CDF/DOC/EXOTIC/CDFR/10227
Draft 1.0
July 23, 2010

Detection of New Physics Using Density Estimation Based Anomaly Search

Mikael Kuusela[1,2], Eric Malmi[1,2], Tommi Vatanen[1,2],
Risto Orava[2], Timo Aaltonen[2], Yoshikazu Nagai[3]
[1] *Aalto University*
[2] *University of Helsinki*
[3] *University of Tsukuba*

Abstract

We introduce a model-independent anomaly search approach for detection of new physics signals in high-energy physics and study three different variants of this approach. All these methods are based on multivariate probability density estimation under the assumption of a fairly accurate background Monte Carlo (MC) model. Demonstrations using CDF MC samples for WH Higgs show that anomaly search is able to find the new physics signal without prior knowledge of its signature. As such, these methods are robust against inaccuracies in the signal MC. Comparison to a model-dependent Neural Network classifier (NN) shows that for a foreseen signal anomaly search and NN produce comparable results. On the other hand, in the case of an unexpected signal, NN fails to correctly identify the signal while anomaly search does not suffer from such a limitation.

Contents

1 Introduction

One of the main objectives of high energy physics is the search for new physics signals. Traditionally, searches for such signals are conducted with *model-dependent* classification methods, such as neural networks (NN). These methods rely greatly on training samples from Monte Carlo (MC) generators to distinguish the desired signal from the background. The obvious drawback of this approach is that it becomes useless if one does not know what to look for, or in the case that the MC generators do not model the signal events accurately.

To overcome these problems, we propose using *model-independent multivariate machine learning methods* for searching anomalies in the particle collision data. These anomaly search methods are based on probability density estimation under the assumption that there exists a fairly accurate representation for the background, i.e., the detector response to a sample containing no signal events. In most cases, the background would be defined using MC although in some situations it might be possible to use real measurements as well. The anomalies found should be investigated further in order to determine whether they result from (i) deficiency or inaccuracy in the background MC generator, (ii) a detector defect or a lack of understanding of the detector, or (iii) a previously unknown physics process. The advantage of this kind of an approach is that we are independent from the distribution of the anomalous events. Furthermore, in the case of a well-understood detector and background, anomaly search is insensitive to uncertainties in the MC models for the new physics signals.

Such model-independent approaches have previously been used at some experiments. In CMS, for example, an algorithm called Model Unspecific Search in CMS (MUSiC) [1] scans the measured data for deviations from the MC expectation. The algorithm does this by looking at single variable at a time as a results of which one may end up missing the dependency structures (i.e., correlations) in the data. Similar algorithms, namely Vista and Sleuth [2], have been devised at the Tevatron with the same limitations. On the contrary, the multivariate methods studied here inherently take correlations into account in the multi-dimensional input space.

Outside physics, anomaly detection has been successfully applied to the credit card fraud detection [3]. This problem is similar to searches for new physics as one would like to detect unforeseen anomalous credit card usage among a large background of proper transactions. Analogously, in high-energy physics, one would like to identify new signals among a background of well-known old physics.

This paper is organized as follows. In Section 2, we present the basic ideas behind density estimation based anomaly search and present two different approaches for locating the signal. Section 3 describes the technical details of the algorithms. We first present principal component analysis as a means to reduce the dimensionality of the data space followed by a description of the density estimation methods employed in this work. We demonstrate the feasibility of the anomaly search approach using CDF MC samples for WH Higgs in Section 4 and end with discussion and conclusions in Section 5 and 6.

2 Anomaly Search with Density Estimation

The task of an anomaly search is to find differences between the measured data and the expected background. The traditional approach in anomaly detection is to estimate only the background distribution and then classify an incoming event as anomalous if it is located in an area of low probability density of the background distribution. This approach, however, becomes useless if the anomalous signal lies among the background.

When the signal is among the background, an event-by-event classification is usually difficult. Nevertheless, one can detect changes in the distribution of the data; there is more events in the signal region than one would expect according to the background distribution. To analyze the changes in the distribution of the data, we utilize density estimation techniques to model the background distribution (p_{B}) from the MC generated events and the actual distribution (p_{M}) from the measured data. Then we compare these two distributions to identify the signal distribution p_{S} which represents the unexpected data.

The measured data distribution p_{M} is assumed to be a linear combination of the background and the signal

$$p_{\mathrm{M}}(\boldsymbol{x}) = (1-\lambda)p_{\mathrm{B}}(\boldsymbol{x}) + \lambda p_{\mathrm{S}}(\boldsymbol{x}), \tag{1}$$

where λ is proportional to the cross section of the signal. We further assume that p_{B}, which is estimated from the MC, is accurate and thus the deviations from the MC are represented by p_{S}.

For an event-by-event classification a discriminant function D is needed. We choose the probability of event $\boldsymbol{x}$ to belong to signal as the discriminant

$$D(\boldsymbol{x}) = \frac{\lambda p_{\mathrm{S}}(\boldsymbol{x})}{(1-\lambda)p_{\mathrm{B}}(\boldsymbol{x}) + \lambda p_{\mathrm{S}}(\boldsymbol{x})}. \tag{2}$$

The decision rule for selecting events is as follows

$$D(\boldsymbol{x}) = \begin{cases} D(\boldsymbol{x}) \geq T \;\Rightarrow\; \boldsymbol{x} \text{ accepted}, \\ D(\boldsymbol{x}) < T \;\Rightarrow\; \boldsymbol{x} \text{ rejected}, \end{cases} \tag{3}$$

where T is a constant threshold which can be used to control the sensitivity of the classifier. As extreme cases, if $T = 0$ all events are accepted, and if $T = 1$ all events are rejected.

We propose two different approaches for revealing the signal distribution, namely (i) fixed background model and (ii) background subtraction. The first and a common step for both approaches is background modeling. After estimating the background distribution $p_{\mathrm{B}}(\boldsymbol{x})$, *fixed background model* tries to find the signal distribution $p_{\mathrm{S}}(\boldsymbol{x})$ that in combination with the background model $p_{\mathrm{M}}(\boldsymbol{x})$ gives the maximum likelihood for the signal and background data (signal+background). In *background subtraction* the second step is to estimate the signal+background distribution after which the two distributions can be subtracted resulting in a candidate for the discriminant function. The two approaches are described in more detail below.

2.1 Fixed Background Model

An underlying assumption in fixed background modeling is that one has a relatively representative MC sample for the background. In terms of particle physics this usually corresponds to a precise MC model. Figure 1a illustrates a one dimensional background sample and a maximum likelihood Gaussian distribution $p_{\mathrm{B}}(\boldsymbol{x})$ estimated using the sample.

After modeling the background distribution one starts estimating the measured data model $p_{\mathrm{M}}(\boldsymbol{x})$ (see equation (1)) in such a way that the shape of the background distribution $p_{\mathrm{B}}(\boldsymbol{x})$ is fixed while the other parameters in $p_{\mathrm{M}}(\boldsymbol{x})$ (i.e., λ and the parameters of $p_{\mathrm{S}}(\boldsymbol{x})$) are varied. The lines in the bottom of Figure 1b illustrate a signal+background sample where the longer lines correspond to a weak signal. Here, the signal model $p_{\mathrm{S}}(\boldsymbol{x})$ is a one dimensional Gaussian with parameters μ and σ. Hence, one is optimizing the values of λ, μ and σ to give the maximum likelihood for the measured signal+background data. The resulting distribution $p_{\mathrm{M}}(\boldsymbol{x})$ is shown with a solid line in Figure 1b.

