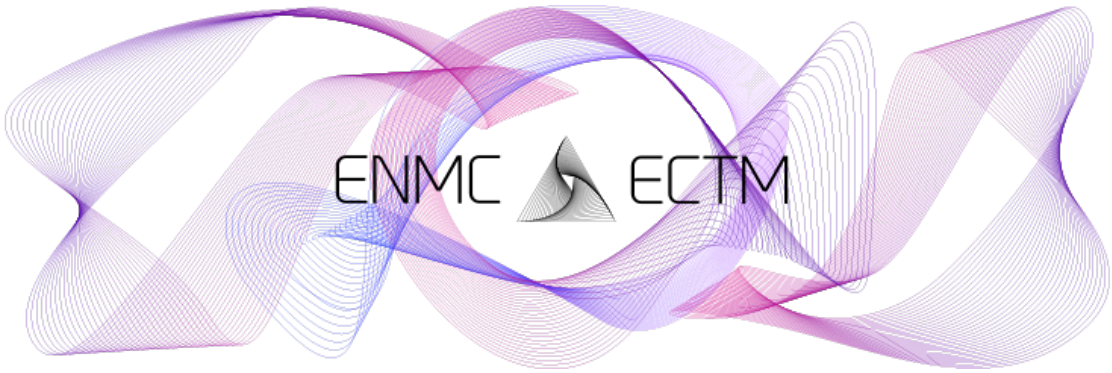

08 a 11 de Outubro de 2019
Universidade Federal de Juiz de Fora
Juiz de Fora - MG

ESTIMAÇÃO DA AMPLITUDE UTILIZANDO O ESTIMADOR DE MÁXIMA VEROSSIMILHANÇA EM CONDIÇÕES DE EMPILHAMENTO DE SINAIS

Sarita de Miranda Rimes[1] - saritamrimes@gmail.com
Lucas de Souza Gomes Nolla[1] - lucasnolla@outlook.com.br
Bernardo Sotto-Maior Peralva[1] - bernardo@iprj.uerj.br
Luciano Manhães de Andrade Filho[2] - luciano.andrade@ufjf.edu.br
Augusto Santiago Cerqueira[2] - augusto.santiago@ufjf.edu.br
José Manoel de Seixas[3] - seixas@lps.ufrj.br
[1]Universidade do Estado do Rio de Janeiro, Instituto Politécnico - Nova Friburgo, RJ, Brasil
[2]Universidade Federal de Juiz de Fora, Faculdade de Engenharia - Juiz de Fora, MG, Brasil
[3]Universidade Federal do Rio de Janeiro, Poli/COPPE - Rio de Janeiro, RJ, Brasil

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.*

Palavras-chave: *Estimação de parâmetros, Filtro ótimo, Empilhamento de sinais, Física de altas energias.*

1. INTRODUÇÃO

Desde que o ser humano se tornou capaz de pensar, é possível registrar sua curiosidade e ânsia por entender como o mundo ao seu redor funciona. Uma forma de se obter esse entendimento é através do estudo daquilo que constitui as coisas que o rodeiam, e, neste sentido, encontra-se uma área da física que estuda as partículas elementares que formam o universo: a física de partículas (Close, 2004).

Na área da física de altas energias, partículas elementares podem ser estudadas em experimentos que geram colisões entre elas, recolhendo os dados provenientes. Tais experimentos são feitos através da aceleração de feixes de partículas, de forma que, em dado momento, elas colidam e, dessas colisões, sejam gerados dados que possibilitem a análise da natureza das mesmas (Cottingham et al., 2001).

O maior e mais energético acelerador de partículas do mundo, atualmente, é o LHC (*Large Hadron Collider*). Localizado no CERN (sigla para a *European Organization for Nuclear Research*), na fronteira entre a França e a Suíça, próximo a cidade de Genebra, o LHC é composto por dois aneis supercondutores, com circunferências de cerca de 27 quilômetros, construídos em um túnel 100 metros abaixo do nível do solo. Dentro desses anéis ocorrem acelerações de feixes de prótons, com colisões acontecendo a cada 25 ns, em locais específicos, a uma taxa máxima de 40 MHz (Evans & Bryant, 2008).

