADA EM PROBABILIDADES PARA RECONSTRUÇÃO DA ENERGIA EM CALORIMETRIA DE ALTAS ENERGIAS

**Sarita de M. Rimes**<sup>1</sup> - smrimes@iprj.uerj.br

**Bernardo S. Peralva**<sup>1</sup> - bernardo.peralva@uerj.br

**Luciano M. de Andrade Filho**<sup>2</sup> - luciano.andrade@ufjf.edu.br

**José M. de Seixas**<sup>3</sup> - seixas@lps.ufrj.br

<sup>1</sup>Universidade do Estado do Rio de Janeiro, Instituto Politécnico - Nova Friburgo, RJ, Brasil

<sup>2</sup>Universidade Federal de Juiz de Fora - Juiz de Fora, MG, Brasil

<sup>3</sup>Universidade Federal do Rio de Janeiro, COPPE/Poli - Rio de Janeiro, RJ, Brasil

**Resumo.** *Em diversas aplicações de engenharia, a qualidade da reconstrução da informação de interesse interfere diretamente no sucesso da tomada de decisão. Em experimentos modernos de física de altas energias, como o ATLAS, do LHC, a reconstrução dos eventos é essencial para a observação precisa de partículas e descobertas importantes para a ciência. Neste contexto, os sinais provenientes do sistema de calorimetria fornecem a informação sobre a energia das partículas produzidas nas colisões. No calorímetro de telhas (TileCal) do ATLAS, a qualidade da reconstrução da energia é medida utilizando o método do  $\chi^2$ . Entretanto, tal abordagem não é capaz de operar em condições de empilhamento de sinais, conseguindo somente detectar a presença do efeito do empilhamento no sinal recebido. Portanto, neste trabalho, uma nova abordagem para avaliar a qualidade da reconstrução da energia em condições de empilhamento de sinais é apresentada. O método proposto se apoia no cálculo de probabilidades e é independente do método de reconstrução empregado. Utilizando dados simulados do TileCal em diferentes condições de empilhamento de sinais, os resultados mostram que o uso do método proposto aumenta a eficiência de seleção de sinais.*

**Palavras-chave:** *Qualidade de dados, filtro ótimo, empilhamento de sinais, calorimetria de altas energias.*

### 1. INTRODUÇÃO

A reconstrução de sinais está presente em diversas aplicações e soluções de problemas na ciência e nas engenharias (Mahfoudia, Horlin & Neyt; Mitra, 2001). Usualmente, tem-se fenômenos que podem ser analisados através de sinais que são produzidos por experimentos específicos, os quais precisam ser reconstruídos e, muitas das vezes, se apresentam imersos em ruído, dificultando uma reconstrução eficiente. Áreas que trabalham com esse tipo de

problema e podem ser citadas como exemplo são a de comunicações, radares, sonares e análise de imagens (Kay, 1993).

Na física experimental de altas energias, experimentos são construídos visando acelerar partículas, em velocidades próximas a da luz, e colidi-las (Hauptman, 2010). Neste contexto, os calorímetros possuem papel fundamental na coleta e fornecimento de dados relativos à energia de tais partículas (Wigmans, 2017). Essa área trabalha com sistemas que operam com altas taxas de eventos, produzindo uma grande quantidade de informações. Calorímetros de altas energias absorvem e amostram a energia das partículas que interagem com seu material, produzindo sinais que se relacionam a essa energia (The ATLAS Collaboration, 2008).

Contudo, devido à alta quantidade de dados gerada, frequentemente, esses sinais se apresentam corrompidos por ruído e se torna necessária a utilização de métodos que auxiliem em suas reconstruções. Esse é o caso do TileCal (The ATLAS Collaboration, 2010; Peralva, 2022), principal calorímetro hadrônico do experimento ATLAS (The ATLAS Collaboration, 2008) (do inglês, *A Toroidal LHC ApparatuS*), o maior experimento do LHC (Evans & Bryant, 2008) (do inglês, *Large Hadron Collider*), que fica no CERN (CERN, 2023) (do francês, *Organisation Européenne pour la Recherche Nucléaire*). Diversas descobertas e avanços tecnológicos já foram observados nos experimentos do LHC, tal como a descoberta do bóson de Higgs (The ATLAS Collaboration, 2012). No LHC, as colisões ocorrem a uma frequência de 40 MHz com crescente aumento da sua luminosidade (Herr & Muratori, 2003). O aumento da luminosidade é inevitável e essencial para os experimentos buscarem por física nova, e acontecerá de forma ainda mais acentuada nos próximos anos de operação (Bruning & Rossi, 2015). Por outro lado, a consequência desse aumento é o crescimento da quantidade de informações recebida pelo sistema de calorimetria, fazendo com que os sinais amostrados estejam imersos em condições de empilhamento de sinais cada vez mais severas. O efeito de empilhamento de sinais introduz desafios para a área de processamento de sinais e vários estudos buscam mitigá-lo na etapa de estimação da energia (Gonçalves, 2022; de Andrade Filho, 2015; Barbosa, 2017).

