

Semi-Supervised Detection of Collective Anomalies with an Application in High Energy Particle Physics

Tommi Vatanen*, Mikael Kuusela*[†], Eric Malmi*, Tapani Raiko*, Timo Aaltonen[†] and Yoshikazu Nagai[‡]

*Aalto University School of Science, P.O. Box 15400, FI-00076 Aalto, Finland, first.last@aalto.fi

[†]Helsinki Institute of Physics, P.O. Box 64, FI-00014 University of Helsinki, Finland, first.last@helsinki.fi

[‡]University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8577, Japan, nagai@fnal.gov

Abstract—We study a novel type of a semi-supervised anomaly detection problem where the anomalies occur collectively among a background of normal data. Such problem arises in experimental high energy physics when one is trying to discover deviations from known Standard Model physics. We solve the problem by first fitting a mixture of Gaussians to a labeled background sample. We then fit a mixture of this background model and a number of additional Gaussians to an unlabeled sample containing both background and anomalies. This way we not only detect but also perform pattern recognition of anomalies. Such mixture model allows us to perform classification of anomalies vs. background, estimate the proportion of anomalies in the sample and study the statistical significance of the anomalous contribution. We first verify the performance of the method using artificial data and then demonstrate its real-life applicability using a data set related to the search of the Higgs boson at the Tevatron collider.

Index Terms—Anomaly detection, semi-supervised learning, EM algorithm, Gaussian mixture models, high energy physics

I. INTRODUCTION

Anomaly detection (outlier detection, novelty detection) refers to the problem of detecting patterns in the data that deviate from the expected, normal behavior so much that they arouse suspicion of having been generated by a different mechanism [1]–[4]. Such methods have applications in, e.g., credit card fraud detection [5], network intrusion detection [6] and aircraft engine fault detection [7].

Since in many application domains it is impractical or even impossible to obtain a representative sample of anomalous data, most work on anomaly detection is focused on the unsupervised problem setting where one does not assume the existence of a labeled sample of training data. Anomalies can be detected within an unlabeled sample by making the implicit assumption that they are produced by rare, infrequent processes.

One can introduce a lot more structure to the problem by assuming the existence of a labeled sample of normal data patterns in which case the problem translates into a semi-supervised anomaly detection problem. In the anomaly detection literature, such training sample is usually called the normal data. However, in order to conform to the terminology used in our application domain and to avoid confusion with a sample generated by a Gaussian distribution, we will be using throughout this paper the term *background sample* to denote a labeled sample of normal observations.

Most semi-supervised anomaly detection techniques use the background sample to produce a model for the normal behavior. In what follows, we call this model the *background model*. These methods then look at individual observations of an unlabeled test sample one-by-one and classify an observation as an anomaly if it seems unlikely to have been produced by the process corresponding to the background model. One could for example produce a density estimate for the background data and then classify observations as anomalies should they fall in the low probability density regions of the data space [3].

A common limitation of existing anomaly detection techniques is that they are unable to detect anomalies which lie within the domain of the background data. This is because a single such observation looks as if it could have been produced by the background process. This paper overcomes such a limitation in the semi-supervised problem setting in situations where anomalies occur as a cluster among the background data. We call such observations *collective anomalies* in accordance with the definition given by Chandola et al. [2]: “The individual data instances in a collective anomaly may not be anomalies by themselves, but their occurrence together as a collection is anomalous.”

Existing work on collective anomaly detection requires auxiliary relationships to connect the observations, such as sequential, spatial or graph information [2]. However, to the best of our knowledge, detection of collective anomalies where collections are characterized by distance in the data space has not been studied before. This is perhaps because of the fact that in the unsupervised setting it does not make sense to talk about collections in the data space. Indeed, in the absence of labeled observations of normal behavior, the more the observations cluster together, the less anomalous they become. But in the semi-supervised setting, a cluster or a “bump” of data which is not seen in the background sample is certainly anomalous even if it lay among the background.

Our considerations are motivated by a real data analysis problem arising in experimental high energy physics where one is trying to look for signs of new physical phenomena not described by the Standard Model of particle physics. Traditionally, such analyses are conducted in a model-dependent mode [8], where one trains a supervised classifier to search for phenomena expected to be seen under some hypothetical model of new physics, such as supersymmetry, extra dimensions, etc.

However, such analyses could completely miss deviations not described by any of the proposed new theories. Because of this, it would be important to complement such searches with model-independent techniques which are sensitive to various types of deviations – or anomalies – with respect to known physics [9]. Although the motivation for the use of such techniques is clear, they have been employed very little in practice because of the apparent lack of existing computational techniques for solving such problems. The work presented in this paper aims at providing one such technique to be used in future model-independent searches of new physics.

In addition to requiring a labeled sample of background observations, our algorithm works under the assumption that: (1) the anomalies occur collectively as an excess over the distribution of the background, (2) there is a large enough number of anomalies in order to enable collective inferences, (3) the dimensionality of the data is or can be reduced to be small enough to allow density estimation with mixture models and (4) the background has a stationary distribution.

