Systematic Effects and Nuisance Parameters in Particle Physics Data Analyses

Dalil Koheeallee Supervisor: Prof. Anthony Davison

EPFL

Introduction

An approach using the Expectation-Maximization (EM) algorithm is introduced to estimate the rate of Higgs boson production events amidst background noise in high-energy physics. By focusing on two critical feature variables (x_1, x_2), respectively the missing transverse energy (MET) and the energy around a lepton candidate (ISO), we aim to distinguish the signal from the background.

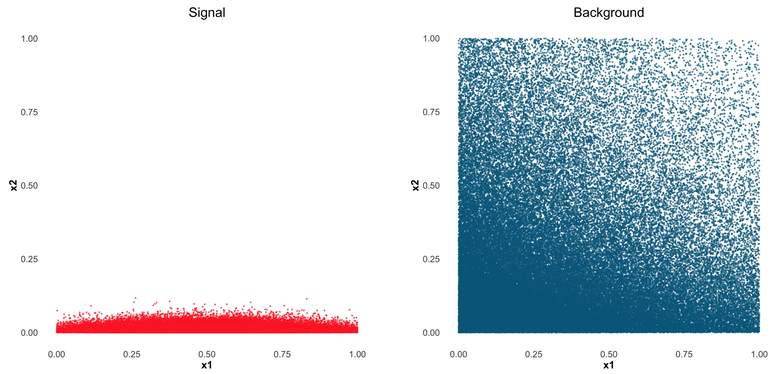

Figure 1. Distribution of signal and background in the (x_1, x_2) feature plane for the training set.

The data processing involves spatially discretizing the dataset into 10000 equal squares based on the coordinates, which range from 0 to 1. Each data point is assigned to a square, with the square's center coordinates and the count of occurrences in that square forming the basis of the new dataset.

Generalized additive model

In the generalized additive model (GAM), we consider each response variable Y_i as following an exponential dispersion family distribution with mean $\mu_i \equiv \mathbb{E}(Y_i)$. The model is structured such that the link function g relates to the linear predictor, which is a sum of smooth functions of covariates and parametric model components. It is expressed as

$$g(\mu_i) = A_i\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{1i}, x_{2i}) + \dots,$$

where A_i represents a row of the model matrix for the parametric components, θ is the parameter vector and f_j are the smooth functions of the covariates.

In this study, we conducted a preliminary analysis using GAMs with a log link function in a Poisson framework to evaluate the suitability of these models for our data. We fitted separate GAMs for signal and background data. Initial results, as shown in Figure 2, reveal a better model fit for signal data.

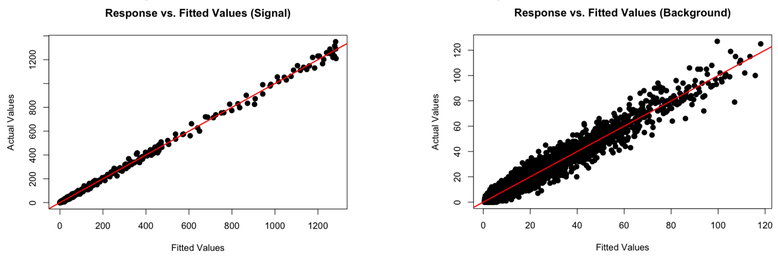

Figure 2. GAM fitted to signal and background. The red line corresponds to the identity function.

EM algorithm

The Expectation-Maximization (EM) algorithm is employed to compute the maximum likelihood estimate (MLE) in scenarios involving incomplete data, characterized by the probability density function $g(y;\theta)$. Considering y as incomplete data and x as complete data, with $y = y(x)$, the algorithm iteratively maximizes the expected complete-data log likelihood $\log L_c(\theta)$ through a two-step process. The E-step computes

$$Q(\theta;\theta^{(k)}) = \mathbb{E}_{\theta^{(k)}}[\log L_c(\theta) \mid y],$$

and the M-step then seeks to maximize this expectation. The iteration continues until convergence, effectively handling datasets with missing or unobservable components.

EM Algorithm for GAM Parameter Estimation

In our context, the aim of the EM algorithm is to find the rates by estimating the parameters of the GAM model using the incomplete data. The incomplete data comprise only the sum of the rates of signal and background within each square of the discretized dataset. The EM algorithm presented for our data analysis involves a model where observed data w_i, $i = 1, \dots, 10000$ are sums of two latent variables Z_{i1} (signal) and Z_{i2} (background), each following a Poisson distribution. The expected values of these distributions are given by functions λ_{i1} and λ_{i2}, structured as GAM with the formula

$$\lambda_{ij} = \exp(\alpha_0 + s_{j1}(x_{1i}) + s_{j2}(x_{2i})), \quad j = 1, 2,$$

where $s_{jl}(x)$, $l = 1, 2$ are smooth functions.