Para a coleta de dados das colisões, o LHC possui quatro grandes experimentos com detectores de partículas, posicionados em locais estratégicos a fim de aproveitar ao máximo as informações obtidas nas colisões. O maior desses experimentos é o ATLAS (*A Toroidal LHC ApparatuS*) (The ATLAS Collaboration, 2008), que possui seis componentes principais, sendo um destes o seu Calorímetro Hadrônico, no qual encontra-se o calorímetro de telhas (The ATLAS Collaboration, 2010), ou TileCal, sendo este o objeto de estudo deste trabalho.

O TileCal é responsável por absorver e medir a energia das partículas que com ele interagem (Wigmans, 2000). Durante a operação do experimento, a eletrônica de leitura do TileCal leva cerca de 150 nanossegundos para responder a uma dada colisão. Como as colisões acontecem em intervalos de tempo menores (a cada 25 ns), tem-se a ocorrência de um fenômeno chamado de *empilhamento de sinais*, ou seja, antes que um sinal seja totalmente processado, uma próxima colisão ocorre e o sinal desta colisão acaba se misturando com o daquela que ainda estava sendo amostrada. Isso sobrevém devido as altas taxas de luminosidade (número médio de interações entre prótons a cada colisão) sob as quais o LHC opera, e que tende a aumentar, visando ampliar o espectro de descobertas do experimento.

O sinal recebido é digitalizado na taxa de 40 MHz, e sua energia é medida através da estimação da amplitude das amostras disponíveis. Com a sobreposição dos sinais, essa análise fica comprometida, aparecendo, portanto, a necessidade de filtragem da informação gerada. O sinal sobreposto, que não é de interesse, corresponde ao ruído e através da modelagem do mesmo é possível melhorar a eficiência de estimação da energia.

Atualmente, o método empregado no TileCal para a estimação de energia não prevê a presença de empilhamento de sinais, e utiliza uma distribuição multivariada Gaussiana para modelar o ruído. No entanto, têm-se observado que, sob altas condições de empilhamento, essa distribuição não fornece mais uma descrição adequada. Logo, em condições de empilhamento de sinais, a eficiência de tal abordagem é comprometida. Desta forma, o presente trabalho visa projetar um método alternativo para a estimação da energia baseado no estimador de máxima verossimilhança (MLE), utilizando uma função multivariada que melhor descreve o ruído em condições de empilhamento de sinais.

Na próxima seção, o calorímetro de telhas do ATLAS e seu método de estimação de energia são brevemente descritos. Em seguida, o método proposto neste para a estimação da energia em condições de empilhamento de sinais é apresentado em detalhes. O banco de dados utilizado neste trabalho e os resultados são apresentados na seção 4. Por fim, as conclusões são derivadas na seção 5.

2. O CALORÍMETRO DE TELHAS

O calorímetro de telhas, chamado de TileCal, é o principal calorímetro hadrônico do experimento ATLAS (Francavilla, 2012). Como os dados utilizados no presente trabalho foram coletados a partir dele, uma breve descrição de seu funcionamento será feita aqui. O TileCal é de extrema importância, pois através dele são coletados dados que serão utilizados, *online* e *offline*, para análise das partículas observadas nas colisões.

O TileCal tem formato cilíndrico e é dividido em quatro partes: dois barris externos (EBA e EBC), um em cada extremidade, e um barril central, que é, por sua vez, particionado em dois barris (LBA e LBC). Cada um desses barris é divido em 64 módulos, sendo cada um desses módulos dividido em 23 células em cada barril central e 16 células em cada barril estendido. As células, em sua maioria, possuem dois canais cada, que são responsáveis pela leitura dos dados que chegam. Como material absorvedor, são utilizadas placas de aço, enquanto que telhas cintilantes trabalham como material ativo. Através de um processo específico, a energia da partícula detectada é descrita por um pulso de certa amplitude, e com o cálculo correto dessa amplitude é possível estimar a energia da partícula, podendo-se dizer qual partícula gerou aquele sinal (Francavilla, 2012).

2.1 Estimação do Sinal

Através de um conversor, o pulso gerado pelos equipamentos do TileCal, inicialmente analógico, é convertido em sinal digital, utilizando-se sete amostras discretas, com intervalos de 25 nanossegundos, formando uma janela de 150 nanossegundos totais. Desse sinal, é subtraída uma variável chamada pedestal, que corresponde a dados coletados antes de iniciadas as colisões e serve como linha de base para a medida da amplitude, fazendo com que a fase, que é a diferença de tempo entre a amostra central (quarta amostra) e o pico do sinal, seja nula. Tal procedimento é feito de forma que o pico do sinal se encontre no zero da abscissa e possa ser calculada a amplitude e estimada a energia da partícula. A Fig. 1 mostra um exemplo de pulso analógico que pode ser obtido e suas respectivas amostras digitais.