Embora os novos métodos de estimação consigam minimizar o problema do empilhamento de sinais, faz-se necessário uma métrica que represente a qualidade da reconstrução do sinal, visando descartar sinais mal reconstruídos ou corrompidos. Para tal, atualmente, o TileCal faz uso do método do  $\chi^2$  (Valero, 2012; Papoulis & Pillai, 2002), que avalia a qualidade da reconstrução dos sinais de interesse. No entanto, este método não prevê o efeito não-linear do empilhamento e, portanto, não é capaz de discriminar sinais com altos erros de estimação ou corrompidos (Seixas, 2015). Vale ressaltar que o empilhamento diminui também a eficiência dos métodos lineares tipicamente utilizados na reconstrução de sinais (Rimes, 2021).

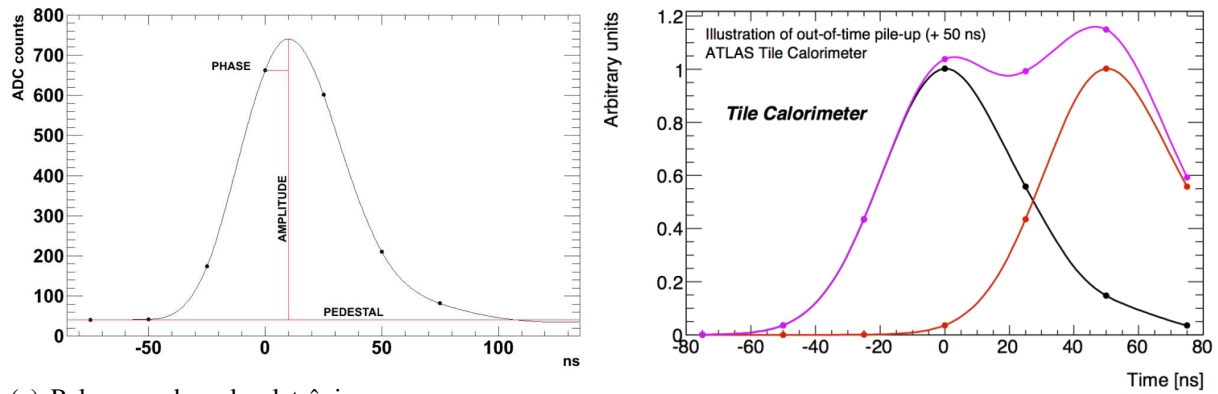
Sendo assim, este trabalho propõe um método de qualidade de dados para o TileCal que seja capaz de descartar sinais mal reconstruídos ou corrompidos, visando melhorar a eficiência dos métodos de reconstrução da energia, independente da abordagem utilizada. Como os algoritmos tipicamente empregados para reconstrução da energia assumem que o sinal é corrompido por ruído gaussiano, o método proposto se baseia em uma probabilidade multivariada gaussiana do ruído.

Na Seção 2., será apresentado o ambiente de trabalho utilizado como base para o estudo feito neste artigo. Na Seção 3., são explorados os métodos utilizados para a reconstrução dos sinais do TileCal, enquanto a seção 4. apresenta a estratégia proposta para qualidade de dados na reconstrução da energia. A Seção 5. traz os resultados obtidos tanto de análises de eficiência dos métodos quanto do estudo da qualidade dos dados. Finalmente, a Seção 6. aborda as discussões finais sobre o trabalho e perspectivas futuras.

## 2. RECONSTRUÇÃO DO SINAL EM CALORIMETRIA

Na calorimetria de altas energias, o sinal é produzido pela eletrônica de leitura do calorímetro. O experimento ATLAS do LHC tem propósito geral e visa cobrir um amplo programa de física. Este experimento possui um sistema de calorimetria composto por dois tipos de calorímetros: eletromagnético e hadrônico. O calorímetro eletromagnético absorve energia das partículas que interagem de forma eletromagnética (elétrons e fótons), enquanto o calorímetro hadrônico absorve energia de partículas que interagem através da interação forte (principalmente hádrons neutros). O TileCal é o principal calorímetro hadrônico do ATLAS e é utilizado como base para as análises deste trabalho.

No TileCal, quando ocorre uma colisão no LHC, as partículas provenientes são absorvidas por placas de aço e suas energias são amostradas em forma de luz por telhas cintilantes (Aad, 2010). O resultado dessa amostragem se dá em forma de sinais, os quais a eletrônica do TileCal amostra a uma frequência de 40 MHz. O tempo de digitalização é ajustado de forma que a amostra central corresponda ao pico do sinal e esteja no tempo  $t = 0$  ns. A janela completa possui 7 amostras e um total de 150 ns. O modelo do pulso do TileCal pode ser visto na Fig. 1a.



(a) Pulso gerado pela eletrônica do TileCal (Peralva, 2013).