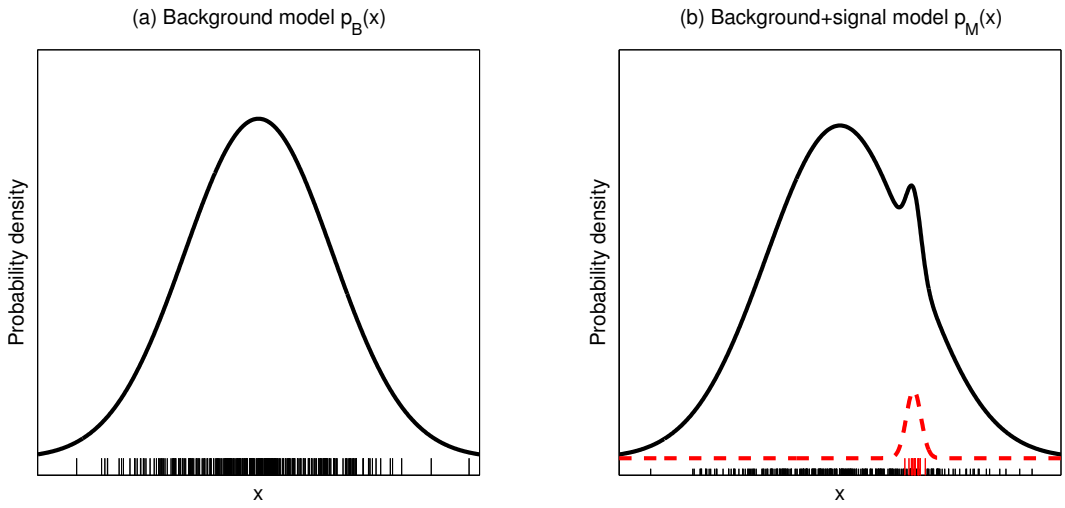

Figure 1: An illustration of the fixed background model in a one dimensional space. Figures show (a) the background sample (lines at the bottom) and an estimated distribution for the sample, and (b) the signal+background sample (longer lines at the bottom denote the anomalous signal) and an estimated signal distribution $p_{\mathrm{S}}(\boldsymbol{x})$ (dashed line). The resulting distribution for the measured data $p_{\mathrm{M}}(\boldsymbol{x})$ is show with a solid line in (b).

2.2 Background Subtraction

Background subtraction is based on calculating the difference of the background and the measured distributions. This has been done for one dimensional data using histograms as density estimators in [4]. The term background subtraction is also used

in vision systems research where the problem is to detect moving objects from static cameras [5, 6]. This problem has many similarities to that of finding new physics signals. However, the difference is that vision systems research deals with a time series of consecutive images of the same object and therefore the methods used are usually somewhat different.

We conduct the background subtraction by solving p_{S} from equation (1)

$$p_{\mathrm{S}}(\boldsymbol{x}) = \frac{p_{\mathrm{M}}(\boldsymbol{x}) - (1-\lambda)p_{\mathrm{B}}(\boldsymbol{x})}{\lambda}. \tag{4}$$

This result implies that we need to have an estimate for the cross section λ in order to estimate the signal distribution p_{S}. Nevertheless, if we only want to do an event-by-event classification, it turns out that a prior knowledge of the cross section is not necessarily needed. Substituting p_{S} from equation (4) into equation (2) we obtain the decision rule

$$D(\boldsymbol{x}) = 1 - (1-\lambda)\frac{p_{\mathrm{B}}(\boldsymbol{x})}{p_{\mathrm{M}}(\boldsymbol{x})} > T, \tag{5}$$

where $D(\mathbf{x})$ is again the discriminant function and T the classification threshold. After rearrangements we get

$$\frac{p_{\mathrm{B}}(\boldsymbol{x})}{p_{\mathrm{M}}(\boldsymbol{x})} < \frac{T-1}{1-\lambda} = T_{\mathrm{final}}. \tag{6}$$

This means that λ can be treated as a part of the threshold T_{final} and is therefore discarded from the decision rule.

3 Methodological background

Our data-driven techniques use multivariate probability density functions (PDF) to summarize the data. When new measurements are collected, their similarity with the known background distribution can be both quantitatively and qualitatively evaluated. A limitation of this approach is that one has to rely on the background predictions made by the MC generators.

If the dimensionality of the data increases, the number of events required for an accurate density estimation grows exponentially. This is known as *the curse of dimensionality* [7]. In the following subsections, we will describe a dimensionality reduction method for dealing with the curse of dimensionality. Then we will shortly introduce some advanced density estimation techniques that produce continuous probability distributions.

3.1 Dimensionality Reduction with Principal Component Analysis (PCA)

To tackle the curse of dimensionality, we conduct a dimensionality reduction using principal component analysis (PCA). In dimensionality reduction, the task is to find

a mapping from the original D-dimensional space to a d-dimensional subspace where $d < D$. The mapping is such that a minimum amount of information is lost. In addition to dimensionality reduction, PCA is widely used for other applications such as feature extraction, lossy data compression and data visualization [8].

Probably the most commonly used definition of PCA is the maximum variance formulation [9] according to which PCA is an orthogonal projection of the data onto a lower dimensional linear space, *principal subspace*, in which the variance of the data is maximized, that is, the maximum amount of information is preserved. The optimal projection onto the d-dimensional subspace is such that we choose d eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$ of the data covariance matrix $\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n - \bar{\boldsymbol{x}})(\boldsymbol{x}_n - \bar{\boldsymbol{x}})^{\mathrm{T}}$ corresponding to the d largest eigenvalues $\lambda_1, \ldots, \lambda_d$. Now, a linear transformation of a data vector $\boldsymbol{x}_n$ onto the principal subspace defined by the d eigenvectors is simply the product

$$\boldsymbol{z}_n = \boldsymbol{U}^{\mathrm{T}}(\boldsymbol{x}_n - \bar{\boldsymbol{x}}), \tag{7}$$

where $\boldsymbol{z}_n$ are called the z-scores for the data vector and the d columns of $\boldsymbol{U}$ are the d leading unit eigenvectors of $\boldsymbol{S}$. A detailed description and derivation of PCA can be found, e.g., in [8, 10].

3.2 Parametric Density Estimation

3.2.1 Mixture of Multivariate Gaussian Distributions

The objective of probabilistic modeling is to approximate the data set with some known probability distribution. In other words, the modeling task is to estimate an unknown probability distribution based on a finite number of observations. The underlying assumption is that the data is drawn from some unknown but well-defined distribution and the task is to estimate the parameters of the distribution. That is why these methods are sometimes called *parametric methods*. The advantage of this approach is that the model can be defined with a small number of parameters, e.g., mean and covariance matrix in the case of a Gaussian distribution. The parameters of the distribution are estimated from the data using *maximum likelihood* (ML) estimation, i.e., we select such a model that it maximizes the likelihood of the data.