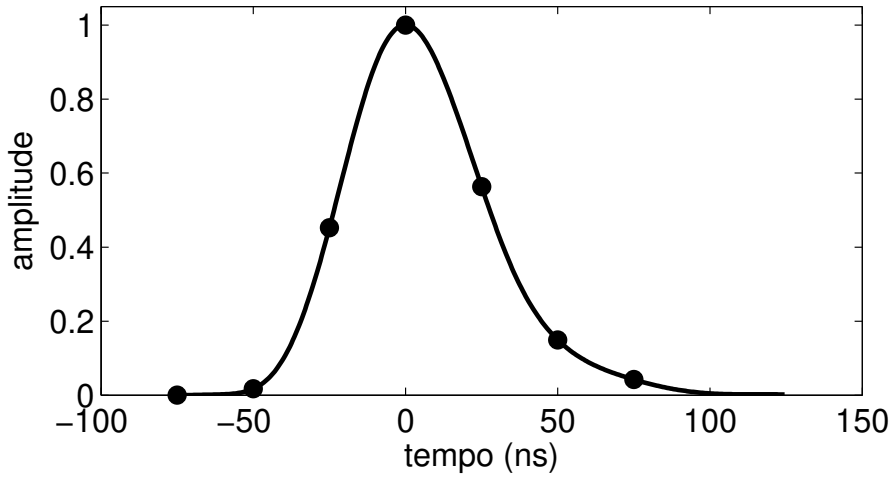

Figura 1- Pulso analógico do TileCal com amostras digitais.

O pulso apresentado na Fig. 1 pode ser muito bem descrito por uma distribuição Gaussiana.

No entanto, sob condições de empilhamento de sinais, o sinal obtido não possui esse formato. A Fig. 2 ilustra como ocorre a sobreposição dos sinais e a consequente dificuldade na medição da amplitude relacionada a uma partícula específica.

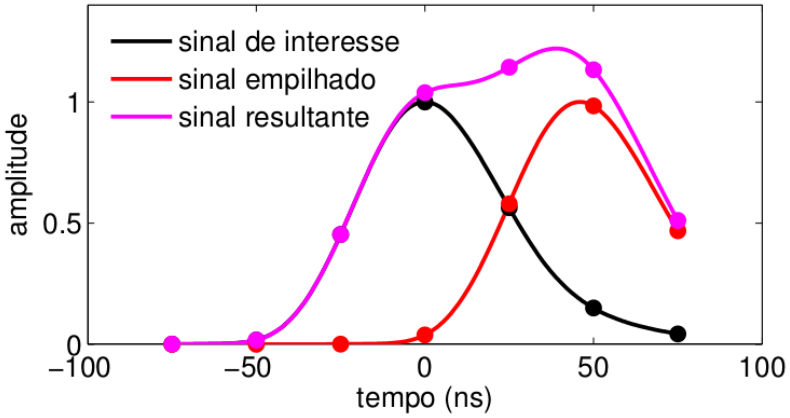

Figura 2- Sinais sobrepostos devido a alta taxa de eventos.

Na Fig. 2, o sinal de interesse é aquele que possui seu pico no do tempo $t = 0$ ns. É possível notar que, antes que este sinal pudesse ser completamente gerado, um segundo sinal chega aos canais de leitura, inserindo à janela mais uma curva e amplitude relacionada. Como consequência, o gráfico final produzido não descreve apenas a energia da partícula de interesse, mas uma junção dos dois sinais recebidos.

2.2 Filtragem do Sinal

Para resolver o problema de empilhamento de sinais, o TileCal possui um sistema de filtragem, que consiste em tratar o sinal indesejado como ruído, retirá-lo da amostra e analisar apenas o sinal de interesse.

Atualmente, o ATLAS utiliza um método chamado de Filtro Ótimo, com a sigla OF (*Optimal Filter*), que é um algoritmo que, através das amostras digitais, reconstrói o sinal, minimizando o ruído presente, fornecendo sua amplitude e a consequente estimação da energia da partícula descrita (Fullana et al., 2005).

