Research Article

A Nonparametric Peak Finder Algorithm and Its Application in Searches for New Physics

S. V. Chekanov[1] and M. Erickson[1,2]

[1] *HEP Division, Argonne National Laboratory, 9700 S. Cass Avenue, Argonne, IL 60439, USA*
[2] *Physics Department, The College of New Jersey, 2000 Pennington Road, Ewing, NJ 08628-0718, USA*

Correspondence should be addressed to S. V. Chekanov; chekanov@anl.gov

We have developed an algorithm for nonparametric fitting and extraction of statistically significant peaks in the presence of statistical and systematic uncertainties. Applications of this algorithm for analysis of high-energy collision data are discussed. In particular, we illustrate how to use this algorithm in general searches for new physics in invariant-mass spectra using pp Monte Carlo simulations.

1. Introduction

Searching for peaks in particle spectra is a task which is becoming increasingly popular at the Large-Hadron collider that focuses on new physics beyond TeV-scale. Bump searches can be performed either in single-particle (such as p_T distributions) or invariant-mass spectra. For instance, searches for new particles decaying into a two-body final state (jet-jet, gamma-gamma, etc.) and multibody decays are typically done by examining invariant masses of final-state objects (jets, leptons, missing transverse momenta, etc.). For example, assuming seven identified particles (jets, photons, electrons, muons, taus, Z-bosons, and missing p_T), a search can be made for parent particles decaying into 2, 3, or 4 daughter particles. This leads to 322 unique daughter groups. Thus, the task of analyzing such invariant-mass combinations becomes rather tedious and difficult to handle. Considering a "blind" analysis technique for scanning many channels [1], any cut variation increases the number of channels that need to be investigated. Finally, similar challenges exist for automatic searches for new hadronic resonances combining tracks [2].

The task of finding bumps is ultimately related to the task of determining a correct background shape using theoretical or known cross sections. However, a theory can be rather uncertain in the regions of interest, difficult to use for background simulation, or entirely nonexistent. Even for a simple jet-jet invariant mass, finding an analytical background function that fits the QCD-driven background spanning many orders in magnitude and which can be used to extract possible excess of events due to new physics requires a careful examination. Attempts to fit two-jet and three-jet invariant masses have been discussed in CMS [3, 4] and ATLAS [5] papers; while both experiments have reached the necessary precision for such fits using initial low statistical data, the used analytical functions are rather different and have many free parameters. This task becomes even more difficult considering multiple channels (invariant-mass distributions) with various cuts or detector-selection criteria (like b-tagging). Each such channel requires a careful selection of analytical functions for background fit and adjustments of their initial values for convergence of a nonlinear regression while determining an expectantly smooth background shape. A fully automated approach to searches for new physics has been discussed elsewhere [6].

One technically attractive approach is to find a nonparametric way to extract statistically significant peaks without *a priori* assumptions on background shapes. Such approach is popular in many areas, from image processing to studies of financial market, where a typical peak-identification task

is reduced to data smoothing in order to create a function that attempts to approximate the background. The smoothing can be achieved using the moving average [7], Lowess [8], and Splines [9] algorithms. Statistically significant deviations from smoothed distributions can be considered as peaks. Such technique is certainly adequate for the peak extraction, but it does not pursue the goal of peak identification with a correct treatment of statistical (or systematic) uncertainties. The later can be asymmetric.

The closest peak-search approach for high-energy-physics applications has been developed for studies of γ-ray spectra where the usual features of interest are the energies and intensities of photo-peaks. Several techniques have been developed, such as those based on least squares [10], second differences with least-squares fitting [11], the Fourier transformation [12], Markov chain [13], and convolution [14], (just naming a few). While such approaches are well suited for counting-type observables, they typically focus on narrow peaks on top of small and often flat-shaped background.

For example, the ROOT analysis framework [15] used in high-energy physics contains the TSPECTRUM package based on a smoothing method developed for γ-ray spectra [14]. The latter typically have narrow peaks on a smooth background. This algorithm is efficient in finding sharp peaks, while detection of wide peaks requires a visual examination of data to adjust several free parameters of this tool. Thus this approach is not well suited for a completely automatic peak search. In addition, systematic uncertainties on data points are not easy to incorporate in this approach.