Finite mixtures of distributions are a flexible method for modeling complex distributions [11]. The idea of a mixture model is that its components can represent different parts of the true distribution, which would be difficult or impossible to estimate by a single parametric distribution. In this work, we use mixtures of multivariate Gaussian distributions [10] or shortly mixtures of Gaussians (MoG) to represent the distribution of the measurements from particle collisions.

After the dimension reduction (see Chapter 3.1) the particle collision events can be represented by d-dimensional vectors $\boldsymbol{x}$, where d is the dimensionality of the subspace. For a single multivariate Gaussian distribution, the PDF value of the observed vector $\boldsymbol{x}$ is

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\Big(-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\Big) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{8}$$

Now, the finite mixture of multivariate Gaussian distributions is defined by

$$p(\boldsymbol{x}|\Theta) = \sum_{j=1}^{J} \pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \tag{9}$$

where π_j are mixture proportions (or mixing coefficients) such that $\pi_j \geq 0$ and $\sum_{j=1}^{J} \pi_j = 1$ and $\Theta = \{J, \{\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^{J}\}$ represents the parameters of the mixture model with J components.

3.2.2 Expectation-Maximization (EM) Algorithm with Fixed Background

In this section, we outline how to use the *expectation-maximization* (EM) algorithm to estimate models of the form (1) when the shape of the background distribution is fixed. Let us first consider the case of fitting a MoG model with J components to the background sample with N observations $\boldsymbol{x}_n, n = 1 \ldots N$. The log-likelihood of the parameters $\{\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^{J}$ can be written as

$$l = \sum_{n=1}^{N} \log \left(\sum_{j=1}^{J} \pi_j \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right). \tag{10}$$

Here we have assumed that the collision events are independent and identically distributed (i.i.d.). The ML estimates of the parameters can be obtained by maximizing (10) which is carried out by using the EM algorithm [12, 13]. The detailed derivation of the EM algorithm for the MoG model can be found in [10]. Next, we show the update equations of the parameters. In the expectation step (E-step), the posterior probabilities

$$p(j|\boldsymbol{x}_n, \Theta^k) = \frac{\pi_j^k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_j^k, \boldsymbol{\Sigma}_j^k)}{\sum_{j'=1}^{J} \pi_{j'}^k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_{j'}^k \boldsymbol{\Sigma}_{j'}^k)} = p(z_{nj} = 1|\boldsymbol{x}_n) \equiv \gamma(z_{nj}) \tag{11}$$

are calculated. Here, Θ^k contains the parameter estimates at the iteration k and $\boldsymbol{z}_n$ is a J-dimensional binary variable having 1-of-J representation in which a particular element z_{nj} is equal to 1 and all other elements are equal to 0. The vector $\boldsymbol{z}_n$ can be interpreted as an explicit latent variable describing which component of the mixture model generates the sample n. Equation (11) gives the posterior probability that data point $\boldsymbol{x}_n$ is generated by the j^{th} component. In the maximization step (M-step), the parameter values are updated according to following equations

$$\pi_j^{k+1} = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nj}), \tag{12}$$

$$\boldsymbol{\mu}_j^{k+1} = \frac{\sum_{n=1}^{N} \gamma(z_{nj}) \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma(z_{nj})}, \tag{13}$$

$$\boldsymbol{\Sigma}_j^{k+1} = \frac{\sum_{n=1}^{N} \gamma(z_{nj})(\boldsymbol{x}_n - \boldsymbol{\mu}_j^{k+1})(\boldsymbol{x}_n - \boldsymbol{\mu}_j^{k+1})^{\mathrm{T}}}{\sum_{n=1}^{N} \gamma(z_{nj})} \tag{14}$$

It has been shown, that each iteration of the EM algorithm increases the log-likelihood of the data until a local maximum of the likelihood is found [13].

Secondly, after estimating the distribution of the background sample, we utilize the EM algorithm to search any unmodeled anomalies in the collision data. Now, the first term, $p_{\mathrm{B}}(\boldsymbol{x})$ in equation (1) is fixed and both λ and the parameters of $p_{\mathrm{S}}(\boldsymbol{x})$ are optimized to maximize the log-likelihood of the data. Here, $p_{\mathrm{S}}(\boldsymbol{x})$ can be either a single Gaussian or a MoG. We can now write equation (1) as follows

$$\begin{aligned} p_{\mathrm{M}}(\boldsymbol{x}) &= (1-\lambda)p_{\mathrm{B}}(\boldsymbol{x}) + \lambda \sum_{q=J+1}^{J+Q} \tilde{\pi}_q \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \\ &= \pi_B p_{\mathrm{B}}(\boldsymbol{x}) + \sum_{q=J+1}^{J+Q} \pi_q \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \end{aligned} \tag{15}$$

where the latter term is another MoG with Q components representing the anomalous signal and $\pi_B = \pi_{J+Q+1}$ is the mixture proportion of the background mixture model $p_{\mathrm{B}}(\boldsymbol{x})$. The mixture proportions of this MoG satisfy $\sum_{q=J+1}^{J+Q+1} \pi_q = 1$, $\sum_{q=J+1}^{J+Q} \pi_q = \sum_{q=J+1}^{J+Q} \lambda\tilde{\pi}_q = \lambda$ and $\pi_B = 1 - \lambda$.

By straightforward analogy to standard EM, the update equations of the EM algorithm in the case of model (15) are as follows. In the E-step we update the posterior probabilities of the background model and the components of the signal MoG as follows

$$\begin{aligned} p(B|\boldsymbol{x}_n, \Theta^k) &= \frac{\pi_B^k p_{\mathrm{B}}(\boldsymbol{x})}{\pi_B^k p_{\mathrm{B}}(\boldsymbol{x}) + \sum_{q'=J+1}^{J+Q} \pi_{q'}^k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{q'}^k, \boldsymbol{\Sigma}_{q'}^k)} \\ &= p(z_{nB} = 1|\boldsymbol{x}_n) \equiv \gamma(z_{nB}), \end{aligned} \tag{16}$$

$$\begin{aligned} p(q|\boldsymbol{x}_n, \Theta^k) &= \frac{\pi_q^k \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_q^k, \boldsymbol{\Sigma}_q^k)}{\pi_B^k p_{\mathrm{B}}(\boldsymbol{x}) + \sum_{q'=J+1}^{J+Q} \pi_{q'}^k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_{q'}^k, \boldsymbol{\Sigma}_{q'}^k)} \\ &= p(z_{nq} = 1|\boldsymbol{x}_n) \equiv \gamma(z_{nq}). \end{aligned} \tag{17}$$

In the first equation, $z_{nB} = 1$ denotes that the n^{th} data vector was generated by the background model $p_{\mathrm{B}}(\boldsymbol{x})$. In the second equation $q = J+1, \ldots, J+Q$. The corresponding M-step updates the means and covariance matrices of the signal MoG together with mixture proportions of the background model (π_B) and the signal MoG components with the following equations

$$\pi_q^{k+1} = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nq}), \quad q = J+1, \ldots, J+Q+1, \tag{18}$$

$$\boldsymbol{\mu}_q^{k+1} = \frac{\sum_{n=1}^{N} \gamma(z_{nq}) \boldsymbol{x}_n}{\sum_{n=1}^{N} \gamma(z_{nq})}, \quad q = J+1, \ldots, J+Q, \tag{19}$$

$$\boldsymbol{\Sigma}_q^{k+1} = \frac{\sum_{n=1}^{N} \gamma(z_{nq})(\boldsymbol{x}_n - \boldsymbol{\mu}_j^{k+1})(\boldsymbol{x}_n - \boldsymbol{\mu}_j^{k+1})^{\mathrm{T}}}{\sum_{n=1}^{N} \gamma(z_{nq})}, \quad q = J+1, \ldots, J+Q. \tag{20}$$