Mixture models have been studied in the context of anomaly detection in the earlier literature, but not in our semi-supervised problem setting. For example, Eskin [10] employs a similar mixture model for unsupervised anomaly detection under the assumption that anomalies can be modeled using a uniform distribution with a small mixture proportion, while Lauer [11] fits a Gaussian mixture model to an unlabeled data set by setting the anomalous component of the model to consist of a widespread Gaussian. Ritter and Gallegos [12] identify outliers based on a mixture model in the context of a chromosome classification problem. The common limitation in all of these algorithms is that they are unable to identify anomalies within the domain of the background data, which is an essential requirement in our application domain.

This paper is organized as follows. In Section II, we introduce a model for semi-supervised anomaly detection, which we call the *fixed-background model*. We then discuss the training of the model using the expectation-maximization algorithm in Section III. We verify the performance of the method using artificial data in Section IV. In particular, we show that the algorithm is able to achieve nearly optimal classification performance and gives accurate estimates for the proportion of anomalies given that there is a large enough number of them in the sample. We then demonstrate the real-life applicability of the method in model-independent searches of new physics by applying it to a data sample related to the search of the Higgs boson at the Tevatron collider. We discuss the proposed framework in Section VI before concluding in Section VII.

II. FIXED-BACKGROUND MODEL FOR ANOMALY DETECTION

To detect collective anomalies among the background, we proceed in two steps. First, we use parametric density estimation to learn a *background model*, $p_B(\mathbf{x})$, using the labeled background data. The next step is to model the unlabeled data with a *fixed-background model*, $p_{FB}(\mathbf{x})$, which is a mixture of

the background model and a new *anomaly model* $p_A(\mathbf{x})$:

$$p_{FB}(\mathbf{x}) = (1 - \lambda)p_B(\mathbf{x}) + \lambda p_A(\mathbf{x}). \quad (1)$$

The fixed-background model is fitted to the unlabeled data by maximizing its likelihood under the constraint of keeping the background model $p_B(\mathbf{x})$ fixed. Hence, the anomaly model, $p_A(\mathbf{x})$, captures any unexpected deviations from the distribution of the background.

As a simple illustration of the fixed-background model, Figure 1(a) shows a univariate data set of background data generated from a Gaussian distribution and a maximum likelihood Gaussian density $p_B(x)$ estimated using the data set. Figure 1(b) shows a very simple anomalous pattern that can be modeled with a single additional univariate Gaussian. Given a sample contaminated with these anomalies, our goal is to find an optimal combination of the parameters of the anomaly model (μ_A, σ_A) and the mixing coefficient λ in Eq. (1). The resulting model $p_{FB}(x)$ is shown with a black line and the anomaly model $p_A(x)$ with a gray line in Figure 1(b).

The fixed-background model enables a variety of data analysis tasks:

- 1) *Classification*: Observations can be classified as anomalies using the posterior probability as a discriminant function

$$p(\text{anomaly}|\mathbf{x}) = \frac{\lambda p_A(\mathbf{x})}{(1 - \lambda)p_B(\mathbf{x}) + \lambda p_A(\mathbf{x})} \equiv \mathcal{D}(\mathbf{x}). \quad (2)$$

The decision rule is then

$$\mathcal{D}(\mathbf{x}) = \begin{cases} \geq T \Rightarrow \mathbf{x} \text{ is an anomaly,} \\ < T \Rightarrow \mathbf{x} \text{ is background,} \end{cases} \quad (3)$$

where the constant $T \in [0, 1]$ is a threshold which is used to control the sensitivity of the classifier.

- 2) *Proportion of anomalies*: The mixing proportion of the anomaly model λ directly gives us an estimate for the proportion of anomalies in the unlabeled sample.
- 3) *Significance of the anomaly*: The statistical significance of the anomaly model can be evaluated by performing a statistical test for the background-only null hypothesis $\lambda = 0$. This enables us to discriminate between statistical fluctuations of the background and real collective anomalies. The test is performed using the likelihood ratio test statistic [13], [14]

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} \mathcal{L}(\theta)}{\sup_{\theta \in \Theta} \mathcal{L}(\theta)}, \quad (4)$$

where Θ_0 refers to the set of parameters allowed by the null hypothesis and $\mathcal{L}(\theta)$ is the likelihood of parameters θ . In our case, the nominator is simply the likelihood of the background model and the denominator the likelihood of the fixed-background model. Following Wang et al. [15], we obtain the distribution of the test statistic under the null hypothesis using nonparametric bootstrapping [16]. That is, we sample with replacement observations from the background data, fit $p_{FB}(\mathbf{x})$ to this new sample and compute the corresponding value of Λ .

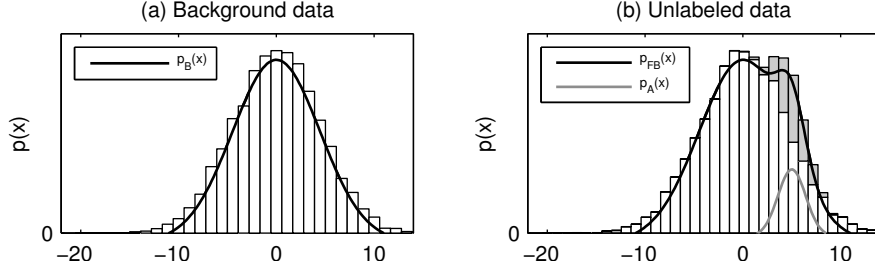


Fig. 1. (a) A histogram of background data from a univariate Gaussian distribution and an estimated background model $p_B(x)$. (b) An illustration of the fixed-background model in a univariate case. The histogram shows the unlabeled data (the gray excess in the histogram denotes the anomalous observations). The estimated fixed-background model $p_{FB}(x)$ is shown with a black line and the anomaly model $p_A(x)$ with a gray line.