A amplitude do sinal de interesse é dada por um somatório de pesos de cada amostra discreta e é descrita pela Eq. (1). Aqui, $w[k]$ são os coeficientes do filtro, obtidos através do pulso de referência e da matriz de covariância do ruído, e $r[k]$ é a amostra do sinal recebido no instante k.

$$\hat{A}_{OF} = \sum_{k=0}^{N-1} w[k]r[k] \tag{1}$$

Para o cálculo dos coeficientes $w[k]$, o sinal recebido $r[k]$ do TileCal pode ser descrito como:

$$r[k] = As[k] - A\tau\dot{s}[k] + n[k] + ped \qquad k = 0, 1, 2, ..., N-1 \tag{2}$$

em que $r[k]$ representa a amostra digital recebida no instante k e N corresponde ao número de amostras disponíveis (sete no caso do TileCal). A amplitude $\boldsymbol{A}$ é o parâmetro a ser estimado,

enquanto $n[k]$ representa o ruído de fundo. Os parâmetros $s[k]$ e $\dot{s}[k]$ correspondem, respectivamente, às amostras do pulso de referência do TileCal e sua derivada (aproximação linear para a fase do pulso), enquanto o parâmetro τ é a fase do sinal. A variável ped corresponde ao pedestal e é um parâmetro constante adicionado ao sinal analógico antes da digitalização.

O procedimento de otimização tem como objetivo minimizar a variância da distribuição da estimação da amplitude. Assim, este método opera próximo de seu ponto ótimo para sinais determinísticos corrompidos por ruído gaussiano. Os coeficientes são calculados através da minimização da variância do estimador em que as seguintes restrições são impostas ao procedimento de otimização:

$$\begin{aligned} &\sum_{k=0}^{N-1} w[k]s[k] = 1 \\ &\sum_{k=0}^{N-1} w[k]\dot{s}[k] = 0 \\ &\sum_{k=0}^{N-1} w[k] = 0. \end{aligned} \tag{3}$$

As restrições minimizam o efeito de flutuações na fase e linha de base do sinal recebido. Por outro lado, tais restrições aumentam a variância do estimador, tornando a medida da amplitude menos precisa.

3. ESTIMAÇÃO DA AMPLITUDE BASEADA NO MLE

Alternativamente, o problema da estimação de parâmetros também pode ser abordado através da maximização da densidade de probabilidade do sinal recebido $p(\mathbf{s}|\hat{A}_{mle})$. O valor de $\hat{A}_{mle}$ que maximiza $p(\mathbf{s}|\hat{A}_{mle})$ é a melhor estimativa para a amplitude do sinal recebido. Este método alternativo é chamado de MLE (do inglês, *Maximum Likelihood Estimator*) (Kay, 1993), e pode ser definido como:

$$\frac{\partial p(\mathbf{n}|\hat{A}_{mle})}{\partial A} = 0. \tag{4}$$

em que o vetor $\mathbf{n}$ corresponde ao processo que está sendo modelado. Visto que o sinal produzido pela eletrônica do TileCal é validado por um complexo sistema de calibração, o desenvolvimento do método MLE assume $\tau = 0$ e o valor do pedestal é subtraído assim que as amostras digitais são recebidas. Desta forma, o modelo utilizado pelo método MLE pode ser representado por:

$$r[k] = As[k] + n[k] \qquad\qquad k = 0, 1, 2, ..., N-1 \tag{5}$$

em que A representa a amplitude do sinal recebido, $s[k]$ corresponde ao pulso de referência do TileCal, $n[k]$ são as amostras do ruído, e $N = 7$ para o caso do TileCal.

3.1 MLE para ruído Gaussiano

Para o caso particular em que as amostras do ruído $\mathbf{n}$ possam ser modeladas por uma distribuição gaussiana multivariável com uma matriz de covariância $\mathbf{C}$ e um vetor de médias

$\boldsymbol{\mu}$, a função densidade de probabilidade é dada pela seguinte expressão (Anderson, 2003):

$$p(\mathbf{n}) = \frac{1}{|\mathbf{C}|^{-1}(2\pi)^{\frac{N}{2}}} \exp\left[\frac{(\mathbf{n}-\boldsymbol{\mu})^T\mathbf{C}^{-1}(\mathbf{n}-\boldsymbol{\mu})}{2}\right] \tag{6}$$

em que $\mathbf{n} = \mathbf{r} - A\mathbf{s}$ e $\boldsymbol{\mu}$ é um vetor nulo, por se tratar de uma distribuição Normal. Extraindo o logaritmo e derivando a Equação (6) em função de A, a expressão para A_{normal} (estimativa de A) pode ser encontrada. Ou seja, a estimativa da amplitude do sinal recebido é dada por:

$$\hat{A}_{normal} = \frac{\mathbf{r}^T\mathbf{C}^{-1}\mathbf{g}}{\mathbf{g}^T\mathbf{C}^{-1}\mathbf{g}}. \tag{7}$$