(b) Empilhamento de sinais no TileCal (Seixas, 2015).

A reconstrução da energia é feita com base na amplitude do sinal, que é definida como a magnitude do pico do sinal com relação a sua base, como pode ser visto ainda na Fig. 1a. Como a amplitude é proporcional à energia da partícula, a reconstrução se torna possível a partir da sua estimação.

Um parâmetro, ou fator de qualidade, que pode ser usado para avaliar a qualidade da reconstrução dos sinais é o  $\chi^2$ , o qual é calculado de acordo com a Eq. (1) (Valero, 2012).

$$\chi^2 = \sqrt{\sum_{i=1}^N \left( r_i - \hat{a}_0 h_{0,i} - \sum_{j=1}^N \hat{a}_j h_{j,i} \right)^2}, \quad (1)$$

onde  $r$  são as amostras do sinal recebido,  $N$  e  $\hat{a}$  são o número de amostras e a amplitude do sinal, respectivamente, e  $h$  é uma matriz formada pelos sinais de interesse reconstruídos. Quanto mais próximo da unidade estiver o valor de  $\chi^2$ , melhor a qualidade da reconstrução.

Atualmente, no TileCal, o método utilizado para reconstrução dos sinais é chamado de OF (Fullana, 2006), uma versão de Filtro Ótimo que é linear e funciona minimizando a variância do ruído e determinando o parâmetro de interesse através de uma combinação linear das

amostras do sinal. Esse método utiliza o  $\chi^2$  como fator para avaliar a qualidade da reconstrução do sinal de interesse (Clement & Klimek, 2011). Outro método que vem sendo usado para a reconstrução dos sinais no TileCal é o COF (do inglês, *Constrained Optimal Filter*) (de Andrade Filho, 2015), que também é linear e se baseia na desconvolução dos sinais digitais recebidos. O  $\chi^2$  também pode ser utilizado para este método (Gonçalves, 2022).

Com o aumento da luminosidade, tem-se um crescimento do empilhamento nos sinais. Isso se deve ao fato de que, quanto mais denso forem os feixes acelerados, mais interações entre partículas (o LHC acelera prótons) ocorrerão a cada colisão e, portanto, mais partículas serão produzidas e interagirão com as células do calorímetro. Visto que a resposta da eletrônica do TileCal tem duração de aproximadamente 150 ns, e o LHC colide prótons a cada 25 ns, múltiplas colisões podem ocorrer dentro da mesma janela de leitura, resultando no fenômeno do empilhamento de sinais. A Fig. 1b ilustra esse fenômeno. O aumento do empilhamento introduz desafios aos métodos de reconstrução de energia e diminui a eficiência do método  $\chi^2$ , tornando necessário o desenvolvimento de novas abordagens.

### 3. MÉTODOS DE RECONSTRUÇÃO DA ENERGIA

Os métodos utilizados neste trabalho serão apresentados de forma resumida nesta seção.

#### 3.1 OF

O OF é um algoritmo que se baseia na minimização da variância do ruído para reconstruir a amplitude (energia) dos sinais. Neste método, as amostras recebidas são dadas pela Eq. (2).

$$r_k = ped + As_k + A\tau\dot{s}_k + n_k, \quad i = 1, \dots, N, \quad (2)$$

onde  $A$  é a amplitude verdadeira do sinal,  $s$  é o valor do pulso de referência e  $\dot{s}$  é a sua derivada,  $\tau$  é a fase do sinal,  $ped$  é o pedestal e  $N$  é o número de amostras (no caso do TileCal,  $N = 7$ ).

A amplitude do sinal, que permite sua reconstrução, pode ser estimada através de uma combinação linear entre as amostras recebidas e os coeficientes do filtro, vista na Eq. (3), aplicando-se ao processo algumas restrições, detalhadas por Fullana (2006).

$$\hat{A}_{OF} = \sum_{k=0}^{N-1} w_i r_i, \quad (3)$$

#### 3.2 COF

O COF é um método linear que aplica uma abordagem de desconvolução nos sinais recebidos, estimando a amplitude do sinal de interesse e as dos sinais possivelmente empilhados numa dada janela de leitura.

Para tal, as amostras do sinal recebido são descritas como um sinal conformado, que contém contribuições de sinais provenientes de colisões adjacentes, somado ao ruído eletrônico:

$$r_i = \sum_k (s_k a_{i-k}) + n_i, \quad (4)$$

onde  $i$  é a amostra de tempo de um sistema linear e invariante no tempo (LTI) e  $s_i$  é sua resposta de impulso (que corresponde ao pulso de referência),  $a$  é um sinal impulso que representa as energias (amplitudes) depositadas e  $n$  é o ruído eletrônico do sistema.