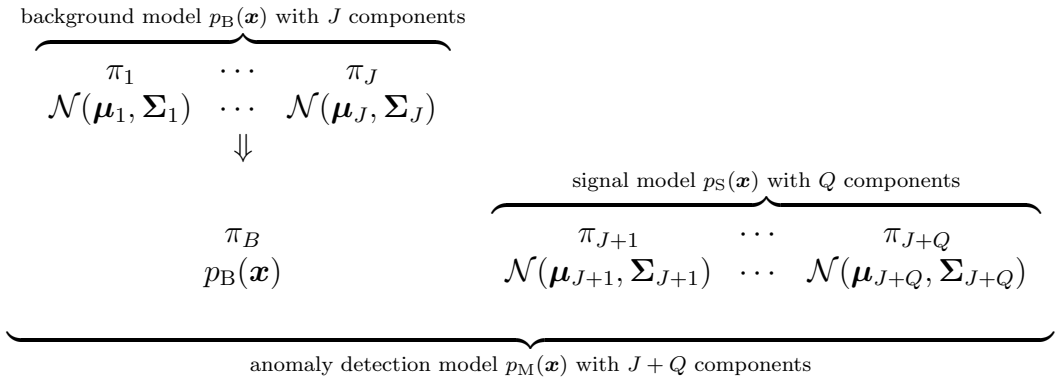

Figure 2: Illustration of the anomaly detection model proposed. Background model $p_{\mathrm{B}}(\boldsymbol{x})$ and signal model $p_{\mathrm{S}}(\boldsymbol{x})$ are mixtures of Gaussians with J and Q components, respectively. The background model is combined with the signal model with an additional mixture proportion π_B.

Figure 2 illustrates the anomaly detection model and its components. We propose calling the ML procedure described above *Expectation-Maximization with fixed background* or *fixed background EM* in short.

3.3 Non-Parametric Density Estimation

The most common density estimator used in high-energy physics is the histogram which is an example of a non-parametric density estimator. The histogram divides the data space into bins of fixed size. Smoothness of the histogram can be controlled by varying the bin width. In addition to the bin width, one has to also define the origin of the bins.

Because of their simplicity, histograms are usually good for visualization purposes—especially if the data is univariate. Drawbacks of the histogram are that it does not produce continuous probability density functions and the choice of the bin width can affect the results [14]. Furthermore, the histogram becomes inefficient for higher dimensional data.

Kernel density estimation (KDE) methods are another type of non-parametric density estimators. The basic idea of KDE is to place a "bump", for example a Gaussian, on each data point. These bumps are summed up and scaled so that they form a continuous and normalized PDF. Formally, the kernel density estimator is defined by

$$p(\boldsymbol{x}) = \frac{1}{N} \sum_{n=1}^{N} K(\boldsymbol{x} - \boldsymbol{x}_n), \tag{21}$$

where $\boldsymbol{x}_n$ iterates the events used for the density estimation and K is a kernel function. In the case of the Gaussian kernel for multivariate data, K is defined by

$$K(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\Big(-\tfrac{1}{2} \boldsymbol{x}' \boldsymbol{\Sigma}^{-1} \boldsymbol{x}\Big), \tag{22}$$

where d is the dimensionality of the data and $\mathbf{\Sigma}$ is the covariance matrix. The selection of $\mathbf{\Sigma}$ controls the smoothness of the density estimate, cf. the number of bins in the histogram. Several methods for selecting the covariance matrix have been developed, see e.g. [15].

The advantage of KDE is that it is able to capture distributions of very complex shapes if large enough training sample is provided. The trade-off, on the other hand, is that KDE becomes computationally expensive when the number of events increases.

4 Demonstration: Higgs Detection

In this section we compare the different anomaly search methods using using CDF MC events for WH Higgs. Here, the task is to correctly classify any given event as background or Higgs signal. As an optimal, *gold standard* classifier we use a neural network trained with "the correct" Higgs signal, i.e., the NN was trained and tested using Higgs MC with same mass m_{H}. Since the Higgs signal is among the background, any reasonable decision boundary yields background rejection less than 1.

4.1 Description of MC samples

We demonstrate our methods using CDF Higgs MC which consists of background and MC generated $q\bar{q} \to WH$ events. The background sample used to train the background models contains both MC and real data and includes 3406 events. The measured sample (signal+background) consist of 400 Higgs and 3406 background events resulting in a sample containing 10.5 % of signal. It has to be emphasized that this is not a realistic Higgs analysis as our signal/background ratio (i.e., cross section of Higgs signal) is not consistent with the actual Higgs cross section. Instead, the motivation of this study is to merely demonstrate our methods in preparation for a more realistic analysis.

The signal is generated with the Higgs masses $m_{\mathrm{H}} = 100, 115, 135, 150$ GeV representing uncertainty about the signature of the signal one is looking for. All the events are required to be tagged for two secondary vertices and each event is characterized by 8 variables. We use the same 7 variables used in the CDF Bayesian Neural Networks Higgs study [16] for double tagged events. In addition, we use KIT, a neural network b tagger variable from [17].

4.2 Dimensionality Reduction

Before doing the dimensionality reduction we normalize the samples. Since PCA is sensitive to outliers [18], we use logarithmic normalization

$$x_i = \mathrm{sgn}(x_i) \log(1 + |x_i|) \tag{23}$$

where the sign of the measurements is preserved.

After normalization, we conduct a dimensionality reduction from eight dimensions into two dimensions using PCA. The principal subspace is calculated using the background sample only. In the projection, 52 % of the information in the background sample is preserved. Figure 3a shows the proportion of the variance explained as a function of number of eigenvectors included in the projection (i.e., number of dimensions in the projection). The scree graph in Figure 3b shows the plot of the corresponding eigenvalues.

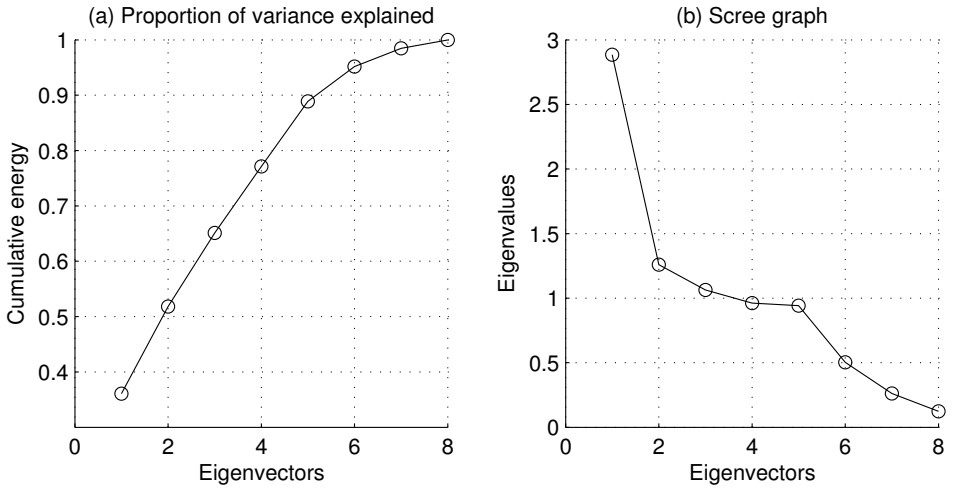

Figure 3: (a) The proportion of variance explained as a function of the number of eigenvectors in the PCA projection for the background sample. (b) Scree graph for the background sample.