3.2 MLE para ruído Lognormal

A modelagem do ruído através da distribuição Gaussiana não é eficiente em condições de alto empilhamento de sinais. Portanto, é necessário que seja utilizada uma outra abordagem, que se assemelhe mais ao ruído nessa conjuntura. O pulso de referência do TileCal é unipolar, ou seja, se mantém completamente acima da linha de base, e, portanto, os sinais de empilhamento recebidos podem ser tratados como uma soma. Esse tipo de comportamento gera uma distribuição conhecida como Gamma (Evans et al., 2000). No entanto, essa distribuição tem como característica o fato de possuir muitos parâmetros que precisam ser determinados, o que dificulta a modelagem do problema.

Com o intuito de solucionar o problema da quantidade de parâmetros, no lugar da distribuição Gamma pode ser utilizada uma distribuição Lognormal. Essa aproximação é comumente feita (quando há a possibilidade de utilização de apenas dados positivos) já que a Lognormal tem propriedades que fazem com que seu uso, neste caso, seja confiável e possui a vantagem de ter poucos parâmetros a determinar (Johnson et al., 1995).

Portanto, a PDF sobre a qual o MLE será aplicado é dada por (Tarmast, 2001):

$$p(\mathbf{n}) = \frac{1}{\left(\prod_{i=0}^{N-1} n[i]\right)|\mathbf{C}|^{\frac{1}{2}}(2\pi)^{\frac{N}{2}}} \exp\left[\frac{-[\ln(\mathbf{n})-\boldsymbol{\mu}]^T\mathbf{C}^{-1}[\ln(\mathbf{n})-\boldsymbol{\mu}]}{2}\right] \tag{8}$$

em que $\boldsymbol{\mu}$ e $\mathbf{C}$ correspondem aos parâmetros de média e matriz de covariância da distribuição lognormal multivariada, respectivamente. Vale ressaltar que $\mathbf{n} = \mathbf{r} - A\mathbf{s}$, e que o objetivo do método MLE é encontrar a estimativa para A que maximize $p(\mathbf{n})$.

No presente trabalho, para a maximização da PDF, optou-se por utilizar um recurso baseado em *busca exaustiva*, em que o parâmetro A é variado dentro de uma faixa de valores factíveis. O valor de A ($\hat{A}_{lognormal}$) que gera o maior resultado para a $p(\mathbf{n})$ é escolhido.

4. RESULTADOS

Para avaliar a eficiência de estimação da amplitude (energia), o método proposto foi implementado computacionalmente juntamente com o método OF. Os métodos foram aplicados a dados simulados com diferentes condições de empilhamento de sinais.

4.1 Simulação

Os sinais utilizados para testar os estimadores descritos neste trabalho foram simulados utilizando os parâmetros encontrados no calorímetro de telhas do ATLAS. O ruído é composto pelo ruído eletrônico (proveniente da eletrônica de leitura) e pelo ruído de empilhamento de sinais (sinais provenientes de colisões adjacentes). O ruído eletrônico possui a mesma descrição, independente da condição de empilhamento de sinais imposta e, no TileCal, possui média zero e desvio padrão de 1,5 contagens de ADC. Já o ruído de empilhamento de sinais depende das condições de operação do LHC e da posição física da célula de leitura do calorímetro. Visando estudar a eficiência de estimação da amplitude em diversas condições de empilhamento de sinais, foi gerado condições de ocupação de 10% a 90%. Na condição de 10% de ocupação, um dado canal de leitura possui 10% de probabilidade de produzir um sinal numa dada colisão, produzindo um empilhamento de sinais.

Para cada condição de empilhamento de sinais, dois conjuntos foram produzidos. O primeiro conjunto contém 50.000 observações de ruído, e foi utilizado para estimar os parâmetros dos ruído tais como média e matriz de covariância, que são utilizados nos métodos utilizados neste trabalho. O segundo conjunto possui 50.000 sinais característicos do TileCal dos quais a amplitude deve ser estimada. As amplitudes destes sinais foram escolhidas de forma aleatória segundo uma distribuição exponencial de média igual a 100. Estes sinais estão imersos em ruído similar ao descrito anteriormente, ou seja, com componentes de ruído eletrônico e de empilhamento de sinais. Este conjunto foi utilizado para a análise de eficiência dos métodos.

