DIANA Fellowship Proposal: Meghan Frate

I propose to develop software that allows the application of Gaussian Processes to problems in high energy physics. Gaussian processes are a mature non-parametric tool with wide applications in other fields, but little use to date in high energy physics.

Motivation
A specific application in high energy physics is the development of a background model for a data spectrum for which no analytic model is available; for example, the search for resonances in dijet events. The current state-of-the-art approach is to use a set of ad-hoc parametric functions. These functions have been chosen simply on the basis that they have historically been able to smoothly fit the data. However, as the dataset grows it becomes more difficult to find simple functions that perform well. As a result, the collection of fit functions has grown over the years as new terms are added to produce a better fit. We are now at a point where these fit function no longer have the flexibility to model the large amounts of data being collected, leaving us a choice of either piecing together new fit functions ad hoc, or finding an alternative method of background prediction.

Description
A Gaussian Processes is a way to describe all possible fit functions, but penalize them based on a notion of smoothness described by a kernel, or a covariance function. This translates to modeling the covariance between points through this kernel, rather than model the point values themselves. By modeling the smoothness and relationship between points, we have more flexibility than the traditional fit function, which will allow this to be used at higher luminosities and across multiply analyses. This method is also more physics driven than a fit function, as we're modeling our understanding of how the data should behave rather than using something that just happens to work.

Product
I have already been working on this idea for about a year, and have done a thorough proof of concept. What will need to be worked on is making this package robust enough to cover many analyses, such that one can simply pass in data or Monte Carlo, and a fit with uncertainties will be produced. This would require a substantial software package that not only provides the mathematical framework to compute and fit the Gaussian Process, but also has a set of comprehensive outputs and plots, as well as many fail safes within the code should something go wrong.

The input to this package would be a ROOT histogram of either data or Monte Carlo. This histogram is read in with rootpy and put into two arrays - bin content and bin centers, which is what is needed to run the Gaussian Process packaged george. Besides an input histogram, I would like to have a configuration file that would allow the user to set things such as fit range (if it is different from the whole histogram), luminosity, and possibly the kernel to use.

Once these inputs are set, george will need to be run to get the background estimation.

The basis of a Gaussian Process is that a set of points in a distribution can be approximated as a multivariate Gaussian. This allows the probabilities and likelihood to be calculated once a kernel function, which calculates the covariance matrix at a set of points, is set. George is able to take care of these calculations once the Gaussian Processes object is initialized with a kernel. This kernel can be built to suit the needs of the analysis, or it can be a common kernel currently programmed in george. The arrays of bin centers and statistical errors are then passed into george, and a covariance matrix will be calculated by evaluating the kernel function at the bin centers. The background fit estimation is the mean conditional distribution evaluated at the bin centers. The uncertainty on this calculation is still something that needs to be addressed. To choose the hyper parameters of the kernel function, I use Minuit minimizer to minimize the negative log likelihood between the bin content and Gaussian Process, which is calculated with george.

The next step in the standard search phase of these analyses is running the BumpHunter algorithm to search for significant excesses in the data. This is already implemented in C++ in the current analyses code, so I would need to rewrite the algorithm in python.

Finally, the output of this package needs to easily integrate with the limit setting code for each analysis. The inputs needed are the background estimation in a ROOT histogram, the uncertainty on the background estimation, and several outputs from BumpHunter. This should be easily done using rootpy. Besides the ROOT output file, there are also several output plots that should be made. In accordance with the current ATLAS standards, these must also be done in ROOT. The current dijet plotting script is written in python, so this should be straightforward to merge into my package.

I believe I can complete the programming of this package in the three months allotted from this fellowship. The usefulness of this fellowship would come in the ability to work closely with experts in the areas of Gaussian Processes and statistics, as understanding of such areas is the current bottleneck. For instance, understanding how to best build a custom kernel that fits the characteristics of the data, doesn't fit any signal excess, and is robust enough to fit various mass spectrum (i.e. diphoton, dijet, different binning, different mass ranges…) is the most important part of this project. Dialog can be started before starting the fellowship, and if the kernel function can be solidified before the beginning of the fellowship I feel confident the rest can be completed within the 3 months. After I complete the project, I will be available for minor updates for several months. Beyond that, I expect upkeep to pass to future graduate students who will be running this package for their analysis.

Location

Finally, I would like to work on this project at NYU to work with Kyle Cranmer who has been a mentor throughout this project,;Dan Foreman-Mackey who is the author of george; and Michael O'Neil who helped developed the HODLR factorization that is currently implemented in george. The ideal time for me to start this project is beginning of March, 2017.

Abstract

A specific application in high energy physics is the development of a background model for a data spectrum for which no analytic model is available; for example, the search for resonances in dijet events. The current state-of-the-art approach is to use a set of ad-hoc parametric functions. These functions have been chosen simply on the basis that they have historically been able to smoothly fit the data. However, as the dataset grows it becomes more difficult to find simple functions which perform well. We are now at a point where these fit function no longer have the flexibility to model the large amounts of data being collected, leaving us a choice of either piecing together new fit functions ad hoc, or finding an alternative method of background prediction. A Gaussian Processes is a way to describe all possible fit functions, but penalize them based on a notion of smoothness described by a kernel, or a covariance function. This translates to modeling the covariance between points through this kernel, rather than model the point values themselves. By modeling the smoothness and relationship between points, we have more flexibility than the traditional fit function, which will allow this to be used at higher luminosities and across multiply analyses.