-- BingxuanLiu - 2018-02-14

Smoothly Falling Background Modeling

Topic Overview

The goal of this page is to centrally document and discuss the methods that have been applied in Exotics searches to perform data-driven background estimations. There are two main questions to address: how to model the background and how to validate the estimates. First we review various modeling approaches with detailed descriptions and supporting documentations provided and then we will mainly discuss two widely used strategies to validate the estimations: the spurious signal method and BumpHunter.

Background Modeling Summary

This section is categorized by the object a search is dealing with, jets, photons and leptons. The underlying logic and methodologies are not very different between each categories but the technical details do vary.

Di-jet Like Resonance Searches

Di-jet like resonance searches have been performed in many experiments. Describing the QCD multi-jet background at a required precision is critical in such searches and has become more and more challenging as the integrated luminosity keeps increasing. Various methods have been proposed and demonstrated to be powerful in recent searches. In this section we will try to get a complete story of the evolution of the methods and discuss the recently developed methods in detail.

Di-jet Fit Function

The following function has been widely used in Di-jet like searches:

$$f(x) = p_1(1-x)^{p_2}(x)^{p_3+p_4 lnx+p_5(lnx)^2}, x = m_{jj}/\sqrt{s}$$

In which p_i are the free parameters. Usually it is referred to as the "x parameter dijet fit function" where x is the number of free parameters in the function. This functional form is motivated by the studies done in the parton distribution fit. In searches carried out by earlier experiments such as CDF and LHC in Run1 or early Run2, the three parameter function was determined to be sufficient. However, in latest searches with much increased luminosity, this functional form seems to reach its limit. In the next two sections we will review the ideas used to deal with this situation. We notice that the first three terms of this functional form, $p_1(1-x)^{p_2}x^{p_3}$, has not changed in the parton distribution fit performed in the past decade while various higher order terms have been probed.

Alternative Fit Functions

The fit function discussed in the previous section is constructed so that the mass distributions have expected behaviors at small or large x with a smooth transition in the intermediate region. It is however not the necessarily the only functional form that satisfies this criterion. Alternative fit functions have been used as well. For instance the following one applied by the UA2 collaboration:

$$f(x) = \frac{p_1}{x^{p_2}} e^{-p_3 x - p_4 x^2}, x = m_{jj}$$

UA2 also explored another similar function:

$$f(x) = \frac{p_1}{x^{p_2}} ln(\frac{p_3}{x}) ln(\frac{p_4}{x^2}), x = m_{jj}$$

If a given fit function fails in describing the background at a desired precision, one possible solution is to increase the number of free parameters in the function. For example, a 6-parameter Di-jet fit function with an additional $x^{p6(lnx)^3}$ term compared with the 5-parameter one. However, as pointed out by many studies, the qualitative similarity of the exponential terms cause the parameters to be strongly correlated in the fit. Therefore adding additional higher order exponential terms introduces instabilities in the fit (multiple minimals) and may not bring the needed degrees of freedom. There are many other functional forms to use such as the Bernstein polynomial and Chebyshev polynomial. Exhausting all possible functional forms is obvious not practical so a critical question for us is what functional form to use and how to determine it.

Statistical tests can be used to determine whether a fit function gives a decent description of the background. In particular, WIlk's test, Kolmogorov Smirnov (K-S) test and F (In honor of Sir Ronald A. Fisher) test are commonly used to compare two fit functions. In the context of Di-jet fit function, they have been used to determine whether additional fit parameters are needed.

Recent ATLAS searches such as Di-jet, TLA and Di-b-jet applied a Sliding Window Fit technique as will be discussed in the next section. In particular, both TLA and Di-b-jet searches are challenged by fine structures introduced by significantly increased integrated luminosity or b-tagging, which are hard to be described in an analytical way. The Sliding Window Fit can be viewed as an approximation to the actual functional form.