A large enough number of such resamplings allows us to recover the distribution of Λ under the background-only null hypothesis and hence to compute the p -value of the observed collective anomaly.

III. METHODS

In this section, we describe the methods used in our experiments. We first review the expectation-maximization (EM) algorithm for multivariate mixtures of Gaussians (MoG) and then describe in detail how to use the algorithm for learning the fixed-background model in Eq. (1).

A. Mixture of Multivariate Gaussian Distributions

Finite mixtures of distributions are a flexible method for modeling complex data sets [17]. In this work, we use multivariate MoGs to represent the distribution of the data. The mixture of J multivariate Gaussian distributions is defined as

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{j=1}^J \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \quad (5)$$

where $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$ denotes the probability density function of a multivariate Gaussian with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$ at \mathbf{x} . The π_j are mixture proportions (or mixing coefficients) which satisfy $\pi_j \geq 0$ and $\sum_{j=1}^J \pi_j = 1$, and $\boldsymbol{\theta} = \{\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\}_{j=1}^J$ represents the parameters of the mixture model with J components.

B. EM Algorithm for the Background Model

Let us first consider the case of fitting a MoG model with J components to the background data with N observations $\mathbf{x}_i, i = 1, \dots, N$. The log-likelihood of the parameters $\boldsymbol{\theta}$ is

$$l(\boldsymbol{\theta}) = \log(\mathcal{L}(\boldsymbol{\theta})) = \sum_{i=1}^N \log \left(\sum_{j=1}^J \pi_j \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right). \quad (6)$$

Here we have assumed that the observations are independent and identically distributed (i.i.d.).

The maximum likelihood (ML) estimate of the parameters can be obtained by maximizing (6), which is carried out by using the EM algorithm [18], [19]. The algorithm proceeds in two steps. In the *expectation step* (E-step), one computes

the posterior probabilities for each data point \mathbf{x}_i to have been generated by the j th Gaussian component

$$p(z_{ij} = 1|\mathbf{x}_i, \boldsymbol{\theta}^k) = \frac{\pi_j^k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_j^k, \boldsymbol{\Sigma}_j^k)}{\sum_{j'=1}^J \pi_{j'}^k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_{j'}^k, \boldsymbol{\Sigma}_{j'}^k)} \equiv \gamma_{ij}^k. \quad (7)$$

Here, $\boldsymbol{\theta}^k$ contains the parameter estimates at the k th iteration and z_i indicates which component generated the i th observation.

In the subsequent *maximization step* (M-step), the parameter values are updated according to the following equations

$$\pi_j^{k+1} = \frac{1}{N} \sum_{i=1}^N \gamma_{ij}^k, \quad (8)$$

$$\boldsymbol{\mu}_j^{k+1} = \frac{\sum_{i=1}^N \gamma_{ij}^k \mathbf{x}_i}{\sum_{i=1}^N \gamma_{ij}^k}, \quad (9)$$

$$\boldsymbol{\Sigma}_j^{k+1} = \frac{\sum_{i=1}^N \gamma_{ij}^k (\mathbf{x}_i - \boldsymbol{\mu}_j^{k+1})(\mathbf{x}_i - \boldsymbol{\mu}_j^{k+1})^T}{\sum_{i=1}^N \gamma_{ij}^k}. \quad (10)$$

A detailed derivation of the EM algorithm for mixtures of Gaussians can be found in [19] where it is also shown that each iteration of the EM algorithm increases the log-likelihood until a local maximum is found.

C. EM Algorithm for the Fixed-Background Model

In this section, we elaborate how to use the EM algorithm to estimate models of the form of Eq. (1). The goal is to search for unmodeled anomalies in the unlabeled data set. Now, the background model $p_B(\mathbf{x})$ in Eq. (1) is fixed and both λ and the parameters of $p_A(\mathbf{x})$ need to be optimized to maximize the log-likelihood. Here, $p_A(\mathbf{x})$ can either be a single Gaussian or more generally a MoG with Q components. We can now write Eq. (1) as follows

$$\begin{aligned} p_{FB}(\mathbf{x}) &= (1 - \lambda)p_B(\mathbf{x}) + \lambda \sum_{q=J+1}^{J+Q} \tilde{\pi}_q \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q) \\ &= \pi_B p_B(\mathbf{x}) + \sum_{q=J+1}^{J+Q} \pi_q \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q), \end{aligned} \quad (11)$$

where we have defined $\pi_B = 1 - \lambda$ and $\pi_q = \lambda \tilde{\pi}_q, q = J + 1, \dots, J + Q$. The mixture proportions satisfy $\pi_B +$