4.3 Fixed Background Model Results

4.3.1 Modeling the Background Signal with a MoG

In order to select a suitable number of components for modeling the background distribution, we performed 5-fold cross-validation procedure which allows us to select optimal model complexity to adequately model the data while avoiding overfitting. In cross-validation the sample was divided into five equal-sized parts and each of these parts was used as a validation set in turn, i.e., log-likelihood of the data vectors in validation set was calculated with a model trained using the other four parts of the sample. Each fold of the cross-validation was performed 10 times resulting 50 log-likelihood values for each model with unique number of components. We ran the whole procedure for models with number of component distributions J ranging from 1 to 15. In all, 750 mixture models were trained.

Figure 4a shows the results of the cross-validation procedure. Out of the 50 log-likelihoods per mixture model, the mean and the interquartile borders for each J are calculated. The mean of the log-likelihood for the training sample is a monotonically

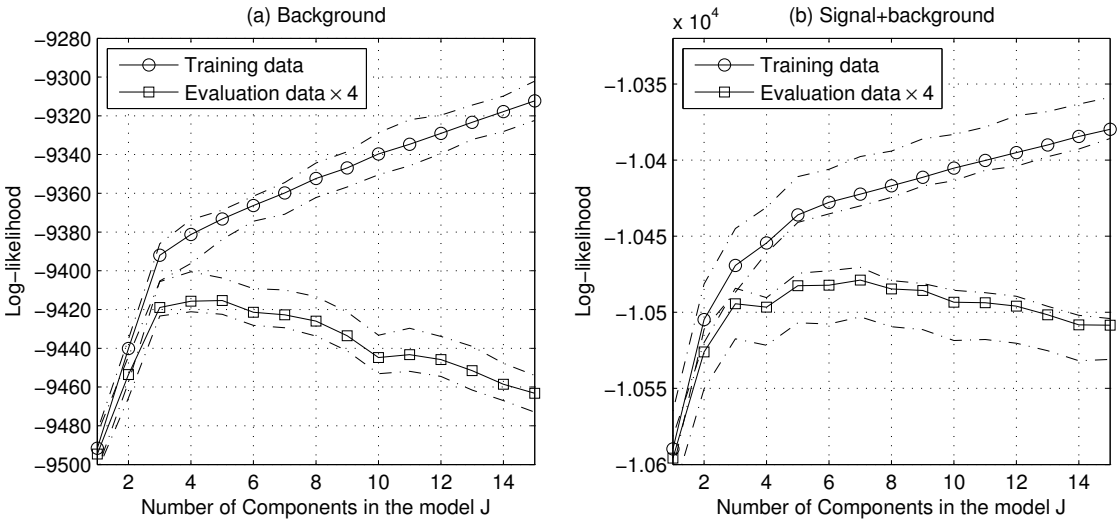

Figure 4: The log-likelihoods for different training and evaluation sets for (a) the background sample and (b) the signal+background sample ($m_{\mathrm{H}} = 150$ GeV) as a function of the number of mixture components J. The training and evaluation sample log-likelihoods are marked with solid lines (evaluation likelihood is multiplied by four to fit in the plot with the training likelihood) and the interquartile ranges for the 50 evaluation runs are drawn with dash-dotted lines.

increasing curve with growing J (increasing model complexity). As the training set was four times larger evaluation likelihoods are multiplied by four to make them fit in the same plot with the training likelihoods. The best model can be chosen based on the mean of the evaluation likelihoods. The model that gives the highest likelihood for the evaluation sample is a model with five component distributions ($J = 5$). As one trains a model with more than five components the training likelihood is increased but the model has overlearned the sample since the evaluation likelihood is decreasing. Model with five components can also be motivated from the parsimony point of view: a simpler model is preferred over the more complex one.

Figure 5a shows an example of the background model with five Gaussian components. The events are plotted in the two-dimensional principal subspace and the solid lines show contours of the PDF estimated using the background sample.

4.3.2 Modeling the Higgs Signal with a MoG

We analyzed the signal+background sample using the same cross-validation procedure explained in the previous subsection. Figure 4b shows that mixing signal with the background makes the sample more complex: the optimal number of components for modeling the signal+background sample is seven. Comparing this to the background model, the number of additional components needed to model the signal+background sample is two. Thus, we chose to use two Gaussian components for the signal model $p_{\mathrm{S}}(\boldsymbol{x})$.

We trained a model for the signal+background sample with $m_{\mathrm{H}} = 150$ GeV Higgs

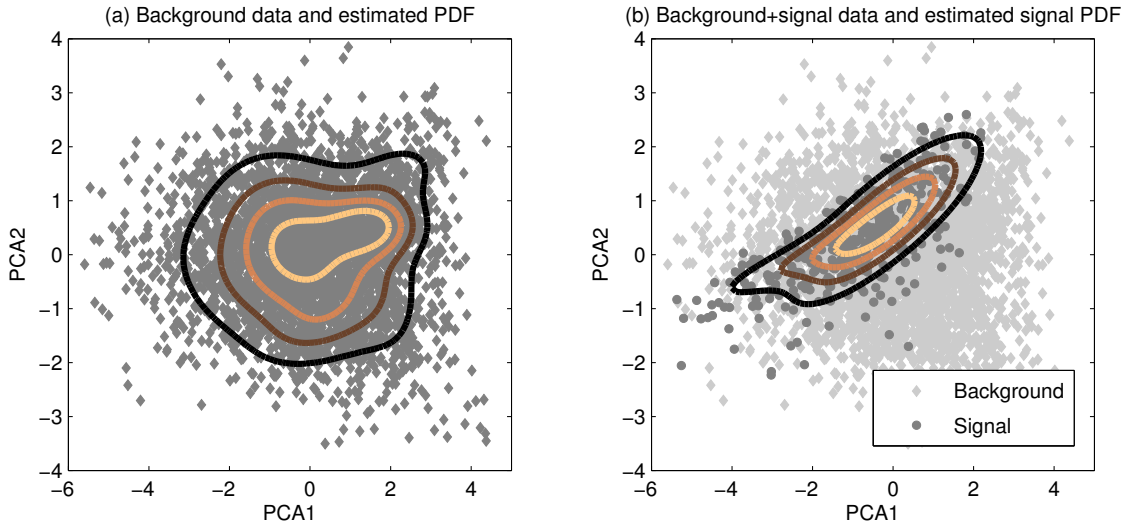

Figure 5: An illustration of the fixed background EM: a projection of (a) the background and (b) the signal+background ($m_{\mathrm{H}} = 150$ GeV) samples into the two-dimensional principal subspace defined by the background sample. The solid lines show contours from the estimated PDFs for (a) the background and (b) the signal. The estimation is done using the fixed background EM procedure.

using the fixed background EM procedure as described in Section 3.2.2. The resulting signal distribution $p_{\mathrm{S}}(\boldsymbol{x})$ with two signal components and the Higgs events projected to the two-dimensional principal subspace are shown in Figure 5b.