Sliding Window Fit (SWiFt)

The Sliding Window Fit tries to obtain the background estimate in each bin by fitting a constrained region which is referred to as a window, instead of fitting the full spectrum. Doing so mitigate the impacts brought by finer structures to the fit. On one hand, it makes it possible to describe the background given finer structures; on the other hand, if the finer structures mainly appear in part of the spectrum this makes the smoother region less affected by the problematic region so that fit functions with fewer parameters can be used.

A report was given by the analyzers from Di-b-jet and TLA teams on the comparison between global fit and SWiFt. Details on this method can also be found in the documentation here.

Two Window Fit

SWiFt performs fit in a constrained region for each bin. The idea of dividing a full spectrum to several regions to perform multiple fits also has been explored. In the Di-jet+Lepton search, the background is divided into two regions. A single fit is performed in each region to obtain the estimate. The estimates from both regions are then combined to yield the final background estimate. The procedure is documented here. A comparison between this method and SWiFt is also done where the estimates are found to be consistent. This brings up an open question: could SWiFt be generalized so that it can cover simple specific cases like this?

Two Sidebands Method (ABCD)

Two Sidebands Method is widely used in estimating the background when it is possible to construct signal-free regions via two uncorrelated variables, x and y. The data events are categorized into four regions: A, B, C and D:

$A : y > y_0, x < x_0$

$B : y > y_0, x > x_0$

$C : y < y_0, x < x_0$

$D : y < y_0, x > x_0$

One of the region, for instance region B, has the same selection as the search region while the other three regions have minimal signal contamination. If the correlations between the observable and x or y are minimal,

the event yields in region B (signal region) can be expressed as:

$$N_B = N_D * \frac{N_A}{N_C}$$

In this simplest version of the ABCD method, the assumption that neither x nor y is correlated with the observable needs to be justified. If y is correlated with the observable, one can not simply take the shape of the spectrum in region D and multiply it by a global transfer factor (Similar argument holds for x, if x is correlated with the observable, one can not simply take the shape of the spectrum in region A and multiply it by a global transfer factor). If only one of the two variables is correlated with the observable, a bin-by-bin estimation can mitigate the impact of the correlation. Instead of taking a global transfer factor, the formula above is applied in each bin of the distribution. Often the time there are correlations between all three variables. The correlation coefficients can be obtained in simulation and used to correct the formula:

$$N_B = R_{x,y} * N_D * \frac{N_A}{N_C}$$

Where $R_{x,y}$ is the correlation coefficient. Doing so makes it a quasi-data-driven method with additional systematic uncertainties introduced by the simulation. The W'->tb full hadronic analysis applies this method with top-tagging and b-tagging used as constructing variables. Details can be found here.

As one can see, the formula can be generalized as $N_B = N_D * \alpha(N_A * \alpha)$, where α is the transfer factor. If region A (D) is a subset of C, α is equivalently the efficiency for events in region C to pass the selection.

Depending on the variables used to construct the regions and the observable, it is possible to obtain α in data. Here we use the method applied in Di-b-jet search as an example. The signal region in this search requires the two jets to pass |y*| < 0.8 to suppress the t-channel QCD processes. The |y*| and b-tagging are used as one of the two variables to construct ABCD regions. The |y*| inverted regions are populated with t-channel QCD events where the jets are more in the forward region. As a result, for a event in a given mass bin, the two jets have vastly different kinematics compared with those of events in |y*| not inverted regions (nominal regions), which yields a very different event level tagging efficiency as the b-tagging performance is kinematic dependent. However, for jets in a given p_T- η bin, the b-tagging efficiency should remain the same as that in |y*| not inverted regions for a specific flavor. As a consequence, b-tagging efficiency maps can be constructed in the |y*| inverted region to be then applied in the nominal regions. The caveat is that the flavor compositions might be affected by the |y*| selection, which introduces a correlation between the efficiency maps and the |y*|. Similar to the example we have been before, some MC based studies are needed to assess this correlation. In the di-b-jet search, this method gives good agreement between the estimation and data. The estimation has even higher precision than data so that one can produce pseudo data samples to train the fit. Details of this method can be found in the documentation here.