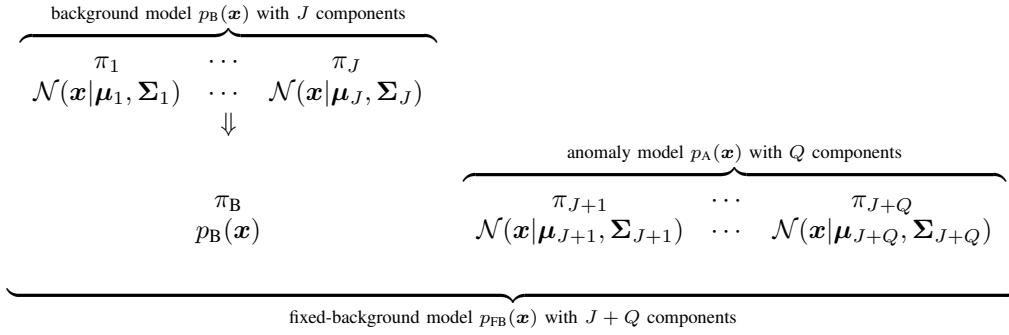


Fig. 2. Illustration of the proposed anomaly detection model. The background model $p_B(\mathbf{x})$ and the anomaly model $p_A(\mathbf{x})$ are mixtures of Gaussians with J and Q components, respectively. The background model is combined with the anomaly model with an additional mixture proportion π_B to give the fixed-background model $p_{FB}(\mathbf{x})$.

$\sum_{q=J+1}^{J+Q} \pi_q = 1$ and $\sum_{q=J+1}^{J+Q} \pi_q = \sum_{q=J+1}^{J+Q} \lambda \tilde{\pi}_q = \lambda$. This anomaly detection model and its components are illustrated in Figure 2.

The EM updates for the model in Eq. (11) are easily found by straightforward analogy to the standard mixture model. In the E-step, the posterior probabilities of the background model and the components of the anomaly MoG are updated as follows

$$p(z_{iB} = 1 | \mathbf{x}_i, \boldsymbol{\theta}^k) \quad (12)$$

$$= \frac{\pi_B^k p_B(\mathbf{x}_i)}{\pi_B^k p_B(\mathbf{x}_i) + \sum_{q'=J+1}^{J+Q} \pi_{q'}^k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_{q'}^k, \boldsymbol{\Sigma}_{q'}^k)} \equiv \gamma_{iB}^k, \quad (13)$$

$$p(z_{iq} = 1 | \mathbf{x}_i, \boldsymbol{\theta}^k) = \frac{\pi_q^k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_q^k, \boldsymbol{\Sigma}_q^k)}{\pi_B^k p_B(\mathbf{x}_i) + \sum_{q'=J+1}^{J+Q} \pi_{q'}^k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_{q'}^k, \boldsymbol{\Sigma}_{q'}^k)} \equiv \gamma_{iq}^k.$$

In the first equation, $z_{iB} = 1$ denotes that the i th observation was generated by the background model $p_B(\mathbf{x})$. In the second equation $q = J+1, \dots, J+Q$. In the subsequent M-step, the means and the covariances of the anomaly model are updated using Eqs. (9) and (10) for indices $j = J+1, \dots, J+Q$. The mixture proportions for these indices are also updated with Eq. (8), while the mixture proportion of the background model follows from the normalization constraint

$$\pi_B^{k+1} = 1 - \sum_{q=J+1}^{J+Q} \pi_q^{k+1} \left(= \frac{1}{N} \sum_{i=1}^N \gamma_{iB}^k \right). \quad (14)$$

D. Additional Remarks

Assessing the number of components in mixture models is a hard problem which has not been completely resolved [17]. We use the cross-validation-based information criterion (CVIC) [20] for model selection in the Higgs experiments of Section V, but take the correct number of components as given in our artificial data experiments. Naturally, any known information criterion can be used to perform model selection for the background model. Further discussion about model selection can be found in, e.g., [17].

The maximization of the log-likelihood function of a Gaussian mixture model is not a well-posed problem due to the

singularities corresponding to one of the Gaussian components “collapsing” onto a single data point, i.e., $\sigma_j \rightarrow 0$ in the one-dimensional case. With multivariate data, this corresponds to the case where the smallest eigenvalue of the covariance matrix $\boldsymbol{\Sigma}$ tends to zero. In this work, we avoid this problem by resetting the mean of a collapsing component to a randomly chosen data point while also resetting its variance or covariance matrix to some large value. We also reset the components with a very small mixture proportion to avoid unnecessary nuisance components.

We assess the “goodness” of the components in the anomaly model using a simple likelihood comparison. Using the likelihood of the background model as a reference, we take components of the anomaly model one at a time and combine them with the background model. Components that have learned some anomalous patterns in the unlabeled data should increase the likelihood compared to the background model. On the other hand, if the component under investigation decreases the likelihood, it is most probably useless. Again, components that do not appear to capture any anomalies in the data are reset to a random data point.