Aqui, a estimação das amplitudes dos sinais presente na janela de leitura é dada pela Eq. (5), onde  $\mathbf{S}$  é uma matriz  $N \times N$  com versões defasadas do pulso de referência.

$$\hat{\mathbf{a}} = \mathbf{S}^{-1}\mathbf{r}. \quad (5)$$

### 3.3 MLE

O MLE (do inglês, *Maximum Likelihood Estimation*) (Kay, 1993) é um método que consiste na estimação do parâmetro de interesse que maximize a função de verossimilhança escolhida. O procedimento é descrito na Eq. (6), sendo  $L(\mathbf{X}|\theta)$  a função de verossimilhança e  $\theta$  o parâmetro a se determinar quando  $\mathbf{x}$  é igual a  $\mathbf{X}$ .

$$L(\mathbf{X}|\theta) = P(\theta|\mathbf{x} = \mathbf{X}). \quad (6)$$

Neste estudo, a função de verossimilhança escolhida foi a PDF (do inglês, *Probability Density Function*). A maximização foi feita igualando-se a derivada a zero:

$$\frac{\partial p(\mathbf{x}|\theta)}{\partial \theta} = 0, \quad (7)$$

onde  $p$  é a PDF,  $\mathbf{x}$  é o vetor que descreve o processo modelado e  $\theta$  é o parâmetro a ser estimado. O objetivo é encontrar o valor de  $\theta$  que faça a derivada se aproximar o máximo possível de zero.

**MLE Gaussiano** No MLE Gaussiano, como o nome sugere, utiliza-se a PDF Gaussiana (Anderson, 2003) como função de verossimilhança. Através da derivada da PDF Gaussiana e igualando esta a zero, chega-se a uma equação fechada para a amplitude:

$$A = \frac{\mathbf{r}^T \mathbf{C}^{-1} \mathbf{s}}{\mathbf{s}^T \mathbf{C}^{-1} \mathbf{s}}. \quad (8)$$

O procedimento completo pode ser observado em detalhes em Rimes (2021).

## 4. MÉTODO PROPOSTO PARA QUALIDADE DE DADOS

Os métodos de estimação de energia aproximam o ruído através de uma distribuição Gaussiana multivariada. Portanto, se a reconstrução da amplitude do sinal for correta, o sinal reconstruído pode ser subtraído do sinal recebido e a informação restante compreenderá somente o ruído multivariado Gaussiano. Neste caso, é esperado que a PDF Gaussiana,

$$p(\mathbf{n}) = \frac{1}{|\mathbf{C}|^{\frac{1}{2}} (2\pi)^{\frac{N}{2}}} \exp \left[ -\frac{(\mathbf{r} - \mathbf{A}\mathbf{s})^T \mathbf{C}^{-1} (\mathbf{r} - \mathbf{A}\mathbf{s})}{2} \right], \quad (9)$$

resulte em um valor alto de probabilidade. Na Eq. (9),  $\mathbf{C}$  é a matriz de covariância do ruído. Neste caso, o ruído corresponde a  $\mathbf{n} = \mathbf{r} - \mathbf{A}\mathbf{s}$ , e as variáveis  $\mathbf{r}$ ,  $\mathbf{s}$  e  $A$  são as amostras recebidas, o pulso de referência e a amplitude que se deseja estimar, respectivamente. Vale destacar que a matriz  $\mathbf{C}$  é estimada por um conjunto de dados específico contendo somente ruído e, portanto, ela é dependente do nível de empilhamento de sinais.

Independentemente do método de reconstrução da energia empregado, a proposta é utilizar o valor da probabilidade para avaliar a qualidade da reconstrução do sinal de interesse. Assim, a amplitude estimada pelo método de reconstrução da energia é utilizada na PDF mostrada na Eq. (9) e o valor da probabilidade resultante é usado como métrica para avaliação da qualidade da reconstrução do sinal. Se a amplitude for bem estimada, o ruído  $\mathbf{n}$  se aproximará da PDF.

## 5. RESULTADOS

Esta seção detalha como os dados para simulação foram gerados e descreve a análise de eficiência empregada neste trabalho.

### 5.1 Dados simulados

Os dados de ruído foram simulados utilizando-se um simulador (Gonçalves, 2023) desenvolvido especialmente para as análises em calorimetria no LHC. Neste ambiente do simulador, é possível definir os parâmetros que descrevem o sinal, assim como a ocupação desejada. A ocupação é uma forma de reproduzir o conceito de luminosidade e é dada em termos de porcentagem. Quanto maior a ocupação, maior a probabilidade da célula do calorímetro receber um sinal quando ocorre uma colisão, ou seja, ocupações maiores aumentam a chance de empilhamento. Para este trabalho, foram escolhidas quatro ocupações: 10%, 30%, 50% e 80%, de forma a analisar a eficiência da reconstrução em diferentes cenários de empilhamento.