Matrix Method and Sytem8 Method

We have seen several different applications of the ABCD method. Indeed it is demonstrated to be quite powerful. However, we also realize that the correlations between variables are hard to deal with. In the past, there were more sophisticated methods developed to deal with more complex situations such as the Matrix Method and the System8 Method. System8 Method is somehow a more complex version of the Matrix Method but we are gonna discuss teach individually.

Matrix Method

There are sophisticated methods developed to estimate the multi-jet background. The Matrix Method was developed in CDF. One can find a note describing this method. The pdf is also uploaded here in case the link breaks. The key idea is to have two regions, T(ight) and L(ight), defined by a given variable. The yields in both regions are parameterized as:

$$N_L = N_{EW} + N_{QCD}$$

$$N_T = \epsilon N_{EW} + f N_{QCD}$$

Where N_{EW} is the electroweak contribution and N_{QCD} is the QCD contribution (for some reasons people just love QCD). ϵ is the efficiency for electroweak processes while f is the fake rate for QCD processes. $N_T^{QCD} = f N_{QCD}$ is the quantity one wants to evaluate eventually, which can be solved and expressed as:

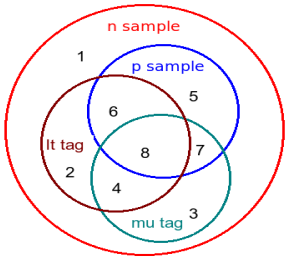
$$N_T^{QCD} = \frac{f}{\epsilon - f}(\epsilon N_L - N_T)$$

When it comes to how to obtain ϵ, f and N_L, it is very analysis dependent. Usually, the simulation of electroweak processes is precise enough to be directly used. f is often estimated in a data-driven way. Now we realize immediately that the so-called ABCD method is in fact a special case of the Matrix Method.

System8 Method

The last example given in the ABCD method section is very similar to the widely used System8 Method. It was applied in the Run1 b-tagging calibration in ATLAS and also D0 and CMS. You can also find a paper on the general implementation of this method.

In the ATLAS b-tagging case, this method construct 8 correlated samples to get a set of 8 equations. The variable of interest, the tagging efficiency of b-jets, is one of the unknown parameters in the functions. As long as there are less than 8 unknown variables, the tagging efficiency of b-jets can be obtained by solving this set of equations. The Venn Diagram is pasted here (from the ATLAS note) to illustrate the logic.



As one can see, the method applied in the di-b-jet search is a much simplified case of the System8 Method where several correlation factors were assumed to be 1. It is equivalently an efficiency measurement in data.

Resonance Searches with Photons

Methods explored by two searches are included in this section, the Di-photon resonance search and the jet-photon search. We will not go over the topics introduced in the previous section in great detail but rather focus on what is done differently.

Fit Functions Explored

In the high mass Di-photon search, a function family adapted from the di-jet function family is used:

$$f(x) = p_1(1 - x^{1/3})^{p_2}(x)^{p_3 + p_4 lnx + p_5 (lnx)^2}, x = m_{\gamma\gamma}/\sqrt{s}$$

Resonance Searches with Leptons

Sliding Window Fit (RooSwift)

RooSWift performs a sliding window fit in which a parametric signal model is utilised (typically a Breit-Wigner convoluted with a resolution model). However, it also supports a BumpHunter-like mode where window configurations are looped over and the Poisson p-value calculated for each window using the expected vs. observed yields.

A RooSwiftTutorial was given at the Exotics workshop 2018 on the use of RooSwift.

Paper Hub

Searches for Dijet Resonances at Hadron Colliders

Presentation Archive

Date	Titles	Link to Presentation	Author
Mar 11, 2017	ATLAS Exotics and SUSY Joint Workshop: Fit-Based Background Estimates	Link to slides	Karishma
Feb 23, 2018	Estimation of the dijet mass shape: global fit vs SWIFT	Link to slides	Karishma, Rui. W

This topic: Sandbox > FunctionalFormsInExoticsSearches
Topic revision: r15 - 2018-05-18 - LydiaBeresford