Event Selection Using Adaptive Gaussian Kernels

A. Askew, H. Miettinen, B. Padley
Rice University, Houston, TX 77271, USA

The Probability Density Estimation method is a technique that uses Kernel Density Estimation techniques to derive a discriminate function which an be used for event selection. This approach has the advantage of handling complex dependencies in data without using the 'black box' approach of neural networks. We present a new variant of the Probability Density Estimation method that allows the use of adaptive kernels. Studies comparing the performance of this method to that of neural networks are presented and prospects for use in physics analysis are described.

1. INTRODUCTION

Neural networks have been used in experimental particle physics analysis with increasing frequency in recent years. However the black box nature of such approaches is worrisome to some. Adaptive Probability Density Estimation was developed at Rice University as an alternative multi-variate approach which provides the flexibility of neural networks with an easily understood and visualizable method. Adaptive PDE is a method of kernel density estimation which uses the distribution of the training data to vary the width of the kernels such that outlying points may be dealt with without discontinuities in the resultant multivariate space.

2. THEORY

2.1. Ordinary Probability Density Estimation

The Probability Density Estimation (PDE) method was implemented by researchers at Rice University for the top quark search. In that analysis, the efficiency of the fixed kernel estimation was found to be similar to that of neural networks. The standard (fixed kernel) probability density estimation method of multivariate data analysis has been previously documented [1]. Unlike neural networks, this method has few free parameters, allowing for less complicated optimization. Given a training sample of data, consisting of sets of signal and background, two functions of the n input variables are formed of the data set. This is accomplished by forming a product of kernel functions in the space of the input variables for each data point. The complete function for the entire sample of events is the sum of all of these product kernels in the training data, normalized by the number of training events used to form the function. This is known as the feature function, because it is an estimate of the important features in the data. Mathematically this function is given by:

$$f(\mathbf{x}) = \frac{1}{N_{tr} h_1 \dots h_d} \sum_{i=1}^{N_{tr}} [\prod_{j=1}^{d} K(\frac{x_i - x_{ij}}{h_j})]. \quad (1)$$

The $\mathbf{x}$ are the set of variables this analysis is being performed on after transformation into a set in which correlations are zero (the new set of variables is a linear combination of the original variables). This transformation is done so that the kernel structure will match the covariance structure of the data, and thus give a better representation of the data points. Here, K is the kernel chosen to suit the data. In this analysis a Gaussian kernel has been chosen:

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}. \quad (2)$$

The remaining items in the feature function are N_{tr} the number of training events, and $h_j = h^0 \times \sigma_j$, where σ_j is the standard deviation of variable j (in the space in which the correlations vanish), d is the number of variables used in the analysis and h^0 is a tunable parameter which must be optimized for the data set. h^0 is chosen such that the functions formed by the sum of the product kernels are smooth representations without losing information about the data.

Using a linearly independent set of testing data, and the feature functions (f_s, f_b), a discriminant function value ($D(\mathbf{x})$) can be found for each event in the testing sample, representing a measure of how well the event matches the feature function for signal ($D(\mathbf{x})$ close to one) or background($D(\mathbf{x})$ close to zero). This discriminant function is simply;

$$D(\mathbf{x}) = \frac{f_s}{f_s + f_b}, \quad (3)$$

where f_s is the feature function for the signal and f_b is the signal function for the background. Then a single cut on the value of $D(\mathbf{x})$ can be made, selecting the signal from the background by the value calculated for each event.

2.2. Adaptive Probability Density Estimation

The PDE adaptive kernel builds on this method with one modification. An additional parameter α is used to further fit the gaussian kernels to the data set. A pilot $f(x)$ is found using h_j and then the analysis is performed using;

$$h_j = h^0 \times \sigma_j \times \left(\frac{f_{pilot}(\overline{x})}{f_{pilot}(x)} \right)^{\alpha}, \qquad (4)$$

where α is another free parameter which must be optimized along with h^0 for the data set, f_{pilot} is the feature function formed using h_j=$h^0 \times \sigma_j$, and $\overline{x}$ is a vector of the mean values of each of the variables chosen for the analysis. This new choice of the width of the product kernels ties the functional form closer to the actual distribution of the data. For sets of data which may have a few outlying events, this now selects a wider gaussian, so that the feature function is smoother for these areas.

2.3. Implementation

The PDE method, as used in this analysis is almost exactly as described in the above section. The only approximation made in the use of the technique is in the method used to diagonalize the covariance matrix of signal and background. For this procedure, a set of Jacobian rotations of the covariance matrices are made to find the eigenvalues and eigenvectors. These rotations are made until the sum of the absolute values of the off diagonal elements of the matrix are sufficiently small. The floating point precision of the machine determines how small the sum of the off diagonal elements must become. For a detailed description of these rotations and the algorithms the reader is referred to [2]. Once the eigenvalues are found, then one can form a transformation matrix which can be used to rotate the covariance matrices such that the correlations in the signal and background covariance matrices vanish. It can be shown that the matrix that achieves this transformation is;

$$A = v^T \times M, \qquad (5)$$

$$M = (U\Lambda^{-\frac{1}{2}}U^T), \qquad (6)$$

where U is the matrix of the eigenvectors of σ_b (the background covariance matrix), $\Lambda^{-\frac{1}{2}}$ is a diagonal matrix with one over the square root of the eigenvalues of σ_b, and v is the matrix of the eigenvectors of $M \times \sigma_b \times M$. As a consequence of this rotation, the diagonal elements of the background covariance matrix become ones. A proof of how this rotation may be performed can be found elsewhere [1]. This rotation may then be performed on the inputs, which results in the PDE analysis being performed in a space that is a linear combination of the input variables. The PDE algorithms have been implemented in C++ code, and prepared in a shared library form for use in ROOT (object oriented data analysis framework). In this analysis ROOT version 2.25/00 was used.