We also exploit the resetting heuristics above in order to remove excess components from the anomaly model. We assume that a component can be removed if it has been reset too many times and, consequently, hinders the convergence of the EM algorithm. Finally, while estimating the fixed-background model, the convergence of the algorithm is denied if the fixed-background model decreases the log-likelihood compared to the background model. Instead, poor components are reset and the EM iteration continues until a model that increases the log-likelihood is found or all anomalous components have been removed.

IV. EXPERIMENTS WITH ARTIFICIAL DATA

We tested the fixed-background model with artificial data sets generated using univariate mixtures of Gaussians. Each data set consisted of 100 000 background samples and 100 000 unlabeled test samples containing some anomalies. The data sets were generated using five components for the background data and three additional anomalous components for the unlabeled test data. All tests were ran using specifically imple-

mented scripts in MATLAB environment in order to exploit the heuristics described above.

We ran the tests using 10 different generative models. The means and variances of the components in each model were chosen randomly in such a way that the anomalies appear as clusters among the background data. For each model, we varied the proportion of anomalies from 1 % to 20 % and generated 10 data sets with each proportion and model. This resulted in 100 tests for each anomaly proportion. Figure 3 shows histograms of two artificial data sets with 10 % and 3 % of anomalies, respectively. Gray excess in the histogram bars denotes the anomalous data.

For each data set, we trained a fixed-background model as described in Sections II and III. The model was then used to classify the data in the unlabeled test data as background or anomalies with different thresholds according to Eqs. (2) and (3). This allowed us to construct the receiver operating characteristic (ROC) curves for each experiment, and use the area under the ROC curve (AUC) as a measure for classifier performance, $0 < \text{AUC} \leq 1$. We used the original generative model as an optimal model to obtain a gold standard AUC for each test data set. Due to our novel problem setting, we were unable to perform a comparison of our method with respect to existing collective anomaly detection algorithms. Instead, we show how a traditional anomaly detection model, where data points at low density areas of the background model are considered anomalies, fails when anomalies lie within the background. The traditional anomaly detection model estimates the density of the background data using a mixture of Gaussians and then labels new instances that fall below a given density threshold as anomalies. By varying the density threshold, we construct the ROC curves for the traditional model.

Figure 4(a) shows the median of the AUC values obtained using the fixed-background model (FBM). The dashed line (Opt) denotes the median AUC obtained using the generative model itself as a classifier and the gray line at the bottom of the figure (Trad) shows the median AUC obtained using traditional anomaly detection (described above). Given that the test data contains a sufficient amount of anomalies, the resulting AUC values for the fixed-background model are practically identical to the optimal results. However, the robustness of the approach starts to suffer when the test data contains less than 3 % of anomalies. The figure also shows that the results obtained using traditional anomaly detection are significantly worse. The reason for the worse performance is that the anomalies appear within the background, that is, at the high density

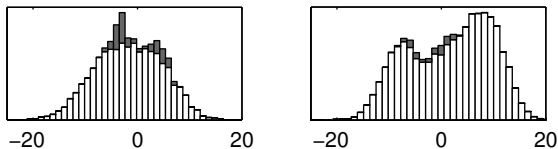


Fig. 3. Two example histograms of the artificial data sets. The gray excess on top of the histogram bars denotes the anomalous data.

areas of the background model. Hence, the traditional anomaly detection method regards them as normal data by definition. Figure 4(b) shows a box plot of the estimated anomaly proportions λ . The small boxes on the diagonal show the interquartile range of the estimated λ s which are in a good agreement with the correct results. The whiskers show the full range of the estimates. The wide downward range results from the algorithm occasionally being able to find only a portion of the anomalous data.

V. DEMONSTRATION: SEARCH FOR THE HIGGS BOSON

We applied the proposed anomaly detection framework to searches of new physical phenomena in particle physics. Such signals usually manifest themselves as tiny excesses of certain types of collision events in particle detectors and the data analysis challenge is to detect and extract these minute signals among a vast background of known physics. The problem can be tackled with the help of machine learning techniques which are an essential tool in improving the signal-to-background ratio in many modern physics analysis scenarios [8].

A. Description of the Data Set

We applied our method to a data set containing a simulated signal produced by the Higgs boson. This is a particle predicted by the Standard Model of particle physics to explain the mass of the other particles in the model but which has yet to be detected experimentally. More precisely, we considered a data set produced by the CDF collaboration [21] containing background events and Monte Carlo simulated Higgs events where the Higgs is produced in association with the W boson and decays into two bottom quarks, $q\bar{q} \rightarrow WH \rightarrow l\nu b\bar{b}$. This signal looks slightly different for different Higgs masses m_H , which is an unknown free parameter in the Standard Model. The goal is to show that semi-supervised anomaly detection is able to identify such a signal without a priori knowledge of m_H . More generally, this could be any set of free parameters in the physical theory under consideration.

Each observation in the data set corresponds to a single simulated collision event in the CDF detector at the Tevatron proton-antiproton collider. As such, the data vectors are statistically independent and consist of 8 variables corresponding to different characteristics of the topology of a collision event. To facilitate density estimation with MoGs, the data were normalized logarithmically and the dimensionality of the data was reduced to 2 using PCA on the background data. The dimensionality reduction also allows better visualization of the results. Dimensionality reduction for the test data was carried out with the same principal components obtained using the background data.