In high-energy collisions, a typical standard-model background distribution has a falling shape spanning many orders of magnitude in event counts. A typical example is jet-jet invariant masses used for new particle searches [3, 5]. For such spectra, the most interesting regions are the tails of the exponentially suppressed distributions where a new high-p_T physics may show up. This means that there should be rather different thresholds to statistical noise, depending on the phase-space region, and, as result, a correct treatment of statistical and systematic uncertainties is obligatory. Unlike the γ-ray spectra where peaks are rather common and subject of various classification techniques, peaks in high-energy collisions are rather rare. As a consequence, relatively little progress has been made to develop a nonparametric fitting technique for high-energy physics applications where an observation of peaks is typically a subject for searches for new physics rather than for peak-classification purposes.

The above discussion leads to the need for a nonparametric way of background estimation together with the peak extraction mechanism which can be suited for high-energy collision distributions, such as invariant masses. The algorithm should be able to take into account the discrete nature of input distributions with their uncertainties. The proposed algorithm is less ambiguous compared to the smoothing methods (such as that used in ROOT [14, 15]), since it uses only one free parameter. In addition, it can take into account systematic uncertainties on data points (that can be asymmetric) and thus can estimate statistical significance of possible peaks in the presence of systematic uncertainties.

2. Nonparametric Peak Finder Algorithm

Due to the reasons discussed above, the program called Nonparametric Peak Finder (NPFinder) was developed using a numerical, iterative approach to detect statistically significant peaks in event-counting distributions. In short, NPFinder iterates through bins of input histograms and, using only one sensitivity parameter, determines the location and statistical significance of possible peaks. Unlike the known smoothing algorithms, the main focus of this method is not how to smooth data and then extract peaks, but rather how to extract peaks by comparing neighboring points and then calling what is left over the "background." Below we discuss the major elements of this algorithm and then we illustrate and discuss its limitations and possible improvements.

For each point i in a histogram, the first-order derivative α_i is found taking into account possible (statistical or/and systematic) uncertainties. This is done by calculating the slope between two points including their experimental uncertainties: if point $i + 1$ is lower than point i, the upper error is used, while if point $i + 1$ is higher than point i, the lower error uncertainty is used. This is done in order to be always on a conservative side while reducing statistical noise. Mathematically, this can be written as

$$\alpha_i = \frac{y_{i+1} + \delta y_{i+1} - y_i}{x_{i+1} - x_i}, \tag{1}$$

where the uncertainty δy_{i+1} is taken with negative sign for $y_{i+1} > y_i$ and with positive sign otherwise. The uncertainty may not need to be symmetric, but for simplicity we assume that they are symmetric as this is usually the case for statistical nature of uncertainties. The derivatives are averaged calculating a running average for any given position N:

$$\overline{\alpha}_N = \frac{1}{N} \sum_{i=0}^{N} \alpha_i. \tag{2}$$

The algorithm triggers the beginning of a peak if the local derivatives satisfy

$$\delta\alpha_{N+1} = \alpha_{N+1} - \overline{\alpha}_N > \Delta,$$
$$\delta\alpha_{N+2} = \alpha_{N+2} - \overline{\alpha}_N > \Delta, \tag{3}$$

where Δ is a free positive parameter that reflects (unknown) slope of the peak. This parameter should be found empirically and we will discuss below a possible range for its value. When the above conditions are true, NPFinder registers a possible peak and begins classifying next points as a part of the peak. The running average equation (2) is not accumulated for the points which belong to a possible peak. Δ is the only free parameter which specifies the sensitivity to the peak finding. This parameter should decrease with increase of sensitivity to the peaks (and likely will increase sensitivity to statistical fluctuations).