3. OPTIMIZATION

Using particle ID Monte Carlo courtesy of the D0 experiment, we examine the case of identification of tau leptons with background from QCD multijet production. The optimization of the two PDE methods is detailed.

3.1. Fixed Kernel Optimization

For the fixed kernel PDE method, there is only one parameter to be determined for each data set, h^0. To determine the optimum value for this parameter, the analysis was performed a number of different times using all of the values for h^0 between (0,1] in increments of 0.05. At each value of h^0 the purity times signal efficiency (for a discriminant value of $D(\mathbf{x}) = .5$) was computed, and then the purity times signal efficiency versus the value of h^0 was graphed. The maximum purity times signal efficiency was taken to be the optimum value of h^0 for that data set. The graph is shown below (Figure 1). For the test case of discriminating tau leptons from QCD background, a value of $h^0 = .55$ was found.

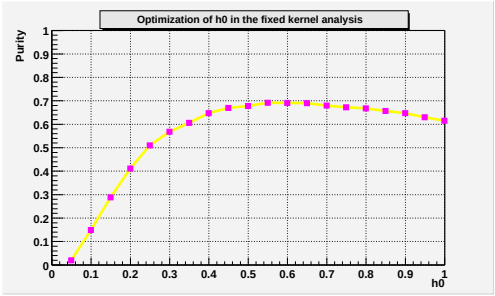

Figure 1: Optimization of h^0 parameter for fixed kernel analysis of τs

3.2. Adaptive Kernel Optimization

The PDE adaptive kernel has two parameters h^0 and α. The optimization of the performance for these two parameters was carried out in a similar way to the fixed kernel. In this case instead of a linear graph, a surface in the space of h^0, α and purity times signal

efficiency was formed, in increments of 0.05 in h^0 and α (for the same discriminant function value as the fixed kernel case). The maximum of this surface was taken to correspond to the optimum values of h^0 and α for the analysis. The maximum of the surface was found to be at $(h^0 = 0.65, \alpha = 0.25)$.

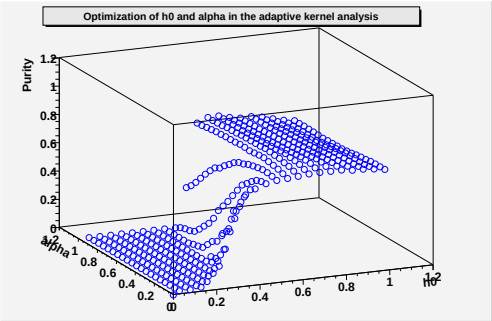

Figure 2: Optimization of h^0 and α: Purity times signal efficiency versus h^0 and α at $D = 0.5$ for adaptive kernel analysis of τs versus QCD jets

4. RESULTS

The above optimized PDE methods were compared with that of a neural network optimized on the same parameter (purity times signal efficiency) and the same signal and background training and testing sets. Figure 3 presents the resultant signal efficiency and background efficiency as a function of a cut on neural network output and PDE discriminant function. The optimum signal efficiencies as determined by the maximum signal efficiency times purity for each method are summarized in Table I.

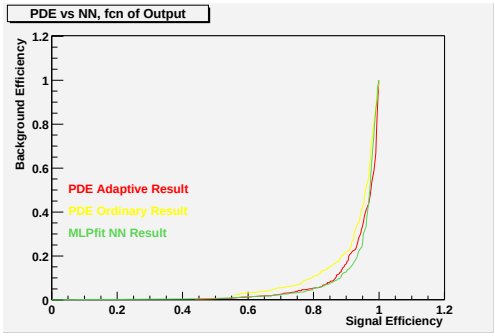

Figure 3: Signal versus Background efficiency for τ versus QCD jets

Table I Summary of Multivariate Performances

Method	Max. Purity $\times\epsilon_s$	ϵ_s	ϵ_b	$\frac{\epsilon_s}{\epsilon_b}$
τ versus QCD jets				
Fixed Kernel PDE	.731	.879	.178	4.94
Adaptive Kernel PDE	.781	.861	.0886	9.72
Neural Network	.793	.883	.100	8.83

Acknowledgments

Work supported by DOE Grant DE-FG05-97ER41031. The authors would like to thank the D0 Experiment for providing the Monte Carlo data used in this study.

References

[1] L. Holmstrom, S. Sain, H. Miettinen" A new multivariate technique for top quark search", Computer Physics Communications 88, 1995, 195-210.

[2] J. Gentle, *Numerical Linear Algebra for Applications in Statistics*. Springer-Verlag (1998).