4.3.3 Using the Model for Anomaly Search

With the fixed background model, the classification can be performed using equations (2) and (3). Figure 7 shows the receiver operating characteristic (ROC) curves for the classifiers with different Higgs masses. These models were trained using five background and two signal components. The curves are obtained using various values for the threshold T and show the background rejection rate (i.e., the proportion of background events correctly rejected) as a function of the signal efficiency (i.e., the proportion of events correctly identified as signal). From the ROC curves, one can see that regardless of the mass of the Higgs the anomaly search method is able to identify the signal with a good efficiency.

We justify our model selection procedure in an indirect way by integrating the ROC curves of classifiers with different background and signal models to get a single performance measure for each model. In general, one is not able to perform such a model selection in a model-independent training scheme. Figure 6a illustrates the mean magnitudes of 10 ROC curve integrals for different models. From the figure, it can be seen that a large group of models less complex than the optimal model suggested by the cross-validation procedure give almost equal results. For example, by choosing a model with one signal and three background components one can obtain almost optimal

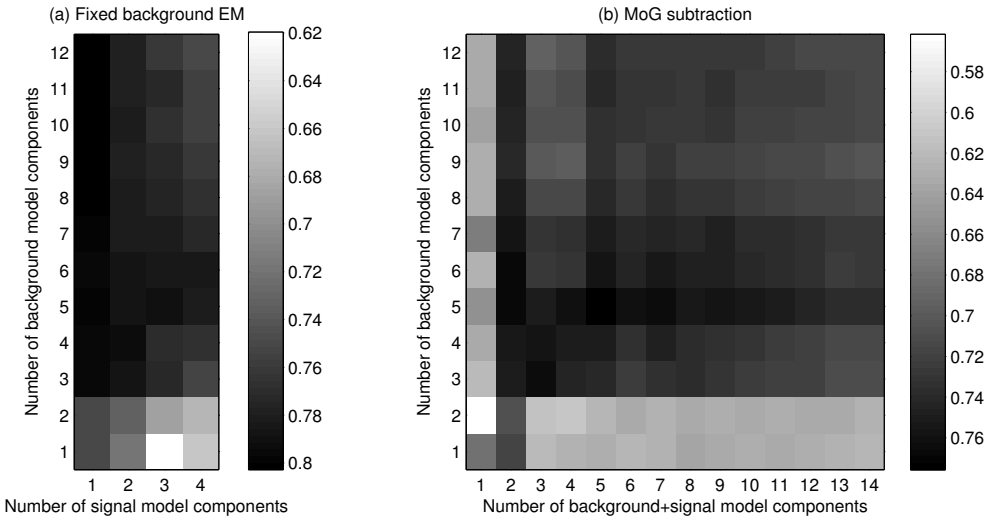

Figure 6: The integrals of the ROC curves for different anomaly search models with 3406 background events and 400 $m_{\mathrm{H}} = 150$ GeV Higgs events. The shade of the cells in the arrays show the mean magnitudes of 10 integrals for different models generated with (a) the fixed background EM and (b) the MoG background subtraction methods. Notice the different scales on the two figures.

classification results, i.e., the ROC curves that an optimal neural network obtains (see Figure 10b). Further comments on model selection can be found in Section 5.

An additional advantage of using fixed background EM is that the signal model weight λ is proportional to the cross section of the signal. Table 1 shows estimated values for λ with models trained using different Higgs masses. Table shows that the method gives estimates for λ which are very close to the real proportion of the signal, $\lambda_{\mathrm{real}} = 0.105$.

Higgs mass m_{H} (GeV)	λ
150	0.122
135	0.118
115	0.121
100	0.106

Table 1: The estimated values for λ with models trained using different Higgs masses m_{H}. The values for λ are mean values from ten different models, although there was no deviation between the models using the same mass. The real proportion of the signal in the sample was $\lambda_{\mathrm{real}} = 0.105$.

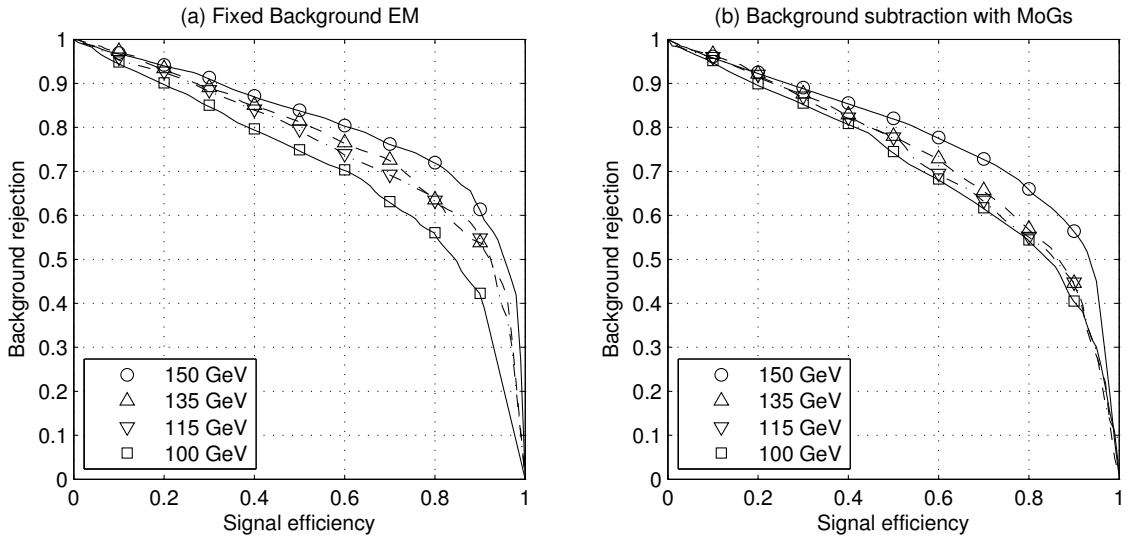

Figure 7: The ROC curves for (a) the fixed background EM and (b) MoG subtraction methods. The plots show the background rejection rate (i.e., the proportion of background events correctly rejected) for different Higgs masses m_{H} as a function of the signal efficiency (i.e., the proportion of events correctly identified as signal).

4.4 Background Subtraction Results

4.4.1 Subtraction of Two MoG Models

We use MoGs for modeling the background and signal+background samples and calculate a discriminant function $D(\boldsymbol{x})$ according to equation 6. The resulting background PDF and discriminant function are illustrated in Figure 8. For the model selection, similar techniques as described in the previous sections and discussed in Section 5 can be applied here, as well.

Figure 6b illustrates the mean magnitudes of 10 ROC curve integrals for different MoG subtraction models. From the figure, it can be seen that there is a fair amount of models that give almost similar results. Figure 7b shows ROC curves for a MoG subtraction model with five background and five signal+background components. In general, the MoG subtraction method gains slightly worse results compared to the fixed background EM method.