NPFinder continues to walk over data points until $\delta\alpha_{N+1}$ and $\delta\alpha_{N+2}$ are both negative, which signifies that the maximum of the peak has been reached. The double condition in (3) is used to reinforce the peak-search robustness. When
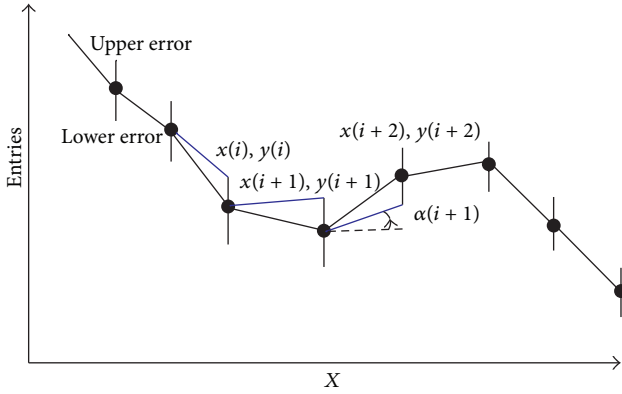
FIGURE 1: A graphical illustration of the NPFinder algorithm. Each data point is characterized by a coordinate (x_i, y_i), with (optional) upper and lower uncertainties on the y_i values. See (1) for the definition of the slopes α_i.
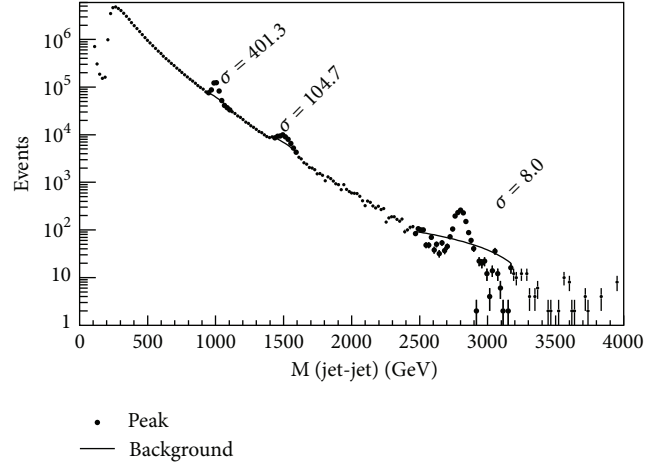

- • Peak
- — Background

FIGURE 2: Invariant mass of two jets generated with the PYTHIA Monte Carlo model. Several peaks seen in this figure were added using the Gaussian distributions with different widths and peak values (see the text). The peaks are found using the NPFinder algorithm which also estimates their statistical significance values as discussed in the text.

this condition is met, NPFinder exits the peak and adds an equal number of points to the right side from the peak center. The requirement of having the same number of points implies that the peaks are expected to be symmetric, which is the most common case. For steeply falling distributions, such as transverse-momenta spectra or dijet mass distributions, this assumption usually means that we somewhat underestimate the peak significance. Figure 1 illustrates the NPFinder algorithm for a falling invariant-mass distribution. Each point of the distribution can have an upper and lower statistical (or systematic) uncertainty.

After detecting all peak candidates, NPFinder iterates through the list of possible peaks in order to form a background for each peak. This is achieved by performing a linear regression of points between the first and last points in the peak, that is, applying the function $y_i = mx_i + b$, where m and b are the slope and intercept of the linear regression, which in this case is rather trivial as it is performed via the two points only. It should be noted that the linear regression is also performed taking into account uncertainties:

$$m = \frac{(y_2 + \delta y_2) - (y_1 + \delta y_1)}{x_2 - x_1}, \tag{4}$$

where y_1 is the first point of the peak, y_2 is the final point of the peak, and δy_1 and δy_2 are their statistical uncertainties, respectively. Here the statistical uncertainties are added in order to always be on the conservative side in estimation of the background level under the peak. The intercept parameter then is $b = y_1 + \delta y_1 - mx_1$.

It should be mentioned that the technique of the peak finding considered above is somewhat similar to that discussed for γ-ray applications [11]. But there are several important differences of NPFinder compared to this algorithm: NPFinder can detect peaks of arbitrary shapes (not only Gaussian-shaped peaks as in [11]), no fitting or smoothing procedure is used, and statistical and systematic uncertainties for data points are included during the peak-finding procedure. The algorithm [11] was not tested since its source code is not publicly available.

Finally, NPFinder uses the background points to calculate the statistical significance of each peak in a given histogram. This is done by summing up the differences r_i of the original points in a peak with respect to the calculated background points and then dividing this value by its own square root. For a given peak, it can be approximated by

$$\sigma = \frac{\sum r_i}{\sqrt{\sum r_i}}, \tag{5}$$

where the sum runs over all points in the peaks. The algorithm runs over an input histogram or graph, builds a list of peaks, and estimates their statistical significance. A typical statistically significant peak in this approach has $\sigma > 5$–7. A first peak is usually ignored as it corresponds to the kinematic peak of background distributions.