We used 3406 data points to train the background model which was then used to detect signals of 400 data points for masses $m_H = 100, 115, 135, 150$ GeV among another sample of 3406 observations of background data. Hence, the unlabeled sample contained 10.5 % of signal events. In reality, the expected signal is roughly 5 to 50 times weaker than this, but due to the limited number of background events available,

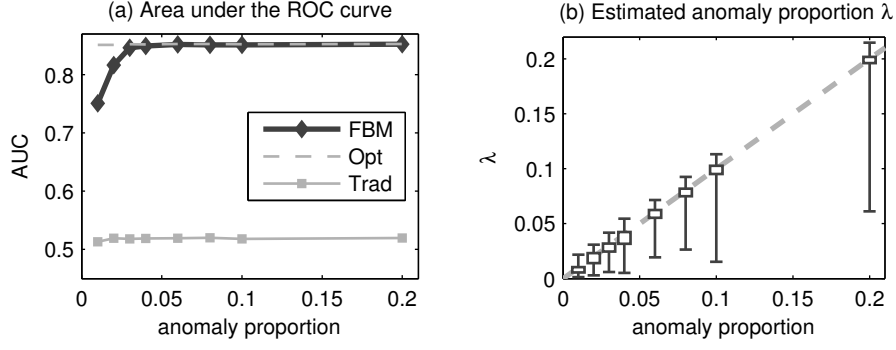


Fig. 4. The results of our artificial data experiments with unlabeled test data sets containing 100 000 data points. (a) Comparison between AUC of the fixed-background model (FBM), the generative model used to generate the data (Opt) and traditional anomaly detection (Trad) with different amounts of anomalies in the test data. Each data point in the plot is the median of 100 runs. (b) Estimation of the anomaly proportion (λ) using the fixed-background model. The small boxes show the interquartile range and the whiskers the full range of the estimates. The gray dashed line shows the correct anomaly proportion.

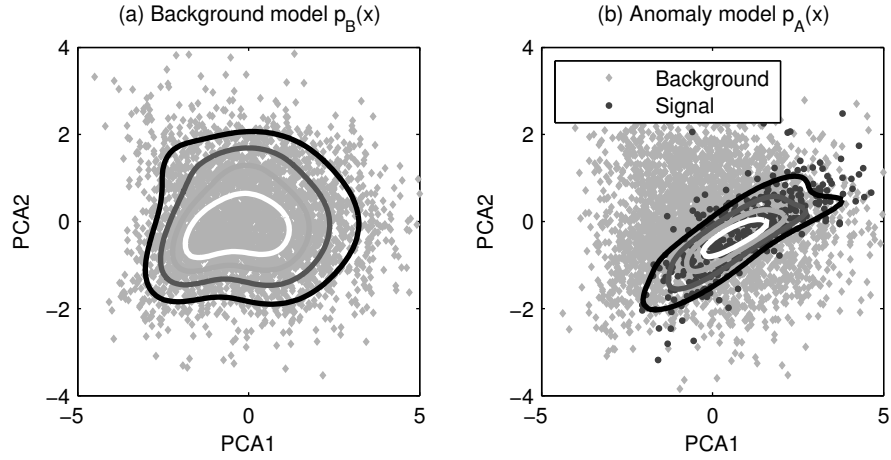


Fig. 5. (a) A projection of the Higgs background data into its two-dimensional principal subspace. The solid lines show contours of the estimated 5-component MoG for the background. (b) A projection of the $m_H = 150$ GeV test data set into the two-dimensional principal subspace. The solid lines show contours of the estimated 2-component MoG for the signal.

the signal had to be amplified for this demonstration. As shown by the experiments with artificial data, we expect to be able to find also weaker signals should more background observations be available. All in all, these experiments should merely be regarded as a demonstration of the potential of the method in physics data analysis and not as a realistic Higgs analysis scenario which remains as one of the greatest experimental challenges of modern particle physics.

B. Modeling the Higgs Data

We used cross-validation-based information criterion (CVIC) [20] in order to select a suitable number of components J for the background model. When a 5-fold cross-validation was performed, the evaluation log-likelihood was maximized with $J = 5$. Figure 5(a) shows contours of the resulting background model in the two-dimensional principal subspace.

We then learned the fixed-background models for the signals with different masses starting with $Q = 3$ anomalous components and allowed for heuristic removal of unnecessary

components as described in Section III-D. The algorithm converged with one anomalous component for $m_H = 100$ GeV and two components for the rest of the masses. The resulting anomaly model for $m_H = 150$ GeV is shown in Figure 5(b).

C. Anomaly Detection Results

The statistical significances of the anomaly models were evaluated using the bootstrap technique with 50 000 resamplings. It was found out that at 5 % significance level the background-only null hypothesis was rejected for all of the considered mass points. Figure 6 shows the distribution of the test statistic and the p -values of the models. It turns out that the higher the mass, the more significant the model becomes. The peak of the test statistic distribution at the origin results from situations where all the components of the anomaly model are correctly removed by the removal heuristics.