4.4.2 Subtraction of Two Kernel Density Estimates

Instead of using MoGs for modeling the probability distribution, we also experimented with using kernel density estimation for modeling the background and signal+background samples. Figure 9a shows the estimated background PDF. In Figure 9b contours from the positive areas of the discriminant function $D(\boldsymbol{x})$ are shown.

At this stage of work, we have not applied any automatic methods for finding the kernel covariance matrix $\boldsymbol{\Sigma}$. Instead, we performed a search over the parameter space

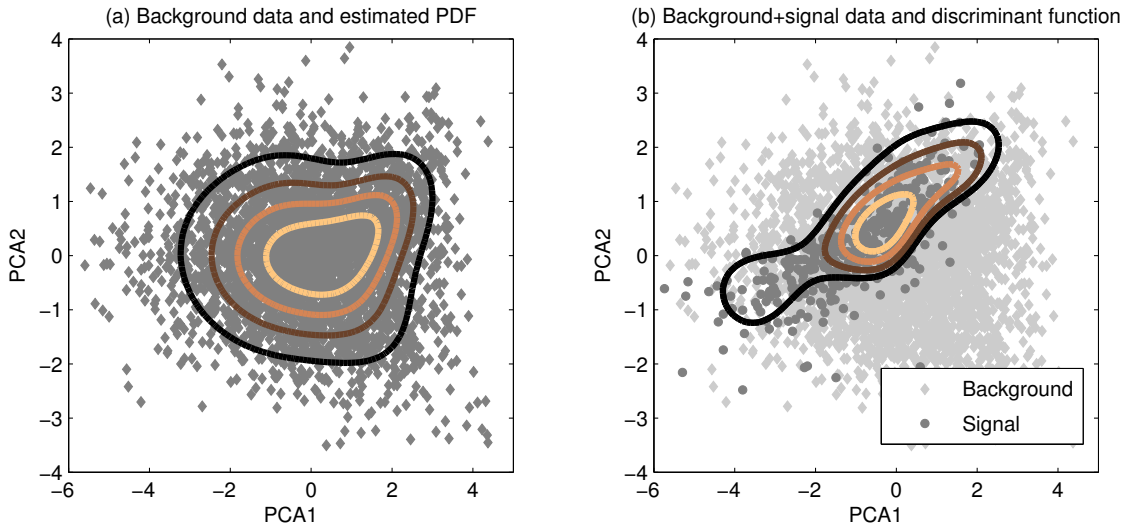

Figure 8: An illustration of background subtraction using MoG models: a projection of (a) the background and (b) the signal+background ($m_{\mathrm{H}} = 150$ GeV) samples into the two-dimensional principal subspace defined by the background sample. The solid lines of (a) show contours from the estimated PDF for the background and the solid lines of (b) show contours from the positive areas of the discriminant function $D(\boldsymbol{x})$.

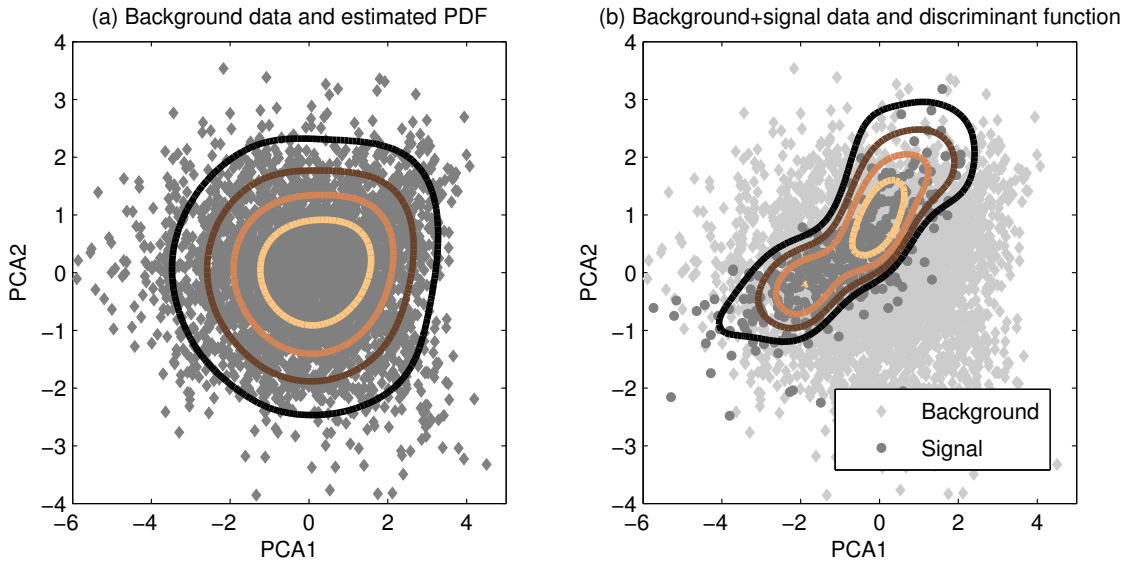

Figure 9: An illustration of background subtraction using kernel density estimation: a projection of (a) the background and (b) the signal+background ($m_{\mathrm{H}} = 150$ GeV) samples into the two-dimensional principal subspace defined by the background sample. The solid lines of (a) show contours from the estimated PDF for the background and the solid lines of (b) show contours from the positive areas of the discriminant function $D(\boldsymbol{x})$.

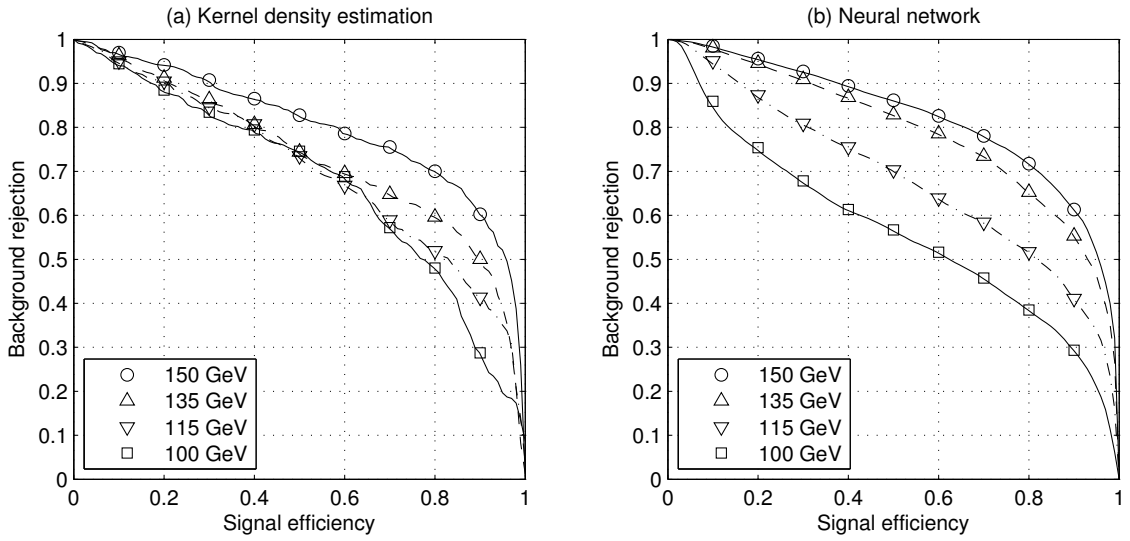

Figure 10: The ROC curves for (a) the kernel density estimation and (b) the neural network. The neural network was trained using the $m_{\mathrm{H}} = 150$ GeV Higgs signal. The plots show the background rejection rate (i.e., the proportion of background events correctly rejected) for different Higgs masses m_{H} as a function of the signal efficiency (i.e., the proportion of events correctly identified as signal).

optimizing the integral of the ROC curve to determine the maximal performance of KDE subtraction. Furthermore, we used the same $\mathbf{\Sigma}$ for both the background and the signal+background distributions and assumed spherical kernels, i.e., $\mathbf{\Sigma} = \sigma\mathbf{I}$. Based on the search, we chose $\mathbf{\Sigma} = 0.6\mathbf{I}$. Figure 10a shows the ROC curves of the KDE experiments. The figure shows that, in the case of an optimal $\mathbf{\Sigma}$, KDE is able to provide classification results comparable to NNs of Figure 10b.