4.2 Análise de eficiência

Para analisar a eficiência dos métodos para as diferentes condições de operação, o erro de estimação foi utilizado. O erro de estimação corresponde à diferença entre o valor estimado e o valor verdadeiro, conhecido da simulação. O parâmetro utilizado foi o desvio padrão da distribuição do erro de estimação. Vale ressaltar que um estimador deve possuir média do erro de estimação igual a zero e um desvio padrão o menor possível. A Figura 3 mostra as distribuições do erro de estimação dos métodos descritos neste trabalho para a condição de 90% de ocupação. Conforme pode ser observado, os métodos MLE possuem uma dispersão menor que o método OF, indicando uma melhor eficiência na medida da amplitude. O método MLE lognormal apresentou o melhor desempenho, levemente superior ao método MLE normal (em torno de 3%).

Na Figura 4 o desvio padrão dos erros de estimação para cada condição de ocupação é mostrada. Podemos observar que o método MLE lognormal possui uma eficiência melhor, visto que visualmente apresenta uma menor dispersão para toda a faixa de ocupação considerada. O método OF apresentou a pior eficiência (maior dispersão), devido às restrições que são impostas ao procedimento de otimização para cálculos de seus coeficientes.

5. CONCLUSÕES

Neste trabalho, foi apresentado o problema de estimação da energia para um calorímetro moderno operando em condições de alta taxa de eventos, em que o fenômeno de empilhamento de sinais é presente. Em condições em que somente o ruído eletrônico (Gaussiano) é presente, métodos lineares são tipicamente empregados, dada a sua simplicidade e resposta rápida. Por outro lado, a eficiência de tais métodos é degradada quando o ruído perde suas características

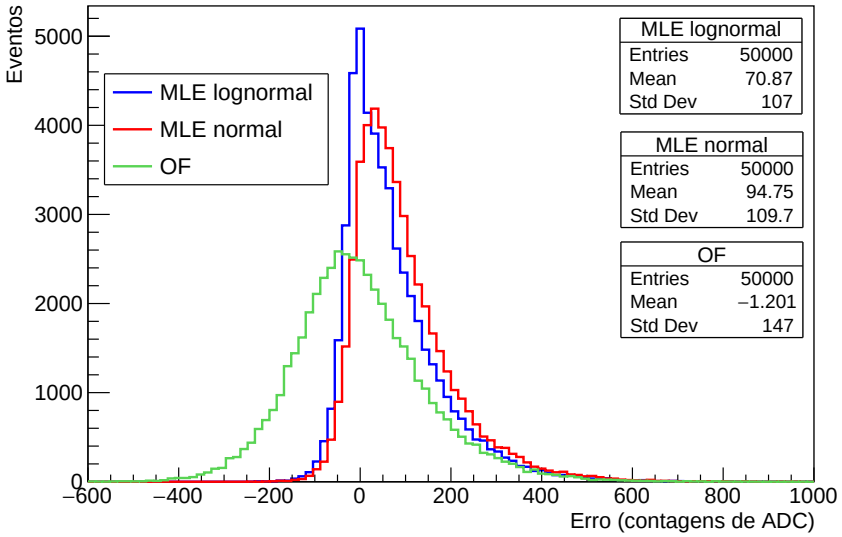

Figura 3- Erro de estimação associado aos métodos MLE lognormal, MLE normal e OF considerando o cenário de 90% de ocupação.