Below we illustrate the above approach by generating fully inclusive pp collision events using the PYTHIA generator [16]. The required integrated luminosity was 200 pb^{-1}. Jets are reconstructed with the anti-k_T algorithm [17] with a distance parameter of 0.6 using the cut $p_T > 100$ GeV. Then, the dijet invariant-mass distribution is calculated and the NPFinder finder is applied using the parameter $\Delta = 1$. As expected, no peaks with $\sigma > 5$ were found.

Next, a few fake peaks were generated using the Gaussian distributions with different peak positions and widths. The peaks were added to the original background histogram. Figure 2 shows an example with 3 peaks generated at 1000 GeV (20 GeV widths, 200000 events), 1500 GeV (50 GeV widths, 30000 events), and at 2800 GeV (40 GeV widths, 1200 events). The algorithm found all three peaks and gave correct estimates of their positions, widths, and approximate statistical significance using the input parameter $\Delta = 1$.

For a comparison, the same distribution was used to test the TSPECTRUM package of the ROOT program discussed in the introduction. It was found that TSPECTRUM can

also detect such peaks, but several iterations with a visual examination of the data have been required to adjust the free parameters of this algorithm, σ (an effective sigma of searched peaks), and the amplitude of the expected peaks. After the first TSpectrum pass, an additional analytic fit was required to determine the statistical significance of each peak. This approach was found to be difficult to implement in a fully automatic peak search.

It should be noted that the peak statistical significance of the proposed nonparametric method might be smaller than that calculated using more conventional approaches, such as those based on a χ^2 minimisation with appropriate background and signal functions. This can be due to the assumption on the symmetric form of the extracted peaks, the linear approximation for the background under the peaks, and the way in which the uncertainty is incorporated in the peak-significance calculation. An influence of experimental resolution can also be an issue [18], which can only be addressed via correctly identified signal and background functions. Such drawbacks are especially well seen for the highest mass peak shown in Figure 2 where a statistical fluctuation to the right of the peak pulls the background level up compared to the expected falling shape. Given the approximate nature of the statistical significance calculations which only serve to trigger attention of analyzers who need to study the found peaks in more detail, the performance of the algorithm was found to be reasonable.

It should be noted that there is a correlation between the peak width and the input parameter Δ: a detection of broader peaks typically requires a smaller value of Δ (which can be as low as 0.2).

In conclusion, a peak-detection algorithm has been developed which can be used for extraction of statistically significant peaks in event-counting distributions taking into account statistical (and potentially systematic) uncertainties. The method can be used for new physics searches in high-energy particle experiments where a correct treatment of such uncertainties is one of the most important issues. The nonparametric peak finder has only one free parameter which is fairly independent of input background distributions. The algorithm was tested and found to perform well. The code is implemented in the Python programming language with the graphical output using either ROOT (C++) [15] or SCaVis (Java) [19]. The code example is available for download [20].
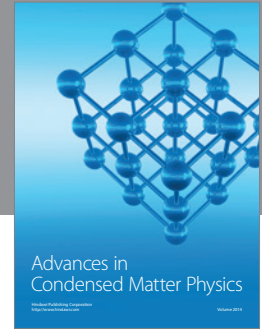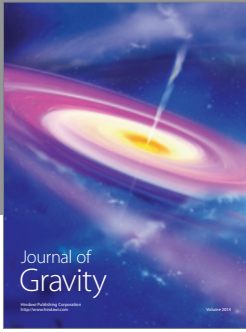
Acknowledgments

References

[1] G. Choudalakis, *Model independent search for new physics at the tevatron [Ph.D. thesis]*, 2008, http://arxiv.org/abs/0805.3954.

[2] S. Chekanov, "A C++ framework for automatic search and identification of resonances," in *Proceedings of the HERA and the LHC: A Workshop on the Implications of HERA for LHC Physics, Part B*, CERN-2005-014, DESYPROC-2005-01, p. 624, 2005.

[3] V. Khachatryan, A. M. Sirunyan, A. Tumasyan et al., "Search for dijet resonances in 7 TeV pp collisions at CMS," *Physical Review Letters*, vol. 105, no. 21, Article ID 211801, 14 pages, 2010.