4.5 Comparison to Neural Networks

In order to compare anomaly search to more traditional model-dependent classification methods, an MLP neural network was trained using the $m_{\mathrm{H}} = 150$ GeV Higgs signal as the training sample. In contrast to the anomaly search experiments, the neural network was trained using the signal-to-background ratio of 1:1.

Figure 10b shows that the neural network classifier is able to produce good signal efficiency and background rejection provided that the real mass of the Higgs is indeed 150 GeV. If it, however, turns out that the mass is different from this, the performance of the classifier degrades as the training signature is different from the actual signal signature. Comparing the $m_{\mathrm{H}} = 100$ GeV curves in Figure 10b and for example Figure 7a shows that in the case of an unexpected signal, anomaly search is able to correctly locate the signal while a model-dependent NN gives suboptimal performance. The neural network is hence only able to reliably identify the signal when its training sample is in good agreement with the true signal.

5 Discussion and Future Work

Anomaly search was demonstrated in this work using Higgs MC with a non-physical cross section for the signal. The signal had to be amplified because of a limited number of background events available. In the near future, we hope to be able to conduct a similar analysis using a more realistic physics scenario involving for example SUSY or other exotic physics processes. For this kind of analysis, the model independent approach is likely to provide a clear advantage over traditional methods which depend heavily on the possibly inaccurate MC model used. Also, in the first stages of commissioning a new detector, such an approach could be used to study detector defects and gain a better understanding of the experiment.

Until now, we have only considered the ROC curves to evaluate the performance of the proposed methods. In addition, it would be important to be able to measure the significance of the signal found by the anomaly search. This could perhaps be achieved by considering distributions of the discriminant function values $D(\boldsymbol{x})$. We are also investigating other methods besides EM for PDF estimation with the MoG model. One possibility is to use Bayesian methods based on Markov Chain Monte Carlo integration or variational approximations [10].

Using PCA for the dimensionality reduction is not a trouble-free solution. Firstly, one has to decide which sample is used to define the principal subspace. If the subspace is spanned by the first n principal components of the background sample only (as we do in our experiments), there is a risk of losing all the important variance in the signal, as it cannot be guaranteed that the signal has variance in the same directions as the background sample. Secondly, the mapping to a principal subspace might be such that it maps a well separated signal onto the background, thus only complicating the anomaly search problem. It is left for future work to investigate possibilities of using adaptive PCA methods (i.e., online PCA), independent component analysis [19] or canonical correlation analysis [20] for the dimensionality reduction.

selection decided to components) for values for Finally, we make a brief comparison between the different methods studied here. The advantages of the fixed background EM method are (i) its lowest computational cost, (ii) large group of well performing models, which makes model selection easier, and (iii) its ability to estimate the cross-section of the signal. Moreover, the method is completely probabilistic, i.e., the results given by the model have a direct probabilistic interpretation. It should be noted, however, that fixed background EM can only handle an excess of data while background subtraction is able to handle deficits as well. Despite this, based on the reasons above, the fixed background EM method looks the most promising from the anomaly search point of view.

6 Conclusions

The goal of this work was to show that density estimation based anomaly search is able to identify new physics signals provided that there exists an accurate representa-

tion for the background. The feasibility of the proposed methods was demonstrated using CDF MC for WH Higgs. In particular, it was shown that the methods can be employed without a Monte Carlo representation for the signal. Thus, the methods are robust against uncertainties in the Monte Carlo generators or parameters of the physical models behind the generators.

Three different variants of the anomaly search approach were studied, namely fixed background EM and both a parametric and a non-parametric version of background subtraction. Out of these methods, fixed background EM turned out to give the best classification performance for a wide range of models while at the same time being computationally the least expensive solution. This method was also able to give fairly accurate estimates for the proportion of signal in the measured data, hence resulting in an estimate for the signal cross section. However, all the methods studied obtained results close to the gold standard given by neural networks in the case of a foreseen signal. For an unexpected signal, is was shown that anomaly search is able to find the signal even in situations where the model-dependent classifiers might fail badly.

We believe that analysis of exotic physics signals such as SUSY would greatly benefit from the proposed anomaly search approach. Hence, we are at the moment trying to identify suitable physics processes for demonstrating the applicability of the approach in a more realistic physics analysis scenario.

References

[1] The CMS Collaboration. Music – an automated scan for deviations between data and monte carlo simulation. Oct 2008.

[2] CDF collaboration. Model-independent (vista) and quasi-model-independent (sleuth) search for new high-pt physics at cdf. CDF note 8763, 2007.

[3] R.J. Bolton and D.J. Hand. Unsupervised profiling methods for fraud detection. In *Conference on credit scoring and credit control*, volume 7, pages 5–7, 2001.

[4] FML D'Almeida and AA Nepomuceno. Subtracting and Fitting Histograms using Profile Likelihood. In *PHYSTAT LHC Workshop on Statistical Issues for LHC Physics*, page 155, 2008.

[5] M. Piccardi. Background Subtraction Techniques: A Review. In *IEEE International Conference on Systems, Man and Cybernetics*, volume 4, pages 3099–3104, 2004.

[6] A. Mittal and N. Paragios. Motion-Based Background Subtraction Using Adaptive Kernel Density Estimation. 2004.

[7] Richard E. Bellman. *Adaptive control processes - A guided tour*. Princeton University Press, 1961.

[8] I. T. Jolliffe. *Principal Component Analysis*. Springer, second edition, October 2002.

[9] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, (24):417–441, 1933.

[10] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

[11] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics. Wiley-Interscience, October 2000.

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[13] Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*. Wiley-Interscience, 2 edition, March 2008.

[14] B.W. Silverman. *Density estimation for statistics and data analysis*. Chapman & Hall/CRC, 1998.

[15] David W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.

[16] Y. Nagai et al. Search for the Standard Model Higgs boson production in association with a W boson using 4.3/fb. CDF/PUB/EXOTIC/PUBLIC/9997, 2009.

[17] Thorsten Chwalek et al. Update of the neural network b tagger for single-top analyses. CDF/ANAL/TOP/CDFR/8903, 2007.

[18] Ethem Alpaydin. *Introduction to Machine Learning*. MIT Press, Cambridge, MA, 2. edition, 2010.

[19] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent component analysis*. John Wiley & Sons, 2001.

[20] Francis R. Bach and Michael I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.