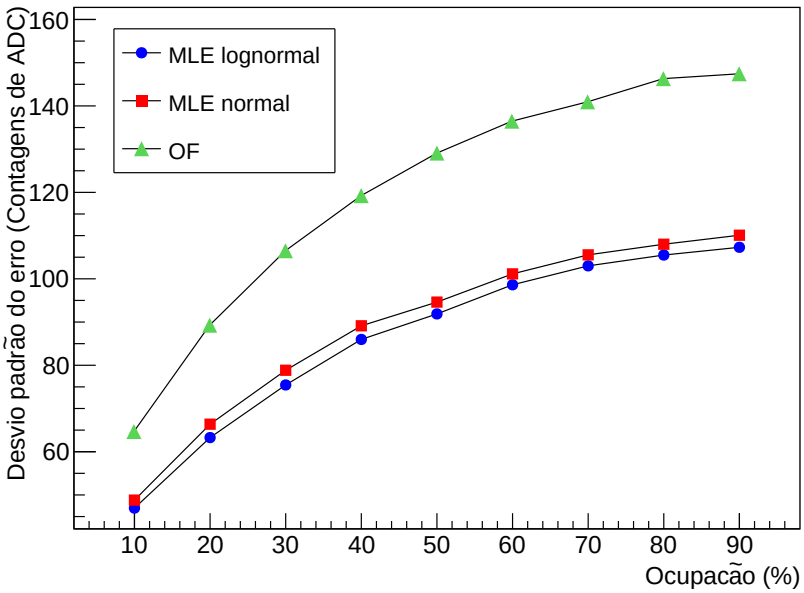

Figura 4- Desvio padrão do erro de estimação para os métodos avaliados considerando vários níveis de empilhamento de sinais (ocupação).

Gaussianas, como é o caso dos cenários de empilhamento de sinais. No ambiente de ruído de empilhamento de sinais, o modelo Lognormal se mostrou mais adequado, visto que o método MLE para este tipo de ruído apresentou a melhor eficiência para toda a faixa de ocupação estudada.

O trabalhos futuros se concentram no uso de um método numérico para a busca da ampli-

tude ótima, que maximiza a função multivariada Lognormal. Além disso, o método MLE será implementado em software no pacote de reconstrução de eventos visando avaliar o impacto da estimação da energia na reconstrução de partículas utilizando dados reais adquiridos durante a operação nominal do LHC.

Agradecimentos

Os autores agradecem a CAPES, FAPERJ, CNPq e RENAVAE pelo apoio para a realização deste trabalho.

Referências

Anderson, T.W. (2003), "*An Introduction to Multivariate Statistical Analysis*", 3º ed., Wiley-Interscience, New Jersey.

Cottingham, W.; Greenwood, D. (2001). *An Introduction to the Model of Particle Physics*. Cambridge University Press.

Close, F. (2004). *Particle Physics: A Very Short Introduction*. Very Short Introductions. OUP Oxford.

Evans, L.; Bryant, P. (2008), *LHC Machine*. JINST 3 S08001.

Evans, M.; Hastings, N.; Peacock, B. (2000), "*Statistical Distributions*", 3º ed., Wiley, New York.

Francavilla, P. (2012), *The ATLAS Tile Hadronic Calorimeter performance at the LHC*. Journal of Physics: Conference Series, v. 404, pp. 012007.

Fullana, E. et. al. (2006), *Digital Signal Reconstruction in the ATLAS Hadronic Tile Calorimeter*, IEEE Transaction On Nuclear Science, v. 53, number 4, pp. 2139-2143.

Johnson, N.L.; Kotz, S.; Balakrishnan, N. (1995), "*Continuous Univariate Distributions*", 2º ed., Wiley, New York.

Kay, S.M. (1993), "*Fundamentals of Statistical Signal Processing, Estimation Theory*", Prentice Hall, New Jersey.

Tarmast, G. (2001), *Multivariate LogNormal Distribution*, ISI Proceedings: 53º Session Seoul.

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*, EPJC 70, pp. 1193-1236.

Wigmans, R. (2000). *Calorimetry: Energy Measurements in Particle Physics*. Clarendon Press.

AMPLITUDE ESTIMATION USING THE MAXIMUM LIKELIHOOD ESTIMATOR IN SIGNAL PILE-UP CONDITIONS

Abstract. *Signal reconstruction systems rely on parameter estimation from received digital time samples. In high-energy calorimetry, the signal is conditioned in such a way that its shape is fixed, and the amplitude corresponds to the parameter to be estimated. Typically, linear methods are employed as the electronic noise may be modeled by a gaussian function. However, in high event rate operation, such as in the LHC experiment at CERN, the received signal is distorted by the signal pile-up effect, degrading the performance from typical linear methods. Therefore, this work evaluates the use of alternative methods based on the Maximum Likelihood Estimator (MLE), where the noise is described by multivariate Gaussian and Lognormal functions. In order to compare the efficiency, a data set was produced considering the main hadronic calorimeter (TileCal) of ATLAS, at LHC. The results show that the MLE methods outperform the currently method used in TileCal.*

Keywords: *Parameter estimation, Optimal filter, Signal pile-up, High-energy physics.*