Figure 7(a) shows the ROC curves for anomaly detection with different Higgs masses. One can see that regardless of the mass of the Higgs, the algorithm is able to identify the signal with a relatively constant accuracy. The classification results

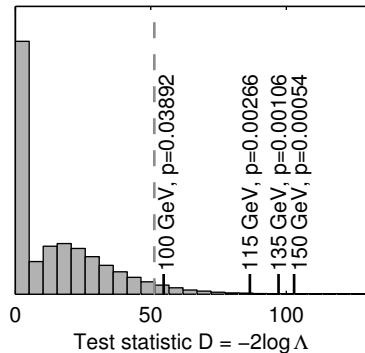


Fig. 6. Likelihood ratio test for the significance of the anomaly model for various Higgs masses. The histogram shows the probability distribution of the likelihood ratio test statistic under the background-only null hypothesis. The vertical dashed line shows the critical value of the test at 5 % significance level and the black markers denote the test statistics for the fixed-background models with respective p -values. All observed test statistics fall on the critical region of the test leading to the rejection of the null hypothesis.

are slightly better with the higher masses because the high-mass signal lies on a region of the data space with a lower background density compared to the low-mass signal. Starting from the lowest mass, the estimated anomaly proportions are $\lambda = 0.100, 0.121, 0.118, 0.122$, which are all in agreement with the real proportion of 0.105. These estimates could be used to compute the cross section of the anomalous physics process.

We also used supervised multi-layer perceptron neural networks (MLPNN) for each of the mass points to act as a reference classifier and compared the ROC curves to the ones obtained with anomaly detection. MLPNNs with two hidden layers containing 10 and 5 neurons and trained using MATLAB’s Neural Network Toolbox were used. Figure 7(b) shows the ROC curves for an MLPNN trained using the Higgs signal at $m_H = 150$ GeV. The resulting ROC curve for a test signal at the same mass is similar to the one obtained using FBM. However, when the mass m_H of the test signal is varied, the classification performance of the MLPNN decreases. Figure 7(c) shows the ROC curves for 4 separate MLPNNs trained and tested using the same mass. The results are comparable with FBM, which shows that semi-supervised anomaly detection is able to achieve similar performance as supervised classification when the mass is known a priori. It should be noted that in this application domain, one does not expect to see perfect classification results as signals are often buried among an irreducible physics background. Instead, the key advantage of anomaly detection is that it is able to identify the signal without the need to specify the mass of the Higgs, while a supervised classifier is able to efficiently identify only the mass it has been trained for.

VI. DISCUSSION

The proposed semi-supervised anomaly detection method is applicable to problems where anomalies lie collectively among the background data, or put in other words, to problems where we want to find an unexpected, unknown or uncertain

signal that does not appear in the known background data. We showed that the method can be applied to model-independent searches of new phenomena in particle physics, but other potential application domains could include, e.g., astrophysics, bioinformatics and electronic surveillance, as long as the stated assumptions on the general problem setting are satisfied.

The general idea of semi-supervised anomaly detection with the fixed-background model can be implemented in a number of different ways. In order not to divert the reader’s attention from the novel problem setting and its application potential, we have deliberately chosen to use simple, well-understood algorithmic techniques with their known limitations related to the curse of dimensionality, model complexity and convergence. We show that even such a simple approach provides useful results in our application domain. However, it is likely that some of these limitations could be alleviated by using state-of-the-art techniques. For example, variational Bayes [22] could provide a natural way of learning the model complexity and incorporating prior information into the problem, while parsimonious mixtures of Gaussians [23] might be a worthwhile alternative to dimensionality reduction with PCA. The framework is also not dependent on the use of Gaussian mixture models. Indeed, other problem-specific parametric models can easily be accommodated. Also non-parametric density estimates of the background model can be implemented in a straightforward manner.

VII. CONCLUSIONS

We have presented a semi-supervised anomaly detection framework based on the fixed-background mixture model. The proposed model assumes that the background data follow a fixed distribution, thus providing the means to detect and perform pattern recognition of collective anomalies that lie within the domain of the background data and manifest themselves as deviations from this distribution. We showed that the algorithm is robust enough to consistently model anomalous patterns that make up only a few percent of an unlabeled data set. We demonstrated that such an algorithm can be successfully used in model-independent searches of new phenomena in high energy particle physics. In particular, the method could be used to find new particles without exact a priori knowledge of their properties. Given the generality of the framework, it should be possible to find future applications also on other fields of science and technology.

ACKNOWLEDGEMENTS

The authors are grateful to the CDF collaboration for providing access to the Higgs signal and background Monte Carlo samples, to the Academy of Finland for financial support and to Matti Pöllä, Timo Honkela and Risto Orava for valuable discussions and feedback on the manuscript.

REFERENCES

- [1] D. M. Hawkins, *Identification of Outliers*, ser. Monographs on Applied Probability and Statistics. Springer, 1980.
- [2] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Computing Surveys*, vol. 41, pp. 15:1–15:58, 2009.