Para simular o ruído de empilhamento, foi utilizada uma distribuição exponencial com média 50 ADC, enquanto a fase e o pedestal do sinal foram de  $[-10, 10]$  ns e 50 ADC, respectivamente. Já o ruído eletrônico foi descrito por uma distribuição Gaussiana de média 0 ADC e desvio padrão de 1,5 ADC, valores típicos de experimentos modernos (Chapman, 2011; Banerjee, 2012). Foram simulados um total de 100.000 sinais no conjunto de ruído, para cada ocupação. Este conjunto foi dividido de maneira aleatória em outros dois conjuntos, chamados de *conjunto de treino* e *conjunto de teste*. O conjunto de treino foi utilizado para estimar os parâmetros necessários para calcular os coeficientes dos filtros de reconstrução de energia, tais como a matriz de covariância do ruído e pedestal.

Já para simular os sinais com efetiva deposição de energia, o primeiro passo foi gerar as amplitudes verdadeiras, as quais deverão ser recuperadas pelos métodos para reconstruir os sinais. Nesse processo, utilizou-se um SPR (do inglês, *Signal-to-Pile-up Ratio*) de 2, típico de eventos selecionados pelo sistema de filtragem (Barbosa, 2017). O SPR é uma relação entre o sinal de interesse e o ruído, e essa relação pode ser feita entre as amplitudes.

Assim, a amplitude verdadeira foi simulada utilizando-se uma distribuição exponencial de média igual ao dobro da média do ruído de empilhamento, ou seja, de 100 ADC. Foram geradas um total de 50.000 amplitudes, as quais foram multiplicadas, cada uma, pelo pulso de referência e o resultado foi somado ao conjunto de teste do ruído. Com essa soma, tem-se o sinal completo recebido, a partir do qual o sinal de interesse precisa ser reconstruído.

Vale ressaltar que a unidade ADC se refere à contagens de ADC, e pode ser traduzida em unidade de energia. No TileCal, 1 ADC corresponde a 12 MeV (Mega elétron-Volt).

### 5.2 Análise de eficiência

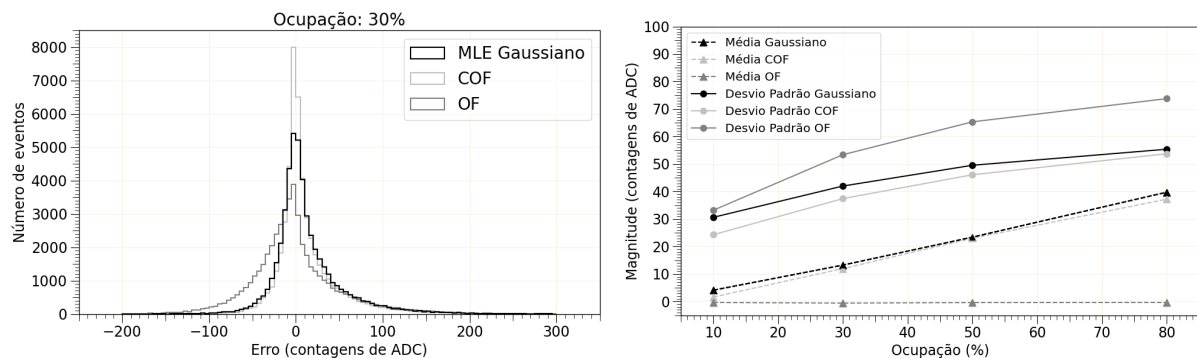
**Cálculos de parâmetros** Para a análise da eficiência da reconstrução do sinal de interesse, foi feita uma comparação entre dois parâmetros: o método  $\chi^2$ , já utilizado e bem difundido, e o método proposto, baseado na probabilidade do ruído.

O método  $\chi^2$  foi calculado a partir da subtração entre o sinal recebido e o sinal de interesse estimado do método em questão, utilizando-se a Eq. (1).

Já a probabilidade pode ser diferente para cada método. Para o MLE Gaussiano e o OF, o ruído utilizado na PDF (mostrada na Eq. (9)) foi simplesmente a subtração entre o sinal recebido e o sinal de interesse estimado de cada método. O COF, por outro lado, é um método capaz de fornecer a amplitude de todos os sinais observados na janela de leitura, não apenas do sinal

de interesse. Assim, para usar a PDF Gaussiana com o COF, foi necessário subtrair do sinal recebido cada uma das amplitudes multiplicada pela matriz  $H$  do método. O resultado dessa subtração foi substituído na PDF Gaussiana normalmente.

**Resultados** Inicialmente, é interessante fazer-se uma análise sobre a eficiência dos métodos na reconstrução do sinal de interesse. A Fig. 2a mostra o histograma do erro de estimação de cada um dos métodos, para a ocupação de 30%. Neste caso, o erro de estimação corresponde à diferença entre o valor estimado da amplitude e o valor verdadeiro conhecido da simulação. É possível notar que, para essa ocupação, o método COF se sobressai aos outros dois, conforme esperado, visto a sua capacidade de operar melhor em condições de empilhamento de sinais (Gonçalves, 2022).