The EM algorithm's E-step calculates the conditional expectation of Z_{ij} given y, and the M-step maximizes the conditional expectation of the log-likelihood. Specifically,

$$Q(\theta, \theta^{(k)}) = \sum_{i=1}^{n} (\mathbb{E}_{\theta^{(k)}}[Z_{i1} \mid y] \log \lambda_{i1} - \lambda_{i1} + \mathbb{E}_{\theta^{(k)}}[Z_{i2} \mid y] \log \lambda_{i2} - \lambda_{i2})$$

where

$$\mathbb{E}_{\theta^{(k)}}[Z_{ij} \mid y] = w_i \times \frac{\lambda_{ij}^{(k)}}{\lambda_{i1}^{(k)} + \lambda_{i2}^{(k)}} = \lambda_{ij}^{(k+1)}.$$

Subsequently, $\lambda_{ij}^{(k+1)}$ is updated to the expected value of Z_{ij} in the M-step. The parameters of the GAM are optimized by fitting separate GAMs for each j using these updated values. The process iterates until parameter convergence. This implementation is a generalized EM algorithm due to the lack of closed-form solutions for the parameters. The algorithm requires an initialization of $\lambda_{ij}^{(0)}$ and employs iterative computation to achieve stability.

Analysis

The primary objective is to ascertain the total signal rate. Under the assumption of a Poisson distribution, the rates from individual squares can be cumulatively summed, which is a characteristic of Poisson processes. The effectiveness of this approach is illustrated through estimates obtained from a test set with varying signal and background distributions, as shown in Table 1. The algorithm generally performs with high accuracy, though it faces challenges with a dataset 5.

Dataset	1	2	3	4	5	6
Estimate	100210	99982	98648	99353	110074	99842

Table 1. Signal Rate Estimates Across Different Datasets. The known true signal rate is 100000. The initialization parameters are the parameters of the GAM of the training set.

Analysis

Figure 3 illustrates the deviance residuals for datasets 1 and 5, highlighting the algorithm's varied performance. The left plot shows deviance residuals between 3 and -3, suggesting a generally good model fit. However, for dataset 5, a distinct pattern with residuals predominantly between -9 and 1 indicates a less adequate fit, especially for small ranges of x_1 and x_2 values.

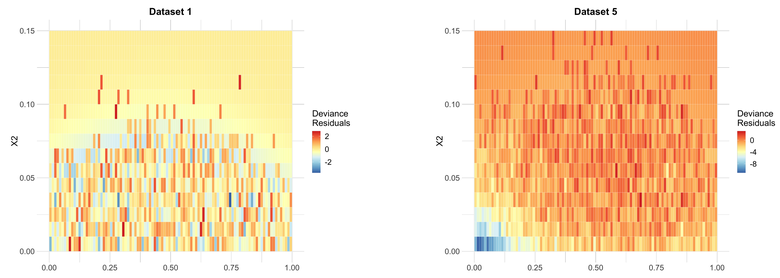

Figure 3. Heatmap of average deviance residuals from EM Algorithm on 100 simulations for each original dataset.

Figure 4 demonstrates the EM algorithm's effectiveness with small values of x_1 and x_2, while also revealing its challenges in adapting to background anomalies. This is further evidenced when comparing results from a dataset with reduced background, where the algorithm shows improved fit, underscoring its sensitivity to background variations.

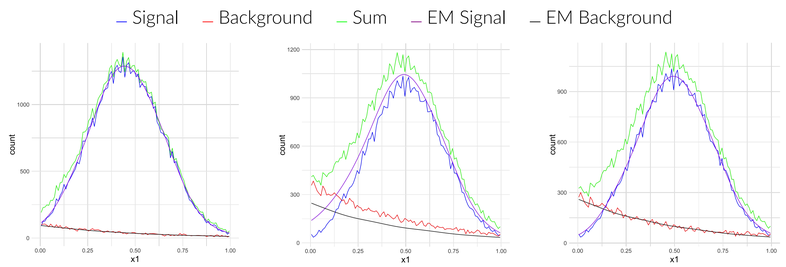

Figure 4. Analysis of the EM Algorithm applied respectively to Dataset 1, Dataset 5, and Dataset 5 with Reduced Background ($x_2 = 0.005$).

Conclusion

In this study, we successfully estimated the parameters of a generalized additive model for signal and background by observing their combined effect, using the Expectation-Maximization algorithm. A key factor in the algorithm's effectiveness is the selection of initial values, crucial for model performance. However, our findings also reveal the model's lack of robustness to background anomalies, highlighting a significant area for improvement.

References

Davison, A. C. (2020) Poisson process. Lecture Notes.

McLachlan, G. J. and Krishnan, T. (2007) *The EM Algorithm and Extensions*. Second edition. Wiley.

Wood, S. N. (2017) *Generalized Additive Models*. Second edition. Chapman and Hall/CRC.