[4] S. Chatrchyan, V. Khachatryan, A. M. Sirunyan et al., "Search for three-jet resonances in pp collisions at $\sqrt{s}$ = 7 TeV," *Physical Review Letters*, vol. 107, no. 10, Article ID 101801, 15 pages, 2011.

[5] G. Aad, B. Abbott, J. Abdallah et al., "A search for new physics in dijet mass and angular distributions in pp collisions at $\sqrt{s}$ = 7 TeV measured with the ATLAS detector," *New Journal of Physics*, vol. 13, no. 5, Article ID 053044, 2011.

[6] T. Aaltonen, J. Adelman, T. Akimoto et al., "Global search for new physics with 2.0 fb^{-1} at CDF," *Physical Review D*, vol. 79, no. 1, Article ID 011101(R), 9 pages, 2009.

[7] E. S. Kenney, *Mathematics of Statistics*, Part 1, 3rd edition, 1962.

[8] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, vol. 74, no. 368, pp. 829–836, 1979.

[9] J. Schoenberg, "Contributions to the problem of approximation of equidistant data by analytic functions, parts A and B," *Quarterly of Applied Mathematics*, vol. 4, pp. 45–99, 112–141, 1946.

[10] F. C. P. Huang, C. H. Osman, and T. R. Ophel, "Line shape analysis of spectra obtained with Ge(Li) detectors," *Nuclear Instruments and Methods*, vol. 68, no. 1, pp. 141–148, 1969.

[11] M. Mariscotti, "A method for automatic identification of peaks in the presence of background and its application to spectrum analysis," *Nuclear Instruments and Methods*, vol. 50, no. 2, pp. 309–320, 1967.

[12] K. J. Blinowska and E. F. Wessner, "A method of on-line spectra evaluation by means of a small computer employing Fourier transforms," *Nuclear Instruments and Methods*, vol. 118, no. 2, pp. 597–604, 1974.

[13] Z. K. Silagadze, "A new algorithm for automatic photopeak searches," *Nuclear Instruments and Methods in Physics Research A*, vol. 376, no. 3, pp. 451–454, 1996.

[14] M. Morháč, J. Kliman, V. Matoušek, M. Veselský, and I. Turzo, "Identification of peaks in multidimensional coincidence γ-ray spectra," *Nuclear Instruments and Methods in Physics Research A*, vol. 443, no. 1, pp. 108–125, 2000.

[15] I. Antcheva, M. Ballintijn, B. Bellenot et al., "ROOT—a C++ framework for petabyte data storage, statistical analysis and visualization," *Computer Physics Communications*, vol. 180, no. 12, pp. 2499–2512, 2009.

[16] T. Sjostrand, S. Mrenna, and P. Z. Skands, "PYTHIA 6.4 physics and manual," *Journal of High Energy Physics*, vol. 2006, no. 5, article 26, 2006.

[17] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-kt jet clustering algorithm," *Journal of High Energy Physics*, vol. 2008, no. 4, article 63, 2008.

[18] S. V. Chekanov and B. B. Levchenko, "Influence of experimental resolution on the statistical significance of a signal: implication for pentaquark searches," *Physical Review D*, vol. 76, no. 7, Article ID 074025, 7 pages, 2007.

[19] S. Chekanov, *Scientific Data Analysis Using Jython Scripting and Java*, Springer, London, UK, 2010, http://jwork.org/scavis/.

[20] M. Erickson and S. Chekanov, "Non-parametric peak finder algorithm," http://atlaswww.hep.anl.gov/asc/packages/NPFinder/.

Journal of
Gravity

Journal of
Photonics

The Scientific
World Journal

Journal of
Soft Matter

Advances in
Condensed Matter Physics

Journal of
Fluids

Journal of
Aerodynamics

International Journal of
Statistical Mechanics

Hindawi

Submit your manuscripts at
http://www.hindawi.com

International Journal of
Optics

Journal of
Computational
Methods in Physics

Advances in
High Energy Physics

Journal of
Thermodynamics

Physics
Research International

Advances in
Astronomy

Journal of
Solid State Physics

Journal of
Astrophysics

International Journal of
Superconductivity

Journal of
Biophysics

Journal of
Atomic and
Molecular Physics