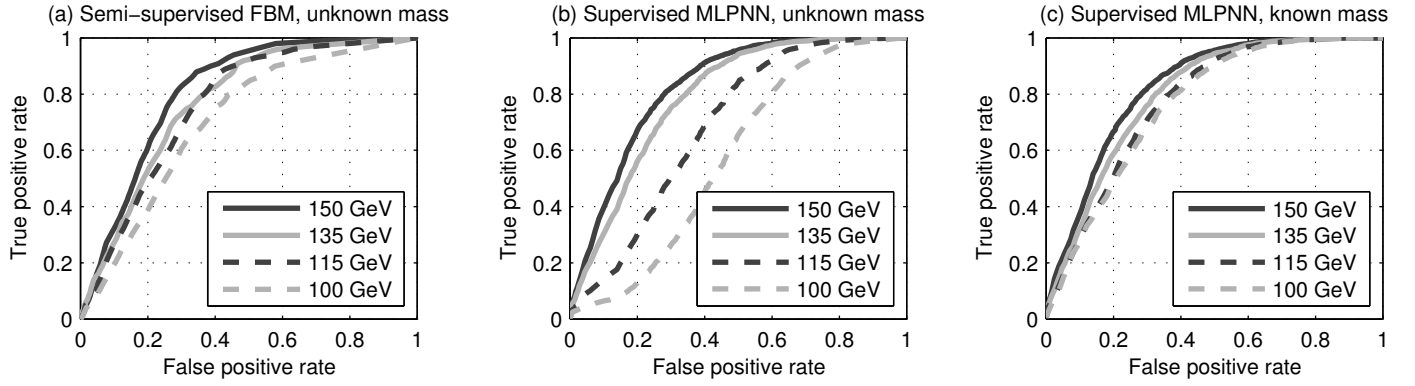


Fig. 7. ROC curves for the Higgs signal with various Higgs masses m_H with (a) the semi-supervised fixed-background model (FBM), (b) the multi-layer perceptron neural network (MLPNN) with unknown m_H and (c) MLPNN with known m_H . The FBM is able to identify the signal without a priori knowledge of the mass. The MLPNN in (b) was trained using the signal $m_H = 150$ GeV, hence the performance decreases when signals with other masses are classified. In (c), a different MLPNN was trained for each signal.

- [3] M. Markou and S. Singh, "Novelty detection: A review - part 1: Statistical approaches," *Signal Processing*, vol. 83, pp. 2481–2497, 2003.
- [4] —, "Novelty detection: A review - part 2: Neural network based approaches," *Signal Processing*, vol. 83, pp. 2499–2521, 2003.
- [5] E. Aleskerov, B. Freisleben, and B. Rao, "CARDWATCH: a neural network based database mining system for credit card fraud detection," in *Computational Intelligence for Financial Engineering (CIFEr)*, 1997. *Proceedings of the IEEE/IAFE*, March 1997, pp. 220–226.
- [6] D.-Y. Yeung and C. Chow, "Parzen-window network intrusion detectors," in *Proceedings of the Sixteenth International Conference on Pattern Recognition*, 2002, pp. 385–388.
- [7] L. Yu, D. Cleary, and P. Cuddihy, "A novel approach to aircraft engine anomaly detection and diagnostics," in *Aerospace Conference, 2004. Proceedings of the IEEE*, vol. 5, March 2004, pp. 3468 – 3475.
- [8] P. C. Bhat, "Advanced analysis methods in particle physics," *Annual Review of Nuclear and Particle Science*, vol. 61, pp. 281–309, 2011.
- [9] T. Aaltonen *et al.*, "Model-independent and quasi-model-independent search for new physics at CDF," *Physical Review D*, vol. 78, p. 012002, Jul 2008.
- [10] E. Eskin, "Anomaly detection over noisy data using learned probability distributions," in *Proceedings of the International Conference on Machine Learning*. Morgan Kaufmann, 2000, pp. 255–262.
- [11] M. Lauer, "A mixture approach to novelty detection using training data with outliers," in *Lecture Notes in Computer Science*. Springer, 2001, pp. 300–311.
- [12] G. Ritter and M. T. Gallegos, "Outliers in statistical pattern recognition and an application to automatic chromosome classification," *Pattern Recognition Letters*, vol. 18, no. 6, pp. 525 – 539, 1997.
- [13] K. Knight, *Mathematical Statistics*. Chapman and Hall, 2000.
- [14] D. R. Cox and D. V. Hinkley, *Theoretical Statistics*. Chapman and Hall, 1974.
- [15] S. Wang, W. A. Woodward, H. L. Gray, S. Wiechecki, and S. R. Sain, "A new test for outlier detection from a multivariate mixture distribution," *Journal of Computational and Graphical Statistics*, vol. 6, no. 3, pp. 285–299, 1997.
- [16] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. Chapman and Hall, 1993.
- [17] G. McLachlan and D. Peel, *Finite Mixture Models*, ser. Wiley Series in Probability and Statistics. Wiley-Interscience, 2000.
- [18] 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)*, vol. 39, no. 1, pp. 1–38, 1977.
- [19] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions (Wiley Series in Probability and Statistics)*, 2nd ed. Wiley-Interscience, 2008.
- [20] P. Smyth, "Model selection for probabilistic clustering using cross-validated likelihood," *Statistics and Computing*, vol. 10, pp. 63–72, 2000.
- [21] 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.
- [22] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Science+Business Media, LLC., 2006.
- [23] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.