(a) Histogramas dos erros de estimação para os métodos MLE Gaussiano, COF e OF, na ocupação de 30%.

(b) Média e Desvio Padrão dos erros, para os três métodos, nas ocupações de 10%, 30%, 50% e 80%.

Para uma análise mais detalhada, a Fig 2b mostra como se comportam a média e o desvio padrão da distribuição de erro para os três métodos, nas quatro ocupações. Essa figura reforça o que foi observado na Fig. 2a e mostra um comportamento semelhante nas demais ocupações.

A Fig. 3 mostra as distribuições de erro versus  $\chi^2$  e probabilidade, para o método OF. Para uma reconstrução eficiente do sinal de interesse, valores grandes de erro devem resultar em  $\chi^2$  grandes e probabilidades pequenas, assim como erros pequenos devem ter  $\chi^2$  pequenos e probabilidades maiores. Dessa forma, distribuições muito dispersas tendem a caracterizar reconstruções menos eficientes.

Observando a Fig. 3, é possível notar que a distribuição de erro versus  $\chi^2$  possui uma maior dispersão do que aquela de erro versus probabilidade. A maior centralização da distribuição, em torno de  $x = 0$ , observada na Fig. 3b indica que a utilização da probabilidade pode ser uma boa alternativa ao  $\chi^2$  quando se deseja avaliar a qualidade da reconstrução de um sinal.

Com a ideia de aperfeiçoar a capacidade de reconstrução dos métodos, foi realizado um tratamento de qualidade de dados, baseado em cortes consecutivos nos valores de  $\chi^2$  e probabilidade de cada método. Foram feitos um total de 100.000 cortes a um passo correspondente ao valor da subtração entre o máximo e o mínimo da distribuição, dividida pelo número de cortes. A ideia é que, a cada corte, sejam encontrados na distribuição apenas os valores que sejam maiores que o passo multiplicado pelo corte  $n$ . Com esses valores, foi possível encontrar os erros correspondentes, nas distribuições de erros, e calcular seu desvio padrão.

A Fig. 4 mostra o gráfico que relaciona a quantidade de cortes e o desvio padrão do erro para o  $\chi^2$  e a probabilidade de cada um dos métodos, nas ocupações de 30% e 80%. Os dados foram

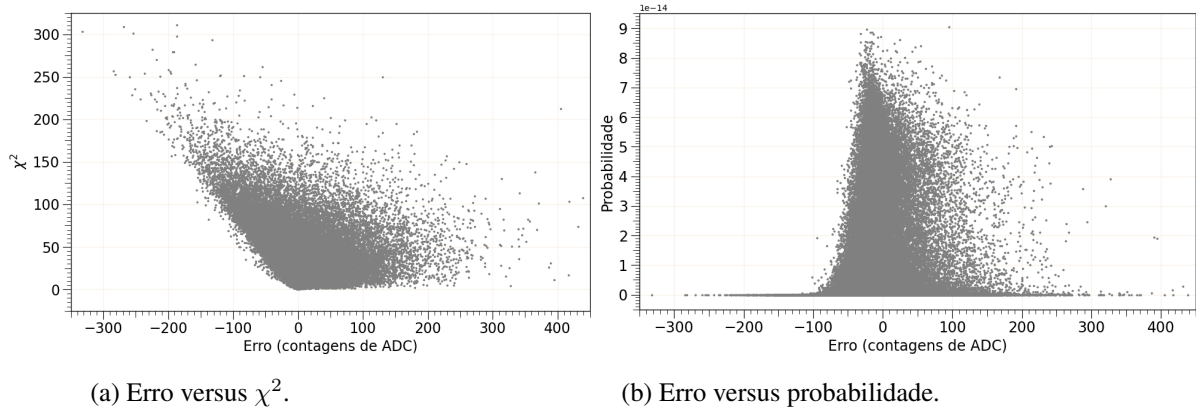


Figura 3: Distribuição de erro versus  $\chi^2$  e erro versus probabilidade para o OF, na ocupação de 30%.

plotados apenas até um total de 400 cortes para facilitar a visualização, sabendo que valores muito grandes de cortes resultam em uma perda razoável de dados.

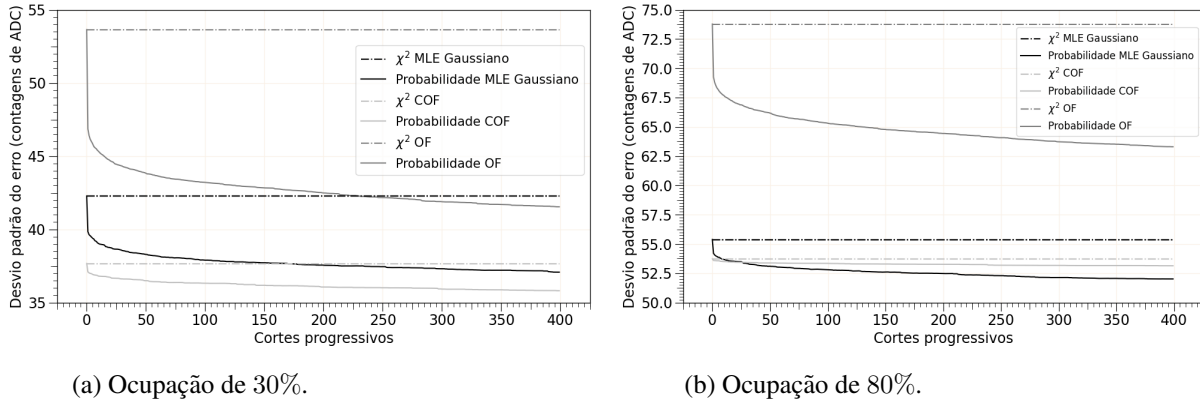


Figura 4: Corte versus desvio padrão do erro, comparando o  $\chi^2$  e a probabilidade de cada um dos métodos.

Analisando a Fig. 4a com relação ao tratamento dos dados, pode-se perceber que, para os três métodos, quando se trata da probabilidade, a qualidade dos dados pode ser apurada conforme os cortes são realizados. Para os valores de  $\chi^2$ , no entanto, independentemente do método, o tratamento aplicado não parece surtir muito efeito. Notou-se que os valores do  $\chi^2$  se mantêm constantes por uma larga faixa de quantidade de cortes, sendo necessário um número maior de cortes para que uma modificação muito menor seja observada.

Na Fig. 4b, é possível notar que o comportamento se repete para ocupação de 80%. A diferença que se destaca é que, nessa ocupação, a qualidade da reconstrução no MLE Gaussiano se distancia ainda mais do OF, mantendo sua superioridade. Ao mesmo tempo, a diferença entre as curvas de  $\chi^2$  e probabilidade para o COF parece diminuir, mostrando que, para este método, em altas ocupações, o tratamento dos dados começa a perder sua eficiência na análise de reconstrução do sinal usando a probabilidade.



## 6. CONCLUSÕES

A eficiência da reconstrução da energia em calorimetria de altas energias está diretamente associada à qualidade dos sinais reconstruídos. Tipicamente, o método do  $\chi^2$  é utilizado como fator de qualidade, embora este método não opere de forma eficiente na presença de empilhamento de sinais. Desta forma, este trabalho apresentou uma abordagem alternativa baseada na probabilidade do ruído presente no sinal. Visto que os métodos de reconstrução da energia aproximam o ruído por uma função Gaussiana multivariada, o método proposto utiliza a amplitude estimada e o sinal recebido para avaliar a qualidade da reconstrução baseado na PDF do ruído. Os resultados mostraram que o uso do método do  $\chi^2$ , atualmente utilizado no TileCal, não foi capaz de descartar sinais com problemas de reconstrução de energia, e consequentemente reduzir o erro de estimação. Por outro lado, o uso da probabilidade se mostrou promissor, uma vez que foi observado uma melhor correlação desta métrica com o erro de estimação.

Os trabalhos futuros visam avaliar outros modelos de ruído, tais como a função multivariada Lognormal, que pode descrever melhor o empilhamento de sinais (Rimes, 2020).

## Agradecimentos

Os autores agradecem a CAPES, CNPq, FAPEMIG, FAPERJ e RENAFAP pelo apoio. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

## REFERÊNCIAS

- Aad, G. et al. (2010), "Readiness of the ATLAS Tile Calorimeter for LHC collisions". *European Physical Journal C*, v. 70, p. 1193 – 1236.
- Anderson, T. W. (2003), *An Introduction to Multivariate Statistical Analysis*. 3. ed. Stanford, CA: Wiley-Interscience.
- Banerjee, S. (2012), *CMS simulation software*, J. Phys., Conf. Ser., vol. 396, no. 2, Art. no. 022003.
- Barbosa, D. P. et al. (2017), *Sparse Representation for Signal Reconstruction in Calorimeters Operating in High Luminosity*, in IEEE Transactions on Nuclear Science, vol. 64, no. 7, pp. 1942-1949.
- Bruning, O. e Rossi, L. (2015), *The High Luminosity Large Hadron Collider. The New Machine for Illuminating the Mysteries of Universe*, volume 24. World Scientific, Advanced Series on Directions in High Energy Physics.
- CERN (2023). Disponível em: <https://home.cern/>. Acesso em 14 fev. 2023.
- Chapman, J. (2011), *ATLAS simulation computing performance and pile-up simulation in ATLAS*, presented at the LPCC Detector Simulation Workshop, Geneva, Switzerland.
- Clement, C. e Klimek, P. (2011), "Identification of pile-up using the quality factor of pulse shapes in the ATLAS Tile Calorimeter", *2011 IEEE Nuclear Science Symposium Conference Record*, pp. 1188-1193, Valencia, Spain.
- de Andrade Filho, L. M. et al. (2015), "Calorimeter Response Deconvolution for Energy Estimation in High-Luminosity Conditions". *IEEE Transactions on Nuclear Science*, v. 62, n. 6, p. 3265–3273.
- Evans, L. e Bryant, P. (2008), *LHC Machine*. JINST 3 S08001.
- Fullana, E. et al. (2006), "Digital Signal Reconstruction in the ATLAS Hadronic Tile Calorimeter". *IEEE Transactions on Nuclear Science*, v. 53, n. 4.
- Gonçalves, G. I. (2023), *Simulation code for analysis*. GitHub repository. <https://github.com/ingoncalves/jcae-2021-analysis>.
- Gonçalves, G. I. et al. (2022), "Performance of Optimal Linear Filtering Methods for Signal Estimation in High-Energy Calorimetry". *Journal of Control, Automation and Electrical Systems*, p. 1-11.
- Hauptman, J. (2010), *Particle Physics Experiments at High Energy Colliders*, Wiley-VCH, 1st edition.
- Herr, W. e Muratori, B. (2003), *Concept of luminosity*. In: CERN Accelerator School and DESY Zeuthen: Accelerator Physics. [S.l.: s.n.]. p. 361–377.

- Kay, S. M. (1993), *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, New Jersey.
- Mahfoudia, O.; Horlin, F. e Neyt, X. (2017), *Optimum reference signal reconstruction for DVB-T based passive radars*, 2017 IEEE Radar Conference (RadarConf), Seattle, WA, USA, pp. 1327-1331.
- Mitra, S. (2001), *Digital Signal Processing : A Computer-Based Approach*, McGraw-Hill Higher Education; 2nd edition.
- Papoulis, A. e Pillai, S. U. (2002), *Probability, Random Variables, and Stochastic Processes*. 4. ed. New York, NY: McGraw-Hill.
- Peralva, B. S. (2013), "The Tilecal Energy Reconstruction for Collision Data Using the Matched Filter". *IEEE Nuclear Science Symposium and Medical Imaging Conference*, p. 1 – 6.
- Peralva, B. (2022), *Upgrade of ATLAS hadronic Tile Calorimeter for the High-Luminosity LHC*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, Volume 1038, 166951.
- Rimes, S. M. et al (2020), *Estimação da amplitude de sinais em calorimetria de altas energias em condições de alta ocupação de eventos no detector*, XXXVIII Simpósio Brasileiro de Telecomunicações e Processamento de Sinais (SBrT).
- Rimes, S. M. (2021), *Filtragem inversa não-linear para estimação de sinais em calorímetros operando a alta taxa de eventos*. Orientador: Bernardo Sotto-Maior Peralva. 2021. 86. Dissertação (Mestrado) - Modelagem Computacional, Instituto Politécnico do Rio de Janeiro (UERJ), Nova Friburgo - RJ.
- Seixas, J. M. (2015), "Quality Factor for the Hadronic Calorimeter in High Luminosity Conditions", *Journal of Physics: Conference Series*, v. 608, pp. 012044.
- The ATLAS Collaboration (2008), *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST, 3:S08003.
- The ATLAS Collaboration (2010), *Readiness of the ATLAS tile calorimeter for LHC collisions*, Eur. Phys. J. C, 70, pp. 1193-1236.
- The ATLAS Collaboration (2012), *Observation of a new particle in the search for the standard model Higgs boson with the ATLAS detector at the LHC*, Physics Letters B, vol. 716, no. 1, pp. 1 – 29.
- Valero, A. (2012), *Implementation and Performance of the Signal Reconstruction in the ATLAS Hadronic Tile Calorimeter*, Physics Procedia, Volume 37, Pages 1765-1771.
- Wigmans, R. (2017), *Calorimetry: Energy Measurement in Particle Physics*, International Series of Monographs on Physics, OUP Oxford.

## PROBABILITY BASED DATA QUALITY FOR ENERGY RECONSTRUCTION IN HIGH ENERGY CALORIMETRY

**Abstract.** In several engineering applications, the quality of the reconstruction of the information of interest directly interferes with the success of decision making. In modern high-energy physics experiments, such as the ATLAS at the LHC, the reconstruction of events is essential for the precise observation of particles and important discoveries for science. In this context, the signals from the calorimetry system provide information about the energy of the particles produced in collisions. In the ATLAS tile calorimeter (TileCal), the energy reconstruction quality is measured using the  $\chi^2$  method. However, such an approach is not able to operate under signals pile-up conditions, being only able to detect the presence of pile-up effect in the received signal. Therefore, in this work, a new approach to evaluate the quality of energy reconstruction under pile-up conditions is presented. The proposed method relies on the calculation of probabilities regardless of the energy reconstruction algorithm employed. Using simulated TileCal data under different signals pile-up conditions, the results show that the use of the proposed method increases signals selection efficiency.

**Keywords:** Data quality, optimal filtering, signal pile-up, high energy